Retrieving Implicit Relations from Text: Hidden Semantics and Natural Language Processing

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 12
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Niko Schenk
aus Sindelfingen

Frankfurt 2019
(D 30)

vom Fachbereich 12 der

Johann Wolfgang Goethe–Universität als Dissertation angenommen.

Dekan: Prof. Dr.-Ing. Lars Hedrich

Gutachter: Prof. Dr. Christian Chiarcos, Prof. Dr. Gert Webelhuth

Datum der Disputation: 20. März 2019

Acknowledgements

This thesis represents not only a written summary of my research, it is the result of many years of extensive work, fruitful discussions and exchange of personal ideas with many very interesting people that I met along the way all around the world.

I would like to reflect in particular on the people who have guided, supported and helped me throughout this period. First and foremost, I wish to express my gratitude to my main advisors. I am extremely grateful to Professor Christian Chiarcos, a passionate scientist, who offered me the opportunity to be a member of his lab and who kindly helped me to develop the thesis topic on implicit information. He kept me motivated and on track during the initial exploratory phase of the dissertation. I am very thankful for Professor Chiarcos' enthusiasm when he explained things to me, and his patient guidance, especially when something was unclear to me, throughout the complete course of my research and writing phase. From him I learned about writing scientific papers, where his exemplary manner taught me how excellent research in the field of computational linguistics is conducted. Moreover, Professor Chiarcos always had a sympathetic ear for me whenever I was in doubt, gave me the freedom to develop my own ideas and to explore further research directions while constantly providing help when needed. I am extremely grateful to him for the unique opportunity to travel and attend the major NLP conferences, which made it possible for me to meet with researchers outside of my office and exchange ideas with them.

In the very same way, I would like to sincerely thank Professor Gert Webelhuth who was the first to invite to Frankfurt. At Goethe University, he was the very first person I made contact with, and I clearly recall the overwhelming warm and friendly atmosphere in his research group, which was the main reason that prompted me to stay in Frankfurt: a city that I still call home today. I really appreciated the occasional paper sessions, especially the one on Centering Theory, which has significantly contributed to the interdisciplinary character of my work. With Professor Webelhuth I found not only an excellent supervisor who has an immense knowledge and enormous experience in the field of linguistics, but also a personal mentor whose competence and friendly coaching personality (the most effective cheering during the JP Morgan Corporate Challenge) gave me substantial support not only during my time as an employee in his department but also outside of University.

Even though I truly could not have imagined having better advisors for my doctoral studies, I am equally indebted to a number of fellow students and people I got to know on various occasions.

Samuel Rönnqvist who I first met at MLSS 2013 and became a very good friend in the meantime. Thank you for the valuable discussions, collaborative programming sessions, the invitation to your defense in Turku, and the fun times we spent together, especially at ACL in Vancouver. I remember the physically most challenging task during that conference was definitely the path up to Grouse Mountain. I guess, Miriam Käshammer would agree as well.

A warm word for my great buddy Philip Schulz who has oftentimes impressed me not only as a research enthusiast but also as a critical thinker from whom I learned a lot. Thanks for the valuable discussions we had on the various conference locations for which we traveled between Berlin, Amsterdam, Frankfurt, and Osaka. I really appreciate the time with someone who knows how to properly combine a healthy balance between focused academic research and parties all night long.

A special acknowledgement goes to Aya Zirikly who I met in San Diego at NAACL in 2016. Aya, you are a lovely person and I am truly grateful for the fun times we shared and also for all your academic advice and help both as a colleague and as a friend.

I also thank my former fellow (ISCL) students from Tübingen who, every time we meet, make me feel that I have true friends: Maria Schmidt, Marisa Köllner and her family, Johannes Wahle, Isabel Trott.

I would like to acknowledge the NLP people in Frankfurt, in particular my office mates for making the ACoLi lab a great place to work. One academic colleague stands out: Armin Hoenen who supported me a lot in reassuring and helping me out with my thousands of administrative questions that I had before the submission of this thesis.

I would also like to acknowledge the support from colleagues from other departments and areas of NLP: Evgeny Stepanov, who I met in Beijing, for his assistance in providing help to run his discourse parser. I would like to thank Joyce Chai and Hendrik De Smet for providing us access to their resources and corpora. I am grateful to Egoitz Laparra for sending evaluation scripts, and would also like to acknowledge Biao Zhang, Jifan Chen, and Sujian Li for sending and granting permission to reuse their graphics and/or code. Finally, the quality of the Chinese translations would not have reached human level performance without the help of Wei Ding.

Last but not least I am indebted to my friends in Frankfurt: Monika, Michael, Matthias, Angela, and, in particular, **Zita**.

And my most heartfelt thanks of course to my family: Bärbel, Alan, Eva, Wolfgang, Manfred, thank you for all the moral support and the amazing chances you have given me over the years.
My **mother**, my **father**, and my little **brother** Alex and his family: Thank you for your love. You made me into who I am.

Lastly, I would like to point out that all the work would not have been possible without the support of the Springer Nature publishing company which provided funding for parts of the work in this dissertation.

Abstract

Human readers have the ability to infer knowledge from text, even if that particular information is not explicitly stated. In this thesis, we address the phenomena of text-level implicit information and outline novel automated methods for its recovery. The main focus of this work is on two types of unexpressed content that arises between sentences (implicit discourse relations) and within sentences (implicit semantic roles). Traditional approaches mostly rely on costly rich linguistic features, e.g., sentiment or frame-based lexicons, and require heuristics or manual feature engineering. As an improvement, we propose a collection of generic resource-lean methods, implemented in the form of statistical background knowledge or by means of neural architectures. Our models are largely language-independent and produce state-of-the-art performance, e.g., in the classification of Chinese implicit discourse relations, or the detection of locally covert predicative arguments in free texts. In novel experiments, we quantitatively demonstrate that both types of implicit information are mutually dependent insofar as, for instance, some implicit roles directly correlate with implicit discourse relations of similar properties. We show that implicit information processing further benefits downstream applications and demonstrate its applicability to the higher-level task of narrative story understanding. In the conclusion of the dissertation, we argue for the need of implicit information processing in order to realize the goal of true natural language understanding.

Kurzzusammenfassung

Beim Lesen und Verstehen von Texten ziehen wir Rückschlüsse auf Informationen, welche nicht explizit formuliert sind. Beispielsweise kann ein Grund auch ohne das Wort *weil* beschrieben sein; ein Imperativ muss nicht ausdrücklich denjenigen benennen, dem befohlen wird. Wir verstehen diese impliziten Verknüpfungen intuitiv, ein Computer kann diese jedoch nicht ohne Weiteres erkennen. Diese Dissertation befasst sich mit der Sprachverarbeitung impliziter Information und ihren Relationen und stellt neue Ansätze vor, um diese automatisiert in Texten zu erkennen. Hierbei werden schwerpunktmäßig zwei Arten unterschieden: sprachlich nicht realisierte Inhalte, welche zwischen Sätzen auftreten (implizite Diskursrelationen) und solche, welche innerhalb eines Satzes hervorgerufen werden (implizite semantische Rollen). Herkömmliche Methoden stützen sich auf aufwendige, manuell erstellte linguistische Ressourcen, beispielsweise Sentiment- oder Frame-basierte Lexika, welche darüber hinaus Heuristiken oder manuelles Feature-Engineering erfordern.

Diese Dissertation stellt Verbesserungsansätze dar in Gestalt von generischen und gleichzeitig ressourcenarmen Techniken. Diese sind zum einen in der Form statistischer Hintergrundinformation implementiert, zum anderen mit Hilfe von künstlichen neuronalen Netzen. Die vorgestellten Modelle zeichnen sich größtenteils durch Sprachunabhängigkeit aus und repräsentieren den Stand der Technik, zum Beispiel in der Klassifikation impliziter chinesischer Diskursrelationen oder bei der Erkennung lokal unrealisierter Argumente in freien Texten. In neuen Experimenten lässt sich quantitativ zeigen, dass es eine wechselseitige Beziehung zwischen beiden Arten an impliziter Information gibt. Dies manifestiert sich darin, dass einige implizite Rollen eine direkte Korrelation mit impliziten Diskursrelationen vom selben Typ aufweisen. Der praktische Nutzen impliziter Informationsgewinnung liegt darin, dass sie Grundlage für Folgeanwendungen, wie beispielsweise Question-Answering-Systeme darstellt oder, wie in dieser Arbeit demonstriert, direkt zur automatisierten Analyse der kohärenten Erzählstruktur eines Textes angewandt werden kann.

Zusammenfassend wird ein Ausblick gegeben und erörtert, dass die Erkennung und Verarbeitung impliziter Information einen wesentlichen Bestandteil in der Realisierung intelligenter Systeme darstellen wird, die natürliche Sprache verstehen.

Die vorliegende Arbeit ist in fünf Teile gegliedert und wie folgt strukturiert.

In Teil I, Kapitel 1 wird das Phänomen der impliziten Information und ihren assoziierten Relationen in natürlichsprachlichen Texten motiviert, sowie deren Bedeutung für die automatisierte Sprachverarbeitung erläutert. Die Beschreibung stützt sich hierbei auf Theorien, Beobachtungen und Erkenntnisse der klassischen (psycho)linguistischen Literatur (Horn, 1984; Givón, 1995; Carston, 2006), welche unter anderem besagen, dass wir im Zuge einer effizienten Sprach- und Textproduktion Äußerungen in bestimmten Kontexten unrealisiert lassen, mit dem Ziel Redundanz zu vermeiden und um Kohärenz zu wahren. Da herkömmliche Techniken zur automatisierten Informationsextraktion aus Texten darin beschränkt sind, dass sie lediglich das verarbeiten können, was explizit ausgedrückt ist, werden diesbezüglich eine Reihe von Anwendungsszenarien beschrieben, die zeigen, wie sich klassische Methoden durch Erweiterung auf Informationsgewinnung impliziter Information qualitativ verbessern lassen. Beispielsweise kann ein Question-Answering-System die Erkennung eines impliziten Kausalzusammenhangs zwischen zwei Sätzen direkt dafür verwenden, einem Benutzer eine Antwort auf die Frage nach einem Grund zu liefern.

In Teil II der Dissertation werden zunächst die Grundlagen für die automatisierte Verarbeitung impliziter Diskursrelationen beschrieben. Unter dem Begriff *Implicit Discourse Parsing* werden schwerpunktmäßig Methoden zur Erkennung und Klassifikation der Bedeutungsrelation zwischen zwei verknüpften Äußerungen in unstrukturierten Texten vorgestellt.

Kapitel 2 liefert hierfür einführend einen theoretischen Überblick und beschreibt die in der Literatur etablierten Diskurs-Frameworks, insbesondere Centering Theory (Grosz u. a., 1995), Rhetorical Structure Theory (Mann und Thompson, 1988), und die Penn- und Chinese Discourse Treebank (Prasad u. a., 2008; Zhou und Xue, 2012). Letztere stellt aufgrund ihres flachen Annotationsschemas eine besonders geeignete Ressource für die computationelle Modellierung impliziter Diskursstruktur dar und bildet deshalb die Datengrundlage der im weiteren Verlauf dieser Arbeit beschriebenen Techniken.

Kapitel 3 stellt ein ressourcenschwaches, jedoch gleichzeitig höchst effizientes und sprachunabhängiges neuronales Modell zur automatisierten Erkennung der Diskursrelationen zwischen zwei Äußerungen vor. Dessen Architektur ist inspiriert von ersten Ansätzen, welche sich ausschließlich auf Word Embeddings zur Modellierung beschränken (Zhang u. a., 2015). Das hier vorgestellte Modell stellt darüber hinaus allerdings eine Verbesserung dar, indem Embeddings im beschriebenen Ansatz strukturell diversifiziert werden, sie durch die direkte Integration syntaktischer Information einen Mehrwert erhalten und letztlich dadurch, dass die Repräsentation der Argumentstruktur weniger Kompositionsoperationen erfordert als in vergleichbaren Vorarbeiten.

In Kapitel 4 wird eine direkte Modifikation des vorangehenden Ansatzes aus Kapitel 3 beschrieben. Hierbei stützt sich die Modellierung auf die Tatsache, dass die kognitive Verarbeitung von Diskursinformation bei uns Menschen faktisch am plausibelsten durch sequentielle Weise zu erklären ist. Die praktische Implementierung dieses Aspekts wird durch ein rekurrentes neuronales Netz mit Attention-Mechanismus realisiert, welches darüber hinaus von einer neuartig eingeführten Technik zum Sampling von Trainingsinstanzen profitiert. Es lässt sich in einer Evaluation auf einer standardisierten Datenmenge zeigen, dass die Klassifikationsgenauigkeit im Gegensatz zu Modellen, welche die Reihenfolge der Erscheinung einzelner Worte außer Acht lassen, effektiv gesteigert werden kann. Die in Kapitel 4 beschriebene Technik zeichnet sich in hohem Maße dadurch aus, dass sie einen Einblick in die während der Klassifikation gelernten distinktiven Features, insbesondere für Kohärenzrelationen zwischen Entitäten, gewährt.

In Teil III der Arbeit liegt der Schwerpunkt auf der automatisierten Erkennung und Auflösung impliziter semantischer Rollen. Unter dem Begriff *Implicit Semantic Role Labeling* werden zunächst die theoretischen Hintergründe und Vorarbeiten in diesem Bereich dargestellt.

Kapitel 5 behandelt Verbvalenz und Argumentstruktur als theoretische Fundamente zur Erklärung und Beschreibung lokal unrealisierter Rollen (Fillmore, 1986; Ruppenhofer, 2005). Es werden diesbezüglich exemplarisch eine Reihe manuell erstellter Ressourcen vorgestellt, insbesondere FrameNet (Baker u. a., 1998), PropBank (Palmer u. a., 2005) und NomBank (Meyers u. a., 2004), welche diese lexikalischen Eigenschaften direkt für individuelle verbale und nominale Prädikate kodieren. Diese Ressourcen bilden die Grundlage für die im Folgenden vorgestellten Methoden.

Kapitel 6 beschreibt einen neuartigen Ansatz zur Bestimmung jener semantischer Rollen, welche in speziellen Kontexten lokal unrealisiert sind. Im Kern werden hierfür statistische Generalisierungen expliziter Rollenmuster im PropBank-Stil erzeugt, welche auf der Grundlage von großen automatisiert ausgezeichneten Korpora beruhen. Die kombinierten Auftrittswahrscheinlichkeiten von Prädikaten mit ihren dazugehörigen semantischen Rollen werden direkt zur Erkennung impliziter Argumente verwendet, was es ermöglicht implizite Rollen auch ohne sprach- und domänenspezifische Lexika zu bestimmen. In einer Evaluation auf handannotierten Daten lässt sich zeigen, dass sich die vorgestellte Methode durch eine größere Flexibilität, beispielsweise in der Modellierung unterschiedlicher Word Senses oder der Erweiterung auf beliebige Modifikator-Rollen auszeichnet, sowie einer höheren Trefferquote im Vergleich zu herkömmlichen Ansätzen mit Lexikon-Templates.

Kapitel 7 stellt daran anknüpfend einen unüberwachten Lernansatz zur Auflösung und Verlinkung adäquater syntaktischer Konstituenten für unrealisierte Rollen im Diskurs dar. Hierfür werden auf dieselbe Art und Weise Generalisierungen über Massendaten erzeugt – in diesem Fall durch Word Embeddings –, welche die Generierung prädikats- und rollenspezfischer Prototypen erlauben. Ist eine Rolle als implizit erkannt, werden anhand der Prototypen und mittels distributioneller Ähnlichkeit Kandidaten im Diskurskontext selegiert. Eine Visualisierung des Modells zeigt, dass mit der vorgeschlagenen Methode interpretierbare Eigenheiten einzelner Rollen gelernt werden können. Darüber hinaus stellt die Methode eine Möglichkeit für eine direkte Baseline-Implementierung für implizite semantische Rollen dar und bietet eine flexible Alternative für Domänen und Sprachen, welche nur in geringem Umfang durch NLP-Tools und Ressourcen abgedeckt sind.

In Teil IV der Dissertation werden zwei Brückenexperimente für die einheitliche Analyse impliziter Diskursrelationen und semantischer Rollen vorgestellt. Darüber hinaus wird eine exemplarische Integration der in dieser Dissertation eingeführten Methoden in eine praktische Anwendung beschrieben.

Kapitel 8 illustriert hierfür ein neuartiges Experiment mit dem Ziel quantitativ zu ermitteln, welchen Einfluss satzinterne implizite Information, realisiert durch implizite Rollen, auf satzübergreifende implizite Diskursrelationen hat. Die vorgestellte Methodik basiert unter anderem auf Experimenten mit impliziten Kausalitätsverben (Asr und Demberg, 2015; Kehler und Rohde, 2017). Diese zeigen, dass individuelle Prädikate auf Wortebene die Erwartung nach Diskurskontinuität in Form spezieller Relationen erzeugen. Das vorgestellte Experiment ist insofern neuartig, als nicht nur einzelne Prädikate oder Rollen als Indikatoren betrachtet werden, sondern vollständige Rollenkonstellationen. Es lässt sich zeigen, dass einige implizite Rollen mit impliziten Diskursrelationen desselben Typs, wie zum Beispiel Kausalität, erwartungsgemäß korrelieren.

Kapitel 9 beschreibt die Adaption der Architektur zur Modellierung von Diskursinformation aus Kapitel 4 auf den lokalen Prädikatkontext. Ziel dieses Experiments ist es zu testen, inwiefern ein Diskursparser zur Klassifizierung semantischer Rollen geeignet ist, welche potentiell koreferente Verbindungen zwischen den Diskursargumenten kodieren. Die theoretische Motivation und Grundlage dieses Experiments entstammt einem Kernaspekt der Centering Theory (Grosz u.

a., 1995), welcher in Form von Transitionsrelationen die Salienz von Entitäten im Diskurs modelliert. Das beschriebene Experiment demonstriert, dass sich die Annahmen der Theorie praktisch mit der vorgestellten Methode implementieren lassen, indem die unterschiedlichen Transitionsrelationen auf Ebene koreferenter semantischer Rollen kodiert und klassifiziert werden können.

Kapitel 10 demonstriert wie sich eine praktische Anwendung zur automatisierten Modellierung der Erzählstruktur eines Textes mittels impliziter Informationsextraktion realisieren lässt. Es wird dargestellt, wie die Diskursarchitektur aus Kapitel 3 zur Erkennung semantisch kohärenter Verbindungen zwischen einzelnen Komponenten einer Kurzgeschichte dienen kann, um eine vollautomatische Unterscheidung zwischen adäquaten und unpassenden Schlüssen für einen narrativen Text zu treffen. Der beschriebene Ansatz stellt einen Mehrwert dar, da er ressourcenschwach ist, und darüber hinaus gute Ergebnisse in einer offiziellen Evaluation gegenüber reimplementierten Baseline-Systemen erzielt, welche auf handkodierte Konzeptinformation wie Skripte zurückgreifen müssen.

Der letzte Teil V der Dissertation befasst sich primär mit einer Methodenreflektion und beinhaltet weiterführende Überlegungen auf Grundlage der vorgestellten Ansätze. Speziell in Bezug auf die besonderen Kohärenzrelationen zwischen Entitäten, welche als solche in der Penn und Chinese Discourse Treebank annotiert sind und welche seit jeher im Fokus der klassischen Diskursframeworks stehen, wird nochmals bekräftigt, dass diese nicht nur einfach einen wiederholten Bezug auf dieselbe Entität darstellen, sondern in gleichen Maße Diskursinformation im Sinne der klassischen Bedeutungsrelationen tragen. Kapitel 11 liefert darüber hinaus in Details beleuchtete, punktuelle Aspekte für Verbesserungsansätze der vorgestellten Techniken. Außerdem werden direkte Anwendungsszenarien beschrieben, beispielsweise inwiefern die Erkennung impliziter semantischer Rollen in nutzergenerierten Inhalten dazu beitragen kann, einen erhöhten Informationsgehalt in Rezensionstexten zu gewährleisten.

Abschließend wird ein Bezug auf Sprachtechnologie für unseren Alltag hergestellt. Die Bedeutung der impliziten Informationsverarbeitung wird nochmals verdeutlicht und hervorgehoben, dass sie nicht nur eine Verbesserung der herkömmlichen Informationsextraktion darstellt, sondern auch eine unerlässliche Komponente dafür ist, intelligente Sprachverarbeitung in der Zukunft zu realisieren.

Contents

Part I

Introduction

Chapter 1

Implicit Information in Text

1.1 Motivation

The development of text-based information retrieval (IR) systems has provably made considerable progress within the last few years. Nowadays, users of web search engines browse the content of massive amounts of online documents reliably using keywords. They benefit directly from structured information in the form of rich knowledge graphs[1] as an intelligent supplement to their research needs while, for example in the medical domain, novel relation extraction techniques support users in clinical decision making (Wang and Fan, 2014). Commonly, Natural Language Processing (NLP) techniques further refine the output quality of IR systems in granting a structured analysis to the raw textual content of a document. These applications provide means to spot those keywords (Beliga et al., 2015) and relations among entities (Surdeanu et al., 2011), but also to recognize location and person names (Nadeau and Sekine, 2007), parse sentences into linguistically motivated syntactic units (Chen and Manning, 2014), or assign thematic role relationships to events and associated participants in a text (Roth and Lapata, 2016), for instance, to distinguish subjects from objects in a sentence or to determine semantically what happened to whom, when, and where.

For the most part, standard NLP tools implement methods of *explicit information acquisition*. This means that a keyword or a relation in a knowledge graph can only be annotated and later on extracted if it can be directly located at some specific position within a given document, i.e. it must be either explicitly expressed in the textual body (of the web page), or explicitly associated to it as meta data. Crucially, however, a large quantity of information is allocatable only in an **implicit** form. In texts, this type of information is unexpressed and thus *cannot* be captured by conventional IR.

One of the reasons why implicit phenomena exist may be due to an evolutionary, natural efficiency in language and text production: for example, in consecutive sentences of a narrative story not every piece of information (e.g., the cause of

[1] Cf. `https://googleblog.blogspot.de/2012/05/introducing-knowledge-graph-things-not.html`, accessed December 2017.

an effect, the name of a protagonist, or the location of an arrival) is maximally explicitly stated or continuously repeated in the sentential description of subsequent events because these missing pieces can be easily inferred and interpreted by the reader—either through world knowledge, or from the context of the story. From a cognitive processing perspective, doing so would in fact lead to redundancy and would make human sentence comprehension unnecessarily complex. In this context, Givón (1995) argues that *"most coherent—interpretable—texts fall somewhere in the middle between the two extremes of total redundancy and utter incoherence"*, whereas Horn (1984) in his famous speaker-based *R Principle* states accordingly that *"you should not say more than you must"*. Carston (2006) refers to this observation by a related phenomenon of *linguistic underdeterminacy* and attributes it to pragmatic factors in communication in which a produced sentence does not necessarily reflect a full encoding of the thoughts or propositions explicitly conveyed by the speaker.

The challenging task of Natural Language Understanding (Allen, 1995, NLU) goes beyond conventional information retrieval and builds on aspects of implicit information in text. Generally, NLU comprises a whole range of more sophisticated, higher-level applications, involving recent advances in question answering (Rao et al., 2016), text generation and commonsense reasoning in stories (Mostafazadeh et al., 2016a), text summarization (Zeng et al., 2016), text simplification (Nisioi et al., 2017), or the related subtasks of entailment recognition (Sha et al., 2016), processing of discourse coherence (Li and Jurafsky, 2017), coreference resolution (Lee et al., 2017), or event detection (Zhou et al., 2017). The methodologies in this strand of research achieve advanced inference capabilities beyond isolated words and sentences by relying on neural information processing techniques and the incorporation of distributed word representations. Their core properties make it possible to capture, associate, and combine latent, textually unexpressed facts in addition to the overt words in a text by virtue of generalizations over syntactic and semantic co-occurrences.[2] Keeping this in focus, I argue that in order to realistically approach the desired goal of NLU, text processing does in fact require a realization of a deeper analysis, striving for more elaborate techniques than the shallow surface processing of explicit information mining. In this thesis, I account for this issue, and propose a collection of neural (and neural-related) methods for the *detection, analysis, and interrelation of implicit information in text*, paving the way for improving the quality and effectiveness of conventional NLP and IR, as these systems would significantly benefit from the recovery and integration of textually unexpressed content and their associated relations.

[2] For example, a distributed word representation for the word *king* would capture aspects of the semantically related word *queen*, even when the latter is not overtly expressed in a given text, cf. Mikolov et al. (2013c).

1.1.1 Sentence-Internal & Inter-Sentential Implicit Information

The peculiarities of linguistically non-overt information have been studied in the literature before. For a description of its specific properties, I refer to a broad subdivision into two variants.

Within sentences, locally unexpressed items are typically addressed on the level of predicate-argument structure. In an early, theoretical account to explain lexically unrealized arguments, Fillmore (1986) states that the omissibility of an argument (as, for instance, in *She promised.*) is due to an idiosyncratic lexical semantic feature of a particular verb. He roughly distinguishes two types of categories: In case of *definite null complements*, the missing information can be (anaphorically) retrieved from the context, whereas *indefinite null complements* can be interpreted existentially and do not need to be resolved in the context; cf. Ruppenhofer (2005); Scott (2006); Németh and Bibok (2010), inter alia. A range of recent computational approaches have been suggested for the recovery of null complements in text. The proposed techniques operate on the more general paradigm of *semantic roles* and underlie different linguistic frameworks of distinct granularities, e.g., FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), or NomBank (Meyers et al., 2004). The recognition of **implicit semantic roles**, i.e. the detection, resolution, and linking of locally uninstantiated arguments of mostly nominal and verbal predicates, is a highly challenging task. Recent comprehensive overviews are given in Gerber (2011), Roth (2014), and Laparra (2015).

Interestingly, PropBank defines a few semantic roles which point to (sentence-initial or final) discourse markers (*But, . . . , too.*) or causality (*because*). These roles are special in the sense that they indirectly connect the local predicate context with other surrounding sentences. However, in the classical literature on null complementation and implicit roles—if unexpressed—these particular roles and their associated relations are left out of consideration, because they pertain to relations of modification or adjunction and are thus not part of the *core* arguments in a sentence. Still, as this thesis will show, the recovery of unexpressed non-core roles (esp. on causality) is highly beneficial and leads to the establishment of informative implicit links from the local context to propositional antecedents or postcedents in the global context.

As a matter of fact, implicit information is not only evoked on the local word or phrase level within a sentence. Crucially, it can also hold globally between longer extended descriptions, for example, between complete sentences or even paragraphs. In any well written text, linguistic expressions (clauses, sentences, etc.) are semantically linked and logically cohere by virtue of an underlying discourse structure; cf. Hobbs (1985); Grosz and Sidner (1986); Polanyi (1988); Lascarides and Asher (1993); Webber (2004), inter alia. Even when a particular discourse relationship between two adjacent sentences is not explicitly signaled by a connective (e.g., using the word *because* in a causal relation), a speaker can nonetheless form the sentences in such a way that this interconnection can be easily inferred by the hearer. These **implicit discourse relations** are special insofar as they cannot be analyzed trivially by means of a computer. In fact, in order to explore the realization of different sense types (e.g., causal or temporal), no explicit markers can be consulted and thus various linguistic features need to be considered in the respective discourse units. Computational approaches accounting for implicit discourse relations in free text rely on distinct frameworks of discourse structure, e.g., hierarchically-shaped Rhetorical Structure Theory (Mann and Thompson, 1988), or the shallow Penn Discourse Treebank (Prasad et al., 2008). State-of-the-art techniques follow the concept of neural representation learning for enhanced inference capabilities beyond explicit word content (Ji and Eisenstein, 2014; Ji et al., 2016).

Although sentence-internal and inter-sentential implicit information have for the most part been treated as two separate types of phenomena, there exist a few attempts to assess the effect of their interrelation. The literature in this strand of research studies the characteristics of (potentially unexpressed) words, phrases, or semantic arguments beyond the sentence boundary and is mainly concerned with coreference, anaphoricity, and the choice of referential expressions that a coherent discourse determines, e.g., when an entity is referenced by subsequent mentions in a text, or how its salience is affected when other entities are introduced, cf. Grosz et al. (1995). These models (almost) straightforwardly apply to discourses of, say English or German, but pose serious limitations in the analysis of languages such as Chinese or Japanese which come with the additional complexity of locally unexpressed pronominalization, i.e. *zero anaphora* (Fillmore, 1986; Tao, 1996). Corpus-based studies, as well as computational approaches have been suggested to resolve those entities (e.g., unexpressed core agent roles) in the discourse (Yeh and Chen, 2001; Chen and Ng, 2013; Iida et al., 2007; Chen and Ng, 2016, inter alia). Most notably, Silberer and Frank (2012) apply a special case of coreference/anaphora resolution to the successful resolution of locally uninstantiated items—a technique which emphasizes the mutual dependence of the higher-level discourse structure on local implicit information.

It is fair to say that, in principle, text coherence is formalized by discourse relations and is in turn licensed by implicit semantic roles. This assumption is supported by Givón (1983) and Tao (1996), respectively, who argue that zero anaphora is on the extreme end of a topic continuity scale: when a referent, topic, or subject is mentioned continuously (in subsequent utterances) and is easily recognizable and accessible to a hearer, *"less overt linguistic coding is necessary"*—in the case of zero anaphora, in fact, no coding at all. Pronouns lie somewhere in between, whereas full noun phrases are on the other far end of the scale. They mostly go along with the introduction of a new referent and consequently with topic switches in discourse.

Closely related psycholinguistic experiments study the interpretation of pronouns under specific models of discourse (Kehler and Rohde, 2017). The underlying idea here is that during comprehension, a hearer postulates certain questions (for example *Why?* or *What will happen next?*) that subsequent sentences in the discourse will provide answers to.[3] It has been shown that some words in the lo-

[3] Cf. the so-called *Question Under Discussion* models of discourse interpretation described in

cal (sentence-internal) context can account for an interpretation bias; for example, verbs of implicit causality (Garvey and Caramazza, 1974) such as *frighten* lead to pragmatic inferences made by the hearer and are thus likely to evoke the expectation towards a *particular type* of discourse relation, for example an explanation or a cause, in the global (cross-sentential) context, cf. Rohde and Horton (2010); Asr and Demberg (2012); Hartshorne (2014). The assumption that the realization of subsequent utterances as well as their cohesive links are affected in the discourse (even though these links are not always explicitly signaled as such) is supported by related works focusing on other locally present cues, for instance, negation markers or sentiment polarity (Webber, 2013).

For the purpose of a holistic treatment and the harmonization of both types within a joint setting, in this thesis, I further explore and propose appropriate computational models for the interrelation of implicit information in local and global contexts and their associated relations. Figure 1.1 schematically illustrates this relationship as an orientation for the methods proposed in the ensuing chapters of this thesis.[4]

1.1.2 Significance of the Problem

Taken together, both implicit semantic roles and implicit discourse relations are two distinct—yet interrelated—forms of unexpressed information in free texts, as typically evoked within but also holding between sentences. Both sources contribute substantially to the structure of coherent texts, yet, the non-local resolution of implicit roles, as well as the recovery of implicit discourse senses require more sophisticated techniques than conventional NLP tools, which are, for the most part, restricted to a within-sentence analysis and cannot account for these complex natural language phenomena. Automatically mining implicit information is highly challenging, yet inarguably advantageous, and will therefore be specifically addressed in this thesis. Crucially, downstream NLP applications would greatly benefit from the recovery of unexpressed content: For instance, detecting an implicit purpose relation between sentences would make it possible to answer *why* questions; determining the unexpressed addressee(s) of an imperative statement enables enhanced processing of multi-party conversations; enriching collective information in knowledge graphs with implicit relations between mentions of the same entity will result in advanced inference capabilities and a more intelligent search. I argue that this type of semantic supplementation is essential to effectively approach the true goal of natural language understanding.

Chapter 8.

[4] Note that this illustration is a simplification as most distinctions between the phenomena are far from clear-cut. Also, local contexts do not necessarily need to be complete sentences. Instead, they can be represented by arbitrary expressions such as clauses, spoken utterances, etc.

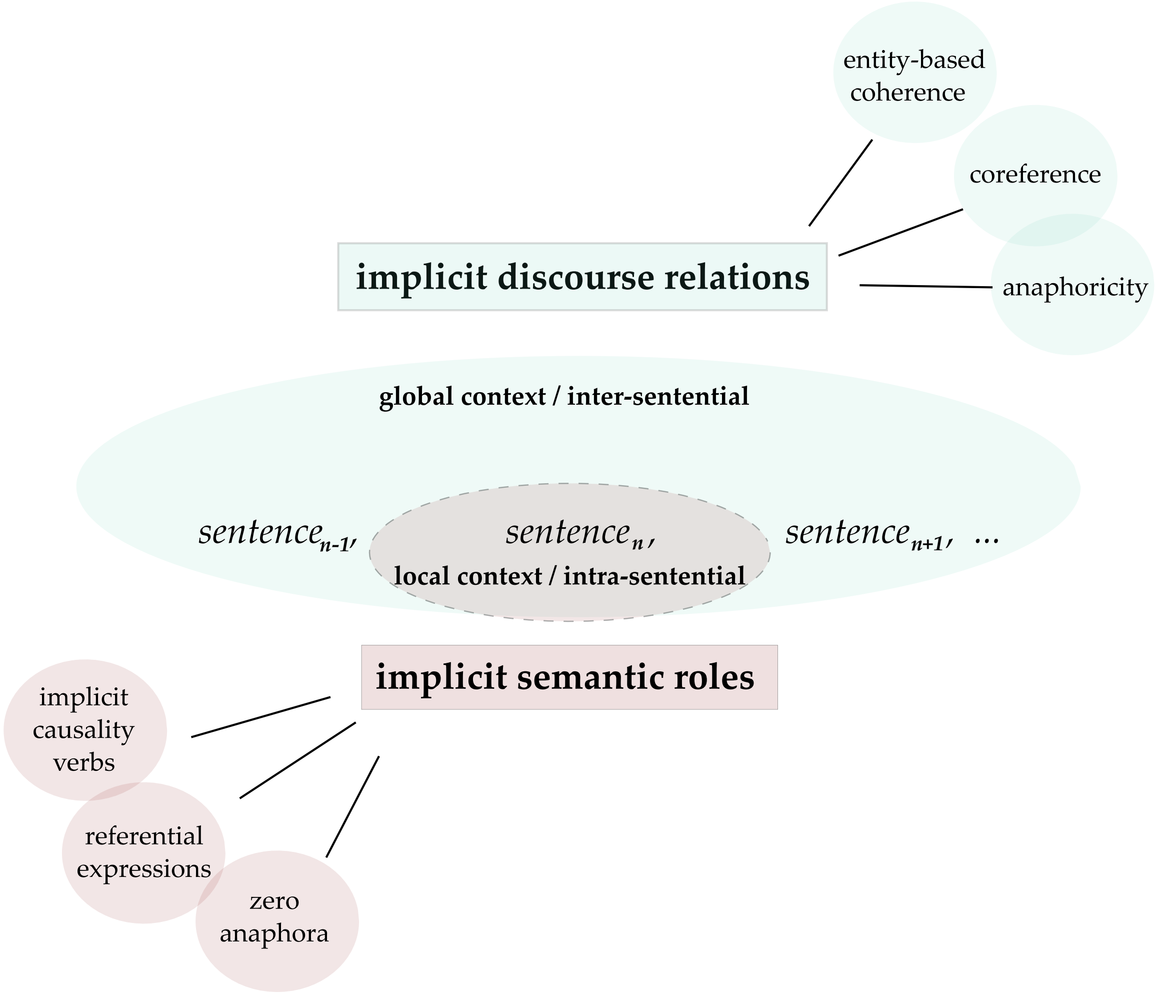


Figure 1.1: Mutual interdependence between local and global implicit information and their associated linguistic phenomena

1.2 An Overview of this Thesis

In this thesis, my main interest lies in developing methods for retrieving implicit information and their relations from unstructured, raw text. To this end, I focus on the two aspects of textually unexpressed phenomena which are evoked within-sentences, i.e. implicit semantic roles, and between sentences, aka. implicit discourse relations.

Implicit discourse parsing is the task of assigning a sense label to the relationship between any logically coherent pair of (non-explicit) discourse units. The presented work starts in this global setting, because most related work in the domain of discourse processing has a computational focus and the parsing task in this chosen framework, as well as the evaluations, are clearly defined.

The technique for the detection of locally uninstantiated arguments and the resolution of appropriate fillers in the context is termed **implicit semantic role labeling**. This task builds on a long tradition of theoretically motivated literature and is, generally speaking, more complex to assess and evaluate, as computational resources (in particular of manually annotated data) are much more scarce and human annotation judgements tend to vary more among individual examples.

Although the techniques proposed in this dissertation are in large part domain and language-independent, I intend to illustrate and evaluate different experiments on two major languages—**English** and **Chinese**: the only two languages, to the best of my knowledge, for which limited amounts of both gold-annotated and computationally sufficiently exploitable implicit relations are available. Until today, English has represented the model language in traditional NLP and computational linguistics, whereas Standard Chinese is globally gaining importance, and features implicit information in the form of zero anaphora, i.e. implicit semantic roles, *to an even greater extent than English*. However, only very recently, efforts have been made to manually construct evaluation resources for English implicit semantic roles and implicit discourse relations (Gerber and Chai, 2010; Prasad et al., 2008), and similarly for Standard Chinese, respectively (Li et al., 2015; Xue et al., 2016). Even though very limited in size, these resources established a basis for training and evaluation of first machine learning systems and serve as direct evaluation criteria for the methods proposed in this dissertation.

1.2.1 The Contributions of this Thesis

Following overviews of both computational discourse and argument structure frameworks, as well as chronological surveys of traditional and state-of-the-art approaches in the two areas, this thesis lays the focus on *technical and implementational practices for modeling implicit information and their relations in text*. It should be noted that all proposed methods are either highly resource-lean, i.e. they do not rely on costly hand-crafted resources, or they infer evidence from statistical generalizations solely obtained from co-occurrence patterns in texts. As a main benefit, the introduced methodology is in large part language-independent and can be easily ported and practically applied to any other domain beyond the ones considered in this thesis. Overall, the systems and presented techniques achieve state-of-the-art or near state-of-the-art performances, either evaluated on official data sets or in independent evaluation frameworks of recent shared tasks. In general, this thesis puts forth the following major contributions and presents innovative

- algorithmic procedures: e.g., for the detection or resolution of implicit arguments.
- modeling concepts: e.g., for improved sequential assessment of discourse structure.
- architectural designs and components: esp. for an efficient representation of implicit discourse structures.
- publicly available parsers, i.e. end-to-end systems from raw text to implicit sense relations.

- resources, e.g., background information obtained from statistical generalizations on predicate-argument structure.

On top of these technical contributions, the work in this thesis aims at finding a principled and coherent explanation for the uniform treatment between both implicit roles and implicit discourse relations. For the purpose of this holistic approach, three theoretically motivated experiments are introduced, including one extension. Different aspects of entity relations and coreferentiality are addressed in the two bridge experiments, and the final extension proposes an algorithm adapted to model entity-based text coherence implemented as a practical downstream application to model narrative text structure. It should be noted that in all three bridge experiments, we directly bring together the related phenomena of entity-based coherence, implicit discourse senses, and (implicit) semantic roles, which makes the studies presented in this thesis distinct from prior research on the topic.

Some of the work illustrated in this thesis has been published previously. A detailed overview of the individual contributions and accompanying papers, partly based on the work with co-authors, is given hereafter. Original publications from which textual descriptions were in parts literally adopted, extended, or improved, and which serve the basis for some of the chapters in this dissertation are indicated in the following. Most of the bridge experiments, in particular the experimental studies described in Chapter 8 and Chapter 9, represent unpublished work.

Implicit Discourse Parsing

A lightweight parser
We describe a structurally lightweight shallow implicit discourse parser based on a feedfoward neural network. It distinguishes itself by a simple network architecture and effective composition of discourse arguments. Our model is resource-lean, benefits from unsupervised pretraining, the integration of syntactic dependencies, and is in large part language-independent. It can be trained quickly, is highly competitive on English data, and ranks second on Chinese implicit discourse relations as evaluated in an official shared task. A detailed description can be found in Part II, Chapter 3; the original publication is Schenk et al. (2016), which we extended, in particular, with the description of the components and the evaluation section.

A state-of-the-art parser for Chinese
We present a recurrent neural network model for the recognition of Chinese implicit discourse relations. Its mode of operation is targeted to analyze discourse relations sequentially, improving upon previous feedforward approaches. The parser's attention mechanism makes possible a thorough inspection of the active features which drive the classification decision. Our approach benefits from a novel partial sampling training technique, and achieves state-of-the-art performance on the Chinese Discourse Treebank. The technique is outlined in Part II,

Chapter 4; the accompanying publication is Rönnqvist et al. (2017), which we extended and revised by including more illustrative examples for the attention weights.

Implicit Semantic Role Labeling

An implicit role detector based on explicit role patterns
We introduce a context-sensitive method to detect whether a particular predicate instance has locally uninstantiated roles. The inference mechanism for implicit roles is grounded on large-scale generalizations of automatically annotated explicit role patterns. Our proposed probabilistic method alleviates the need for rule-based lexicon lookups, is sensitive to distinct word senses for both nominal and verbal predicates, is highly competitive in the recognition of definite null complements, and applies as well to non-resolvable roles. Details can be found in in Part III, Chapter 6; the accompanying publications are Chiarcos and Schenk (2015b) and Schenk et al. (2015), that we revised with motivating examples for the detection of implicit roles.

An unsupervised implicit role resolver based on prototypes
We address the task of implicit role resolution, i.e. linking locally uninstantiated arguments with an appropriate antecedent in the discourse. To this end, predicate and role-specific prototype representations are learned from large-scale, automatically produced annotations, and candidate fillers are determined by distributional similarity. The proposed method learns interpretable patterns, is highly resource-lean, and yet is competitive with supervised systems on two standard evaluation sets across distinct frameworks and parts-of-speech. A detailed description can be found in Part III, Chapter 7; the original publication is Schenk and Chiarcos (2016), which was revised by adding an additional illustration on the creation of protofillers.

Bridge Experiments

An assessment of the effect of implicit roles on implicit discourse
In this quantitative correlation study—the first of its kind—we measure the interrelation between implicit semantic roles and implicit discourse relations. This experiment can be considered a large-scale generalization over prior attempts in the literature, which focused only on a few theoretically-motivated cues. The results are evaluated against implicit gold relations in a discourse treebank, and the insights gained from our experiments demonstrate an expected behavior of a close semantic association between akin local and global information. The study and methodology are outlined in Part IV, Chapter 8.

A discourse-driven semantic role labeler beyond the sentence-level
This experiment aims at setting discourse coherence on the local basis of predicative event structure. We demonstrate that an existent discourse network can serve the additional function of modeling augmented local semantic role patterns by encoding a back-reference to previous utterances. We introduce novel role

labels and distinguish two types of coherence relations, coreferential and non-coreferential ones, both of which can be captured successfully by our method. A visualization of the attention activity of the model unveils that coreferential patterns exhibit similar properties compared to entity-based coherence relations. The architectural design and instance representations are described in Part IV, Chapter 9.

A story coherence model operating on implicit discourse structures
By extension of our lightweight parser, we illustrate how implicit discourse structure can contribute to modeling entity-based coherence in narrative stories. Our adapted system processes consecutive sentences of a coherent story which are semantically linked but whose relationship is generally not signaled by discourse connectives. Our approach achieves competitive performance in a cloze test on automatically predicting appropriate story continuations. Details can be found to in Part II, Chapter 10; the original publication is Schenk and Chiarcos (2017) which we slightly extended.

1.2.2 The Structure of this Thesis

The remainder of this thesis is organized as follows:

Part II is concerned with **implicit discourse parsing**.

Chapter 2 describes the theoretical backbone on which our work is grounded. Besides a description of the historically popular frameworks, Section 2.1.4 places special emphasis on the Penn and Chinese Discourse Treebank (the core resource focused on this thesis) along with an overview of the different annotated sense relations and their proportions in the data set. In Section 2.2, we motivate the need for the analysis of implicit discourse relations and illustrate computational challenges. In particular, we show how a baseline approach functions (Sect. 2.2.1), and describe traditional (feature-rich) approaches in Section 2.2.2. We conclude this introduction with issues related to conventional, rich linguistic features (Sect. 2.2.3.1), and lay the foundations for the ensuing chapters with a summary of recent work on implicit discourse parsing (Sect. 2.2.4) whose methods operate in the novel resource-lean framework.

Chapter 3 introduces the lightweight parser for English and Chinese implicit discourse senses. We elaborate on design principles and its network architecture in Section 3.2 and point out structural differences (argument composition) and key features (incorporation of syntactic dependencies) which distinguishes our system from related works. The system performance in an official shared task is described in Section 3.3.

Chapter 4 describes the Chinese implicit discourse parser. The sequential network model and the partial sampling techniques are part of Section 4.2. State-of-the-art performance and a visualization of its attention activities during classification are illustrated in Section 4.3.

Part III is dedicated to work on **implicit semantic role labeling**.

Chapter 5 first introduces the theoretical motivation on argument structure and semantic roles, followed by a description of well-established computational frameworks (Sect. 5.1). We demonstrate challenges of implicit semantic role labeling (Sect. 5.2), and provide a list of the few hand-crafted resources specifically designed for the task (Sect. 5.2.1). In Section 5.2.2, we report on previous approaches, and highlight key aspects of the respective implementations. Finally, we point out current issues in the way implicit roles are treated (Sect. 5.2.3) and lay the foundations for improvements which we present in the ensuing two chapters.

Chapter 6 deals with implicit role detection. We describe the memory-based method in Section 6.2 and present an evaluation of three distinct experiments to assess its effectiveness on nominal and verbal predicates (Sect. 6.3).

Chapter 7 addresses the resolution of implicit roles. The prototype generation is outlined in Section 7.2, followed by a description of how null instantiations are identified in an unsupervised setting. In an ensuing evaluation (Sect. 7.3), we focus on two standard data sets and demonstrate that our approach is highly competitive with supervised systems.

Part IV of this thesis is concerned with a holistic treatment of both implicit discourse structure and implicit semantic roles, and introduces two **bridge experiments** and an extension.

Chapter 8 motivates the correlation study (Sect. 8.1.1) aimed at assessing the local effect of implicit roles on the superordinate implicit discourse structure in a bottom-up fashion. Along with a collection of illustrating examples, we show how we derive statistical generalizations on local implicit roles (Sect. 8.2), and how we quantitatively compute their contribution to discourse (Sect. 8.3). An ensuing discussion in Section 8.3.4 sets our study into the context of previous works with related observations as well as new findings.

Chapter 9 presents a both linguistically as well as technically motivated study in which we approach the classification of (coreferential) local semantic role patterns by means of a discourse architecture (Sect. 9.2). We demonstrate that our proposed top-down technique is robust enough for the task at hand and shed light on the idiosyncratic properties of entity-based coherence relations (Sect. 9.3).

Chapter 10 describes an extension to our proposed methods. We investigate entity relations in narratives and demonstrate how implicit discourse parsing can benefit the downstream application of modeling story understanding. A comparison of our approach with other implementations based on script learning is given in Section 10.3.

Finally, we reflect on our proposed techniques and reported results, and conclude our work in Part V, Chapter 11 of this thesis.

Part II

Implicit Discourse Parsing

Chapter 2

Theories, Frameworks & Computational Approaches

One of the most promising practices in automated text and speech processing go *beyond* the sentence level. These methods are concerned with the analysis and exploitation of a text's underlying *discourse* properties. In any extended natural language description, it is typically the case that sentences are not simply haphazardly grouped and isolated utterances. Instead—driven by semantic and pragmatic factors—they are logically inter-connected and account for a joint and *coherent* structure of a text.

As a prerequisite to natural language understanding and text comprehension, the proper detection of how meaning units are arranged within discourse can have great benefits for a large number of practical downstream applications. This is a highly challenging task. In the field of Natural Language Processing, these applications include—but are not limited to—text classification (Ji and Smith, 2017), multi-party dialogue processing (Afantenos et al., 2015), spoken dialogue systems (Higashinaka et al., 2003), sentiment analysis (Mukherjee and Bhattacharyya, 2012; Trivedi and Eisenstein, 2013), automated text summarization (Louis et al., 2010; Hirao et al., 2013), natural language inference (Mou et al., 2016), identifying constructiveness in discussions (Kolhatkar and Taboada, 2017), text complexity assessment (Davoodi and Kosseim, 2016), the recognition of textual entailment (Hickl, 2008), question answering (Sun and Chai, 2007; Ferrucci et al., 2010), desire fulfillment modeling in narrative texts (Rahimtoroghi et al., 2017), and various other related fields. Besides these numerous usage scenarios, there have been recent efforts for discourse modeling beyond individual languages on a broader cross-language level.[1]

A few well-established frameworks and formalisms for discourse processing have been proposed in the literature, cf. Hobbs (1985); Grosz and Sidner (1986); Polanyi (1988); Lascarides and Asher (1993); Webber (2004), inter alia. Most of

[1] E.g., by the EU-funded program *Structuring Discourse in Multilingual Europe/TextLink*, `http://textlinkcost.wixsite.com/textlink/`, accessed July 2017. Some of the objectives of the working groups are to assemble, unify, and standardize existing corpora, develop annotation guidelines and tools for discourse phenomena, and to ensure interoperability of multilingual data sources.

the computationally-oriented frameworks, in particular Rhetorical Structure Theory (Mann and Thompson, 1988, RST), model individual, adjacent discourse units as a recursive composition into hierarchical elements, which ultimately represent a text as a tree-shaped pattern. Wolf et al. (2003) question the adequacy of a strictly hierarchical discourse structure in which all textual elements need to be combined as adjacent pieces in order for a text to be coherent. They address the need for more flexible data structures, namely directed graphs in favor of a tree-shaped analysis, which makes their approach especially convenient in capturing long-distance and even crossing dependencies between discourse segments. Both Discourse Representation Theory (Kamp and Reyle, 1993, DRT) and in particular the derived formalism Segmented Discourse Representation Theory (Asher and Lascarides, 2003, SDRT) account for textual discourse in terms of formal semantics allowing for a thorough and precise explanation of important linguistic phenomena (e.g., anaphoricity). Alternative modeling techniques to these stringent logical formalizations exist. For example, the Penn Discourse Treebank (Prasad et al., 2008, PDTB) and the Chinese Discourse Treebank (Zhou and Xue, 2012, CDTB) analytically treat a pair of discourse units in a low-level, "shallow" manner and do not impose a global tree or graph structure on the text, which relaxes several theoretical assumptions and makes them especially suitable for practical implementations.

In the first part of this chapter, we will briefly elaborate on key aspects of different theoretical approaches to modeling discourse structure and specifically address the three aforementioned frameworks of RST, (S)DRT and the PDTB/CDTB in Sections 2.1.2, 2.1.3 and 2.1.4, respectively. With these theoretical foundations as a starting point, we will then quickly move towards applied and computational approaches to modeling discourse information in free text and demonstrate the suitability of PDTB-style relations for practical applications (Section 2.2). We specifically motivate the importance of *implicit* discourse relations. These relations exhibit a latent connection between sentences which is not explicitly signaled by a connective, and thus, makes them highly challenging to deal with. However, as we will see, implicit discourse relations are especially valuable and worth exploring. Drawing on these observations, we review the literature on how previous research has dealt with implicit discourse structure in real implementations. Here, we distinguish three types of computational methods by how strongly they rely on external, hand-crafted data sources: Resource-intensive methods are described in Section 2.2.2 and mildly resource-intensive methods in Section 2.2.3. Finally, the chapter concludes with an outlook of promising generic, resource-lean parsing techniques (Section 2.2.4) which have only recently found their way into automated discourse processing, and which lay the foundations for the implementations described in the ensuing chapters.

2.1 Models of Text Coherence & Discourse Frameworks

2.1.1 Centering Theory

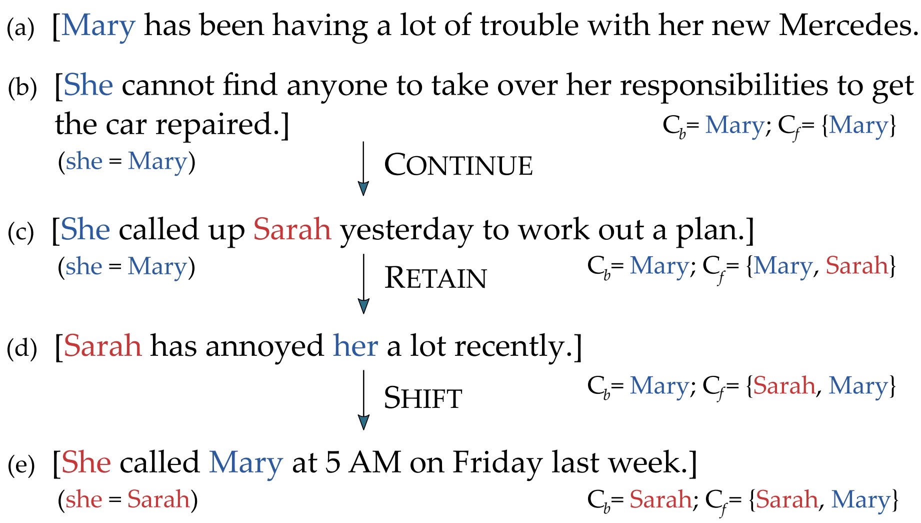
The Centering Theory has been originally proposed by Grosz et al. (1983, 1986) and was further worked out and formalized in Grosz et al. (1995) as an account to model local text coherence and to explain preferences for interpretation. It is an extension and refinement of the early work described in Grosz and Sidner (1986), in which the explanation of the relationship that holds between (sequences of) utterances within a discourse segmented is based on the building blocks of linguistic structure, intentional structure, and the focus of attention. These components in turn influence the use and choice of referring expressions and result in differences in how text coherence is perceived. For example, sometimes a hearer can easily understand and make inferences from a text, in other cases, however, text comprehension can be difficult or even confusing.

Roughly speaking, the *center* of an utterance is the most salient entity, i.e. the discourse referent carrying the focus of attention. Centers can be realized, for instance, by definite descriptions in terms of a noun phrase, or by pronominal reference. Generally, each utterance distinguishes (a single) backward-looking and (a set of) forward-looking centers, denoted as C_b and C_f, respectively. This way, centers serve the purpose of a connection point for linking an utterance with other utterances in the discourse. The focus of attention, i.e. the "aboutness" of an utterance, commonly undergoes changes as the discourse unfolds. Centering Theory models these changes in terms of three transition relations between subsequent utterances: *center continuation* (Continue: same entity still in focus), *center retaining* (Retain: entity still in the center but not as important as before), and *center shift* (Shift: current entity different from previous entity in focus).[2]

The theory imposes restrictions and makes several claims for a text to be coherent, for example, on the realization of centers, or on the transitions between utterances. Regarding the former, coherent texts would prefer repeated pronominal references. In the latter case, sequences of continuations should be given preference over retentions or shifts so as to guarantee *smoother* transitions in general. Crucially, Grosz et al. (1995) pointed out already that these constraints will be particularly important for the implementation of natural language generation systems, and that violations of these rules affect the inference load on the hearer during discourse interpretation.

As an illustration of a center shift and the different types of coherence relations, consider the following (slightly adapted) example from Grosz et al. (1995, p. 217). Figure 2.1 shows an entity-based discourse consisting of a sequence of five utterances. The entities in the centers of C_b and C_f at each stage are shown to the right and the transition relations in between the sentences. Note that in (b) *Mary* is both part of the backward and forward-looking centers. No changes happen (as

[2] Note that Krifka (2006) in addition to plain Shift distinguishes two variants of Smooth-shift and Rough-shift. In both cases the backward-looking centers change, however, in the latter case the backward looking center is not the currently preferred center.



(a) [Mary has been having a lot of trouble with her new Mercedes.]

(b) [She cannot find anyone to take over her responsibilities to get the car repaired.] C_b= Mary; C_f= {Mary}
(she = Mary)

CONTINUE

(c) [She called up Sarah yesterday to work out a plan.]
(she = Mary)

RETAIN C_b= Mary; C_f= {Mary, Sarah}

(d) [Sarah has annoyed her a lot recently.]

SHIFT C_b= Mary; C_f= {Sarah, Mary}

(e) [She called Mary at 5 AM on Friday last week.]
(she = Sarah) C_b= Sarah; C_f= {Sarah, Mary}

Figure 2.1: Illustration of transition types and center shifts.

a result of a Continue transition) until *Sarah* is introduced in (c), who is now part of the forward-looking center. A Retain transition makes *Sarah* the highly ranked element in C_f. As a result of a Shift in (e), she is finally the backward-looking center.

2.1.2 Rhetorical Structure Theory

The framework of Rhetorical Structure Theory (Mann and Thompson, 1988, RST) defines a discourse description of a text in terms of coherence relations. Approximately 30 functional *relation definitions* can be used to model the organization of individual *elementary discourse units* (EDUs), i.e. the minimal building blocks involved, in a *hierarchical* fashion. Through the recursive application of groupings, larger spans are formed from smaller parts and a text is said to be coherent if no gaps are present, all pieces of the text are occupied by their specific function and when all spans are subsumed under a topmost element holding the global discourse structure. The relations and the manner in which discourse units are linked are subject to certain constraints. The constituent spans involved in the composition can either be clauses, sentences, or larger textual units in accordance with the principle of *nuclearity*: Concession relations, for instance, involve a *nucleus* (the role of the core part) and a *satellite* (the contributing, subordinating part). Contrast relations, are multi-nuclear relations, with two or more nuclei of equal importance. More precisely, these two types of relations are related by hypotaxis (subordination) and by parataxis (coordination), respectively (Taboada, 2009). An illustration from Taboada and Mann (2006b) is reproduced in Figure 2.2. The distinct relations are depicted by a curved arrow and two directly connected straight lines, respectively.

It should be noted that the initial goal of RST was to support the creation of practical applications in automated computer-based text summarization and

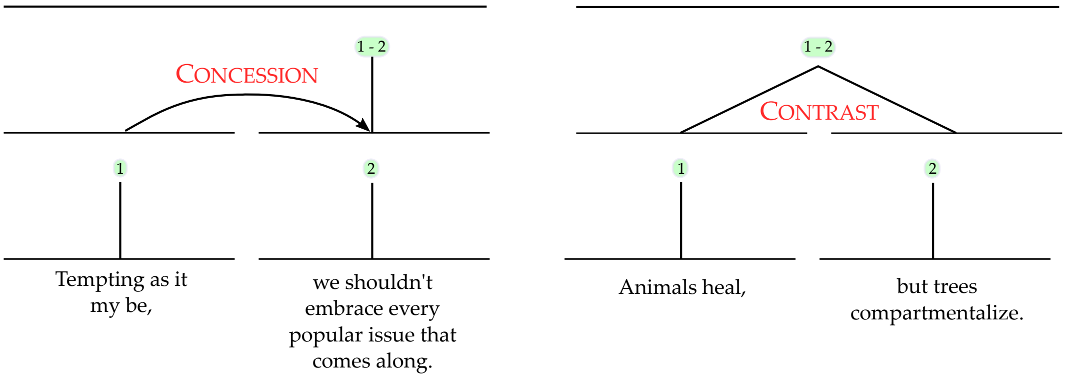

1 - 2
CONCESSION
1
2
Tempting as it my be,
we shouldn't embrace every popular issue that comes along.
1 - 2
CONTRAST
1
2
Animals heal,
but trees compartmentalize.

Figure 2.2: Two example RST discourse relations from the illustration in Taboada and Mann (2006b) with a subordination (left) and a multi-nuclear coordination (right).

generation, and that the framework has been growing in appeal in various related areas (Taboada and Mann, 2006a).

2.1.3 Discourse Representation Theory

An earlier attempt to modeling various linguistic phenomena including, among others, tense, anaphora but also discourse structure, was proposed in Discourse Representation Theory (Kamp, 1981; Kamp and Reyle, 1993; Kamp, 1995, DRT). The conception of DRT is based on formal semantics and for modeling utterances and natural language texts the theory makes use of specific representations, termed *discourse representation structures* (DRS). Put simply, a specific expression is first converted into a standalone DRS and for any subsequent expression, as the context unfolds, information is added to the DRS. This type of *mental* hearer representation is thus constantly updated with new sentences encountered in the discourse, which is a key feature of the theory. The multi-sentence procedure is best illustrated by means of an example.[3]

(1) [x, y: mary(x), mercedes(y), bought(x,y)]

Informally, the DRS in example (1) expresses the following pieces of information. First, that there are two individuals, i.e. two discourse referents (x and y) and that there is a *condition* which states that the former bought the latter. A textual realization of the semantic DRS in (1) would be the sentence *Mary bought a Mercedes*.

Similarly, we could assume an immediately following expression which is directly related to the context of (1) to be represented by another DRS:

[3] The example is a slightly modified version of the illustration from the Stanford Encyclopedia of Philosophy, available at `https://plato.stanford.edu/entries/discourse-representation-theory/`, accessed July 2017.

(2) *She owns it.*
[u, v: owns(u,v)]

The resolution of the anaphoric pronouns *she* and *it* involves the linking to their antecedents by a merging operation. This type of incremental discourse processing[4] results in the final DRS, which encompasses the discourse dependency of both sentences.

(3) [x, y: mary(x), mercedes(y), bought(x,y), owns(x,y)]

DRT in general is highly flexible and allows for a thorough modeling of a great variety linguistic elements and phenomena, including reflexives, plural, presupposition, tense, aspect, and crucially connectives and quantifiers for sub DRS representations which build the connection point to the integration of discourse relations (e.g., binding pronouns across sentence boundaries) in SDRT, a formalism that we sketch in the following.

2.1.3.1 Segmented Discourse Representation Theory

An extension to DRT has been developed and described in Asher and Lascarides (2003); Lascarides and Asher (2007). The semantics/pragmatics interface of Segmented Discourse Representation Theory (SDRT) postulates that in a coherent discourse all subsegments are rhetorically connected. Generally speaking, the logical forms of DRT are equipped with rhetorical relations. The role inventory of discourse relations can, for instance, be borrowed from RST or Wolf and Gibson (2005). An example of a segmented discourse representation structure is given by the discourse graph of Figure 2.3. Note that in SDRT, utterances can be contextually linked to more than one proposition which makes the theory particularly suitable for text and dialogue processing with long-distance relations, as they occur in natural conversations (Ginzburg, 2015).

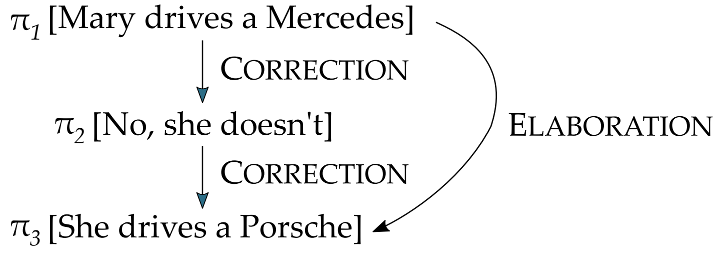

π_1 [Mary drives a Mercedes]
CORRECTION
π_2 [No, she doesn't]
ELABORATION
CORRECTION
π_3 [She drives a Porsche]

Figure 2.3: In SDRT, utterances can be linked to more than one proposition π. Modified illustration from Lascarides and Asher (2007, p. 18).

SDRT provides a mechanism, called *right frontier constraint* (Polanyi, 1985), which takes care of incremental attachments to a given discourse structure and which informally states that new constituents should connect to the rightmost (or else dominating) node in a discourse arrangement. AI-based work on the right frontier constraint has been empirically validated in Afantenos and Asher (2010).

[4] Cf. cross-referential *semantic cohesiveness* (Kamp and Reyle, 1993, p. 59).

2.1.4 The Penn Discourse Treebank & The Chinese Discourse Treebank

The annotation schemes of the Penn Discourse Treebank (Prasad et al., 2008, PDTB) and the Chinese Discourse Treebank (Zhou and Xue, 2012, CDTB) follow a lexically-grounded approach which is centered around discourse connectives. A discourse unit is described as a syntactically motivated character span in the text. It is augmented with relations that point from the second argument (`Arg2`, prototypically, a discourse unit typically associated with an explicit discourse connective) to its antecedent, i.e. , the discourse unit denoted as `Arg1`. Relations between pairs of arguments hold between propositions and are labeled with a relation tag—the *sense*—and the associated predicative discourse marker, either as found in the running text or as inferred by the annotator. Unlike in RST, a discourse unit pair in the PDTB is *flat and shallow*, i.e. it consists of only *two* non-hierarchically linked arguments, which are not annotated with reference to other argument pairs in their immediate context. The PDTB framework can be considered a simplification of RST and (S)DRT in which only local dependencies are considered, yet this assumption makes it considerably easier to annotate running text.

2.1.4.1 Relation Types

PDTB distinguishes *explicit* from *implicit* relations depending on whether such a connective or cue phrase (e.g., *because*) is present. The set of relation types is completed by entity relations (`EntRel`), i.e. entity-based coherence or anaphoric coherence, alternative lexicalization (`AltLex`, discourse marker rephrased), and the absence of any relation (`NoRel`), respectively. For an overview of the type distribution in the PDTB, see Table 2.1.[5]

	`Explicit`	`Implicit`	`EntRel`	`AltLex`
# Instances	14,722	13,156	4,133	524
Proportion (%)	**45.3**	40.4	12.7	1.6

Table 2.1: Distribution of relation types in the PDTB according to Xue et al. (2015)

2.1.4.2 Sense Hierarchy

Sense labels in the PDTB are structured according to a sense hierarchy for explicit and implicit connectives and `AltLex` relations. They were originally introduced in the annotation manuals of the PDTB's first and second version (Prasad et al., 2006,

[5] Statistics are reported in Xue et al. (2015), whose data set provides the basis for all experiments described hereafter.

2007).[6] The top level (class level) has four labels TEMPORAL, CONTINGENCY, COMPARISON, and EXPANSION, which are the most coarse-grained tags. Although different discourse frameworks make different theoretical assumptions, these four senses are roughly shared by most theories in the literature. The sense tag inventory is completed by second level *types* and third level elements (*subtypes*). Figure 2.4 illustrates a slightly modified (but computationally more convenient) sense tag hierarchy in which pragmatic sense tags were removed, and infrequently occurring and semantically closely related tags (especially subtypes) have recently been merged for the purpose of the first shared task on shallow discourse parsing (Xue et al., 2015).[7] The illustration includes statistics for all relation types (i.e. sense information for `Explicit`, `Implicit`, `EntRel` and `AltLex` relations), with EXPANSION:Conjunction being the most frequent sense in the PDTB.[8]

In what follows, we briefly elaborate on the three major relation types.

2.1.4.3 Explicit Relations

Consider the following example of an explicit discourse relation in the PDTB.

(4) `Arg1`: IBM might increase the size of the offering to as much as $1 billion
Connective: <u>if</u>
`Arg2`: investor demand is strong

Explicit discourse relation[9] / sense: CONTINGENCY:Condition

In this explicit relation, `Arg1` and `Arg2` are directly related by the connective *if*. The relation type is CONTINGENCY:Condition that marks the sense relation between the given argument pair.

Note that with approximately 60% of all explicit discourse relations in the PDTB the arguments of the relation appear in the same sentence (like in the example above), whereas in 40% of the cases `Arg1` precedes `Arg2` with arguments being realized in distinct but (almost always) adjacent sentences. In the latter case, connectives typically start the second sentence.

2.1.4.4 Implicit Relations

As an illustrative example without such a marker, consider the following two adjacent sentences from the PDTB in (5).

[6] `https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf`

[7] Absolute frequencies, proportions and sense label information are reproduced in the data set and were originally reported in the accompanying blog by Te Rutherford from `http://conll15st.blogspot.de/2015/02/the-conll-version-of-penn-discourse.html`, accessed July 2017.

[8] It should be noted that the `EntRel` relation type has also the same sense name `EntRel`.

[9] PDTB Document ID `wsj_0351`

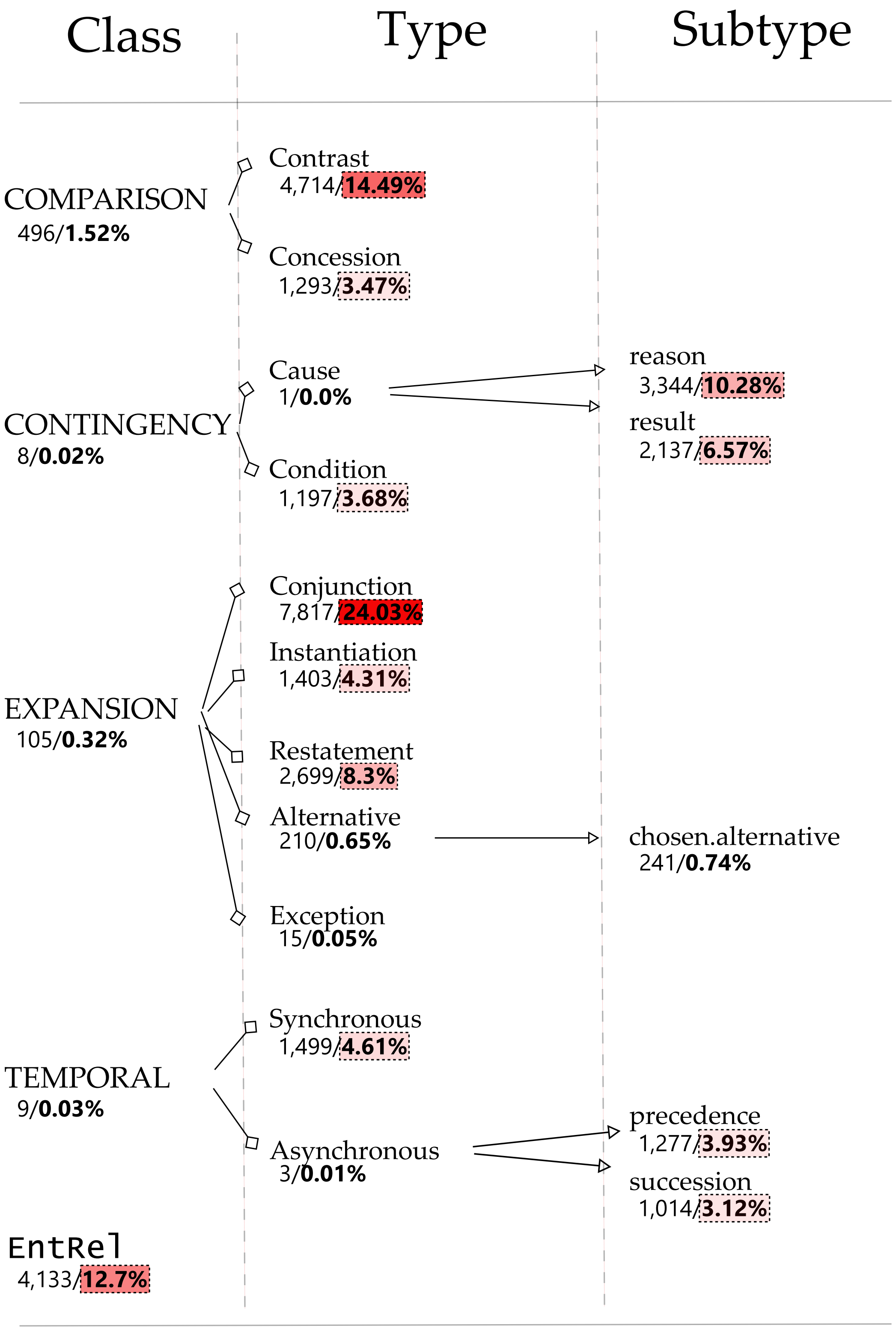


Figure 2.4: Modified PDTB sense hierarchy according to Xue et al. (2015) with merged labels and frequency statistics from all relation types. Proportions are highlighted in red with darker intensities indicating more frequent senses.

(5) `Arg1`: Retail investors nervously sold stock Friday and never returned to bargain-hunt
Connective: –
`Arg2`: Institutional investors were calmer

Implicit discourse relation[10] / sense: COMPARISON:Contrast
Inferred connective: by contrast.

In this implicit relation, `Arg1` and `Arg2` are directly related via the discourse relation COMPARISON:Contrast. Again, note that this time a cue phrase (connective) is not present but only inferred by the annotators of the PDTB. It is supposed to best characterize the underlying sense relation.

Also note that the distribution of argument spans for implicit relations differs greatly from the explicit counterparts. In almost 97% of all cases, `Arg1` precedes `Arg2` across sentences boundaries. In only 3%, both arguments appear in the same sentence.

2.1.4.5 Explicit vs. Implicit Relations

Explicit and implicit discourse relations as annotated in the PDTB differ to a large extent in their underlying sense distribution.[11] Figure 2.5 shows a sample of six contrastive senses. For instance, CONTINGENCY relations of type "Cause" with subtypes "reason" and "result" are more likely to be expressed by implicit discourse relations, i.e. a marker such as *because* or *the reason is* is not present to express these types of relations. The same holds true for EXPANSION:Restatement relations for which it seems natural to not use an explicit connective. On the other hand, temporal relations, e.g. TEMPORAL:Synchrony (*while*) are not as easy to indicate implicitly as opposed to the number of relations which do carry an explicit marker (1.6% vs. 8.6%).[12]

2.1.4.6 Entity Relations

A final example illustrates entity-based coherence relations. Examples (6) and (7) illustrate three consecutive sentences from the PDTB. An `EntRel` sense relation holds between the first two (anaphoric *it*) and the last two discourse arguments (involving a coreferent and shared entity mention of *deficit*). Note that—unlike in explicit discourse relations—there are no connectives involved; that is why according to the PDTB annotation scheme, `EntRels` are variants of implicit relations.

[10] PDTB Document ID `wsj_2379`

[11] Cf. closely related experiments on predicting the presence of a connective in Patterson and Kehler (2013).

[12] In his aforementioned blog, Te Rutherford further elaborates on the difficulty of communicating CONTINGENCY:Condition senses of the form *if, then* and also COMPARISON:Concession relations (*although*) without using explicit markers.

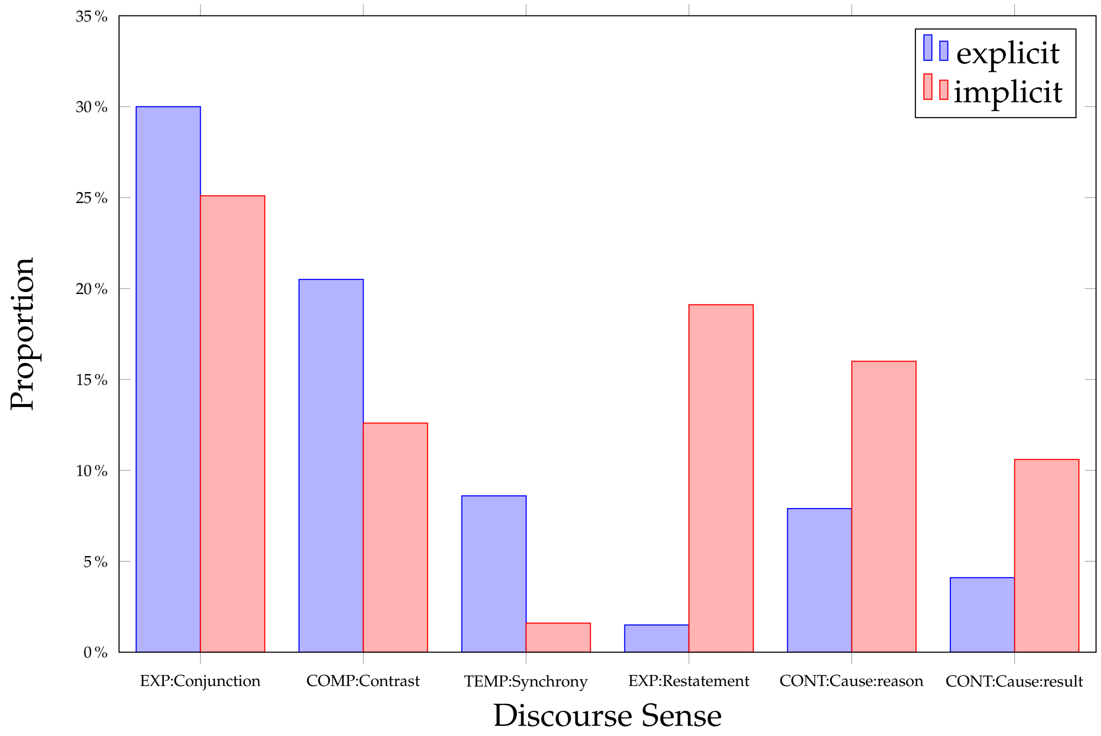


Figure 2.5: A sample of six most contrastive senses from the PDTB with opposing distributions for explicit and implicit relations

(6) `Arg1`: The fiscal 1989 budget deficit figure came out Friday
Connective: –
`Arg2`: It was down a little

`EntRel` discourse relation[13]

(7) `Arg1`: It was down a little
Connective: –
`Arg2`: The next time you hear a Member of Congress moan about the deficit, consider what Congress did Friday

`EntRel` discourse relation[14]

2.1.4.7 Senses in the Chinese Discourse Treebank

The Chinese Discourse Treebank (Zhou and Xue, 2012) closely follows the groundwork and annotation scheme of the PDTB and has contains approximately 73k

[13] PDTB Document ID `wsj_0623`
[14] PDTB Document ID `wsj_0623`

annotated words with 5.5k instances in its version 5.0 (Zhou et al., 2014). The relation type distribution of an augmented version of the CDTB for the purpose of the second shared task on shallow discourse parsing (Xue et al., 2016) is shown in Table 2.2. Note that, compared to the PDTB (cf. Table 2.1), the proportion of implicit discourse relations is much larger (here almost two-thirds of all relations).[15]

	`Explicit`	`Implicit`	`EntRel`	`AltLex`
# Instances	2,398	7,238	1,219	223
Proportion (%)	21.6	**65.3**	11.0	0.2

Table 2.2: Distribution of relation types in the CDTB according to Xue et al. (2016)

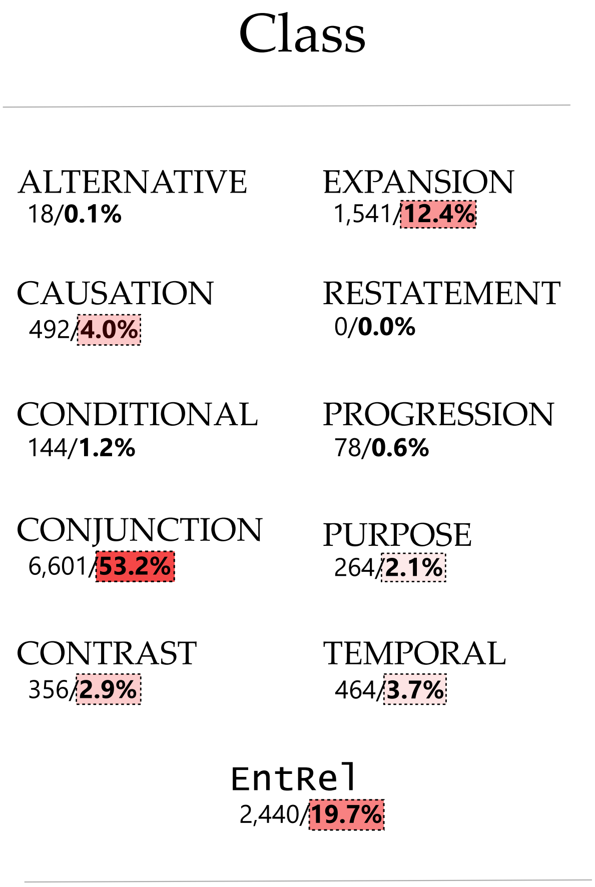


Figure 2.6: CDTB sense labels according to Xue et al. (2015) with frequency statistics from all relation types. Proportions are highlighted in red with darker intensities indicating more frequent senses.

The sense inventory of the Chinese Discourse Treebank follows a flat structure (with no types and subtypes) and contains the sense tags shown in Figure 2.6, with the label distribution according to the shared task data from Xue et al. (2016).[16]

[15] For a detailed analysis of implicit discourse relations and their sense distribution, see Chapter 4, and specifically 4.3.

[16] Note that the originally introduced sense tags RESTATEMENT and PROGRESSION have re-

2.2 Automatic Discourse Analysis & The Challenge of Implicit Relations

The large-scale annotation efforts of the Penn and RST discourse treebanks have quickly initiated the development of automated parsers for the multifaceted problem of discourse parsing, as the manually annotated data has laid the foundations for supervised machine learning for that task. While the number of formal semantics-based implementations of (S)DRT and related theories have been in the minority, more and more full-fledged *end-to-end systems* in the style of RST and PDTB have been realized (Ji and Eisenstein, 2014; Lin et al., 2014; Stepanov et al., 2015; Wang and Lan, 2016). As a minimal requirement, such an end-to-end discourse parser consists of a pipeline with three modular components for

1. Argument (EDU) extraction (shallow/PDTB or hierarchically/RST organized)
2. Relation type detection (implicit, explicit, etc.)
3. Discourse sense classification (e.g., Contrast)

Regarding the first subtask of argument (span) identification, different techniques have been suggested ranging from sequence labeling methods (Ghosh et al., 2011) to constituent-based approaches (Kong et al., 2014). Once arguments have been extracted, in the flat and shallow modeling framework of the PDTB, the determination of the first and second argument, respectively, as well as the relation type classification, is relatively straightforward and dependent on the discourse connective. For RST-style discourse parsing more sophisticated methods are necessary to handle the hierarchical organization and ordering of the elementary discourse units (Hernault et al., 2010b). For the last component, the classification of discourse senses, it has also been shown that, underpinned by the corpus-based observations in Pitler et al. (2008), automated systems can achieve close to human-level performance on *explicit* senses when syntactic features are incorporated (Pitler and Nenkova, 2009). The main bottleneck, however, to any end-to-end discourse parser is the thorough treatment of *implicit* discourse relations.

The identification of the correct implicit discourse sense poses a serious challenge to any automated discourse parser. Here, state-of-the-art performances (ranging between 40-45%, which varies across evaluation and label sets) are not even half as good as for explicit relations. As a consequence, research in the field has paid special attention to these relations without discourse connectives, as potential for improvement is evident. It should be noted that, first, detecting the correct sense is an intricate problem and far from trivial in the absence of an explicit connective given only the occurrences of the bare words in the two arguments. Second, implicit relations are especially worth mining, as they make up the majority of all argument pairs in the PDTB, the *vast* majority in the CDTB,

cently been dropped due to scarcity issues as described in `http://www.cs.brandeis.edu/~clp/cdtb/discourseAnnotationGuidelines.pdf`. Also note that `NoRel` is not part of the shared task data set.

and hence, in most other natural language descriptions. However, only when a discourse parser can reliably recognize implicit senses in a coherent text can it be considered a practical and useful system.

Following the majority of methodological concepts of prior research, the main focus of the work in this dissertation is dedicated to sense recognition for implicit discourse relations, *assuming that the discourse units, i.e. the arguments have already been provided*.[17] As argument span detection (especially for implicit relations) is in principle a solved problem, this specific setup allows us to properly and exclusively focus on the thorough modeling of the underlying sense relations that hold between the postulated discourse spans. As a consequence, this should yield insights into the idiosyncratic properties of these relations without suffering from error propagation introduced by span detection. We chose the PDTB framework for all our following experiments, as its assumption of a facile, shallow discourse structure provides an application-related and computationally attractive basis for relation modeling. Structural issues related to some of the more complex RST-relations, cf. Knott et al. (2001), should thus be avoided. This type of elementary modeling and all related results presented in the next chapters can be considered substantial groundwork for more sophisticated, deeper concepts of discourse parsing.

The remainder of this chapter guides the reader through a roughly chronologically ordered body of literature on the topic of implicit sense classification. Two contrasting methods are presented which we term *resource-intensive* (Section 2.2.2) and *resource-lean* (Section 2.2.4), respectively. Pros and cons of both approaches are outlined as a prerequisite for more sophisticated methods in the two ensuing chapters. Starting with resource-intensive methods, their utilization is best motivated and illustrated by means of a low-generalization baseline approach involving *word pairs*.

2.2.1 Word Pairs—A Baseline Approach

In order to infer that a COMPARISON:Contrast relationship holds between the two arguments in Example (5), a straightforward approach to acquiring features for a supervised machine learning setup would be to first compute the set of all (normalized) *word pairs* from both arguments; this is a list of all pointwise combinations, i.e. the unigram cross-product (×) of tokens in `Arg1` and `Arg2`. As the plain words initially represent the only source of overtly available *lexical* information in a discourse unit, they are generally used as a starting point for advanced modeling. Figure 2.7 illustrates the process of deriving word pairs for the first token *Retail* in `Arg1` with its (five) token combinations in the second argument. The result set contains the word pair elements:
{*Retail–Institutional*, *Retail–investors*, *Retail–were*, *Retail–calmer*, *Retail–.*,...}.[18]

[17] This method is termed *sense-only* classification (Xue et al., 2016) with gold arguments.

[18] Note that tokens in the example are not (lower-case) normalized and that the sentence-final period is also involved in the combination.

As word pairs are typically collected in sets their order of occurrence in both arguments is discarded. Also note that the number of possible word pairs is quadratic in the size of the vocabulary, which makes these categorical and symbolic features highly *sparse* (i.e. #dimensions in the feature matrix >> #instances).

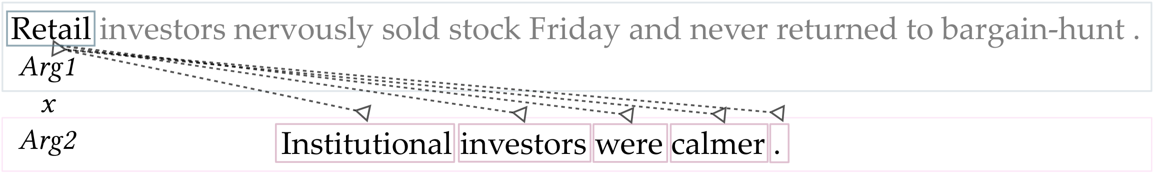

Retail investors nervously sold stock Friday and never returned to bargain-hunt .
Arg1
x
Arg2
Institutional investors were calmer .

Figure 2.7: Word pair combinations based on the first token in the first argument

Marcu and Echihabi (2002): One of the first general attempts to model implicit discourse relations involving word pair features was made by Marcu and Echihabi (2002). The authors present an *unsupervised* approach in which synthetic training examples for implicit discourse relations are generated by first extracting unambiguous explicit relation patterns (including a cue phrase) and then dropping the connective token. A classifier is trained in a Bayesian framework to learn which word pairs are most indicative of a certain discourse relation. The authors demonstrate that learning from larger amounts of unlabeled data can outperform a given baseline by a large margin in their custom setup, which prompted subsequent research to investigate the phenomenon behind word pairs and synthetic training instances in closer detail.

The work in Marcu and Echihabi (2002) has been extended in various directions, for instance to phrasal patterns (Saito et al., 2006), by optimizing the parameter settings, introducing topic segmentation and syntactic heuristics (Blair-Goldensohn et al., 2007), or in order to bootstrap a rhetorical relation classifier on automatically labeled examples (Sporleder and Lascarides, 2008).

Intuitively, some word pairs in the great quantity of potential features might be more powerful in describing a given sense than others, for instance, because they are formed by content words instead of function words or punctuation symbols. In the example, the word pairs *Retail-Institutional* or the positive-negative contrast pattern *nervously-calmer* might provide stronger evidence for the correct contrast sense; others such as *and-were* are probably less likely to be helpful. In addition to these plain surface-level indicators, previous research has proposed the incorporation of *external* resources to help with the identification of the correct sense, e.g., lexicons of sentiment polarities or knowledge bases containing unsupervised word representations. These are typically consulted to bolster the relative contribution of each of the features with respect to the discourse sense by generalization from individual tokens to more *universal* linguistic categories. Figure 2.8 highlights two especially important word combinations for the given argument pair, which are semantically salient word pairs with underlying contrastive properties.

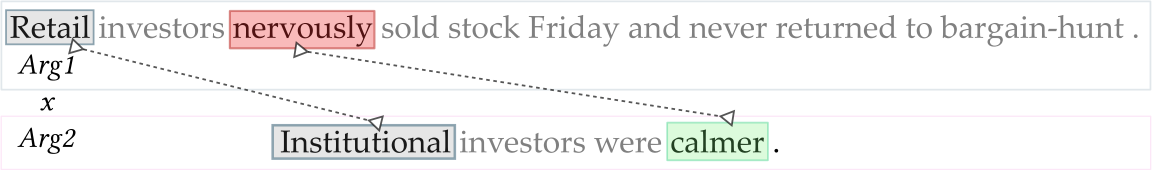

Retail investors nervously sold stock Friday and never returned to bargain-hunt .
Arg1
x
Arg2
Institutional investors were calmer .

Figure 2.8: Illustration of word pairs supported by rich linguistic resources. The word pair *Retail-Institutional* might be supported by an ontology of financial terms, and *nervously-calmer* by a sentiment lexicon.

It should be noted that word pairs per se are not resource-intensive, as they can be generated straightforwardly from the arguments of an implicit discourse relation. The linguistic extensions which build on top, however, are indeed resource-intensive, as lexicons and knowledge bases are typically hand-crafted and costly. Ensuing approaches have at a later stage tried to incorporate add-ons in the form of more abstract word representations for the sparse word pairs and as a (partial) substitute for the rich linguistic features which both come at a cost. We term this interim stage *mildly resource-intensive* which has finally laid the foundations for completely knowledge-free, i.e. *resource-lean* parsing. The overall paradigm can thus best be summarized as follows:

1. **word pairs** baseline (§ 2.2.1)
   → *sparsity issues*
2. **resource-intensive**, rich linguistic features (§ 2.2.2)
   → *cost and flexibility issues*
3. **mildly resource-intensive** methods, incl. abstract representations (§ 2.2.3)
   → *prestage to representation learning*
4. **resource-lean** parsing (§ 2.2.4)

In what follows, we elaborate on a number of resource-intensive and mildly resource-intensive strategies by (a non-exhaustive list of) selected publications from the field of discourse processing. In the light of the motivating example of word pair representations, we discuss pros and cons of various approaches with *rich linguistic resources as well as sparse features* and inspect in closer detail the appropriateness of more *general representations* for implicit relation classification.

2.2.2 Resource-Intensive Implicit Discourse Parsing

Traditional approaches to implicit sense classification are *resource-intensive* and involve careful engineering of *rich linguistic features*. Classic examples are suggested in Huang and Chen (2011) for Chinese discourse relations, by Subba and Di Eugenio (2009) for the PDTB or by Feng and Hirst (2012) for the framework of RST. Given a pair of argument spans of an implicit discourse relation, the prevailing methodology is to integrate additional, external tools and resources (e.g., part-of-speech taggers, dictionaries and knowledge bases) to support the identification of the correct sense.

Subba and Di Eugenio (2009): The methodology in Subba and Di Eugenio (2009) is highly linguistically-motivated. The authors present a feature-rich approach to relational learning from first-order-logic representations. Subba and Di Eugenio (2009) introduce a shift-reduce parser along with a whole repertoire of linguistically-informed resources: compositional semantic information stems from VerbNet (Schuler, 2005), for instance, and the classification of rhetorical relations is guided by a background knowledge base for rule learning. The authors employ part-of-speech tags, linguistic cues, and WordNet (Miller, 1995) information, among others. The easy interpretability of the deduced rules that the inductive model learns is an advantage of their approach.

Pitler et al. (2009): The first study on supervised machine learning for implicit relations from the PDTB was pursued by Pitler et al. (2009). The authors motivate the use of "higher-level" features by first showing the downside of the low-level word pairs involved in prior attempts to modeling implicit discourse relations, for instance, by demonstration of an anomalous effect in the generated synthetic training data described in Marcu and Echihabi (2002). Illustrations in Pitler et al. (2009) reveal that the most distinctive word pairs are of a functional type and—contrary to what would have been expected—do not bear any semantic content. As a consequence, Pitler et al. (2009) introduce a repertoire of linguistically informed features, including, among others, polarity (Wilson et al., 2005) and General Inquirer tags (Stone and Hunt, 1963), Levin verb classes (Levin, 1993), and modality. The set of features is completed by additional ones which are closer to the textual surface level, for example, numerical and temporal expressions, contextual features indicating the presence of a paragraph boundary, language model probabilities for sequences of tokens obtained from implicit relations in the PDTB training set, the first, last and first three words of an argument, or the average length of a verb phrase. The results indicate that word polarity and lexical information are highly indicative of two of the four top level classes, but no single strong effect stands out within their range of diverse features.

Lin et al. (2009): The first work on second-level relation type classification for implicit discourse senses on the PDTB is described in Lin et al. (2009). Their rich feature set builds on three main pillars: contextual features, manifested in discourse dependency structure between arguments in various constellations (e.g., embedded and shared representations), syntactic features (phrase structure and dependency paths in the form of production rules of trees and tree fragments), and, finally, lexical features expressed by word pairs. An ablation study reveals that production rules and word pairs contribute most in the classification scenario. In their discussion, Lin et al. (2009) shed light on the inherent difficulty of modeling shallow implicit discourse relations in the PDTB, and motivate the need for deeper semantic representations, involving inference capabilities based on knowledge bases, general world knowledge, and yet additional context beyond the standard span-based arguments of the PDTB.

Feng and Hirst (2012): Although their study is not explicitly restricted to implicit discourse relations, Feng and Hirst (2012) promote the use of rich linguistic features for RST-style discourse parsing. Their text-level parser is grounded on the previous achievements of Hernault et al. (2010b) and Lin et al. (2009), and their approach supplements the overall feature set with linguistic information which includes, for instance, syntactic tags, production rules, lexical heads of phrase structure trees, contextual features (sequential patterns of discourse units), novel discourse production rules, and semantic similarities derived from VerbNet (Schuler, 2005). In both subtasks of (within-sentence and cross-sentence) structure and relation classification, the methodology presented in Feng and Hirst (2012) is superior to prior works and the authors demonstrate that contextual features are particularly relevant for the task at hand.

2.2.3 Mildly Resource-Intensive Implicit Discourse Parsing

The previously described first attempts to model implicit discourse relations to target a higher level of abstraction put a main emphasis on linguistic background knowledge. On the one hand, this kind of resource-intensive modeling has been well-established for a long time and proven to be adequate and particularly convenient—also in other NLP contexts. In a supervised machine learning setup, when linguistic features are carefully designed and meaningful, they enable engineers to precisely interpret their effect on a specific task, for instance to estimate their predictive impact on the performance of a classifier. Turning certain language cues on and off in ablation studies can provide valuable insights and make possible the linguistic *interpretability* of the results. Instead of having to deal with a "black-box" predictor, manually crafted linguistic features allow for a better understanding of the process behind implicit relation modeling, for example by means of contrasting polarity tags in a COMPARISON relationship.

On the other hand, there are several issues related to high resource intensity: First, these specific features are task-tailored and need to be thoroughly adapted towards the data at hand and most of the time extensive and time-consuming manual feature engineering is required. Rich linguistic features are very costly, both in terms of their acquisition and their creation. Especially the sentiment lexicons, word nets or knowledge bases are assumed to be present for the specific task but, unfortunately, these resources are not available to the same extent for most genres and languages, which makes them highly domain- and language-specific and restricts discourse analysis unproductively to only a subset of available texts.

This section therefore describes a paradigmatic change in more recent approaches to modeling implicit discourse relations. They strive for a higher degree of abstraction by means of external knowledge representations, i.e. abstract word representations as a substitute for the sparse word pairs, which can be acquired more easily. However, these techniques do not fully abandon the use of (rich) linguistic information, which is why we call them *mildly resource-intensive*.

Rutherford and Xue (2014, 2015): A powerful knowledge base integration in the form of data-driven background information has been proposed by Rutherford and Xue (2014). The authors have employed pre-trained Brown clusters (Brown et al., 1992) as an alternative for the sparse word tokens in order to obtain a higher degree of generalization. Specifically, each word in an argument has been replaced by a hard Brown cluster assignment (Turian et al., 2010). The resulting feature set has the advantage of being much smaller (given only a fixed number of Brown clusters) compared to the original word pairs whose theoretical upper bound is dependent on the vocabulary size. According to the distributional hypothesis, tokens classified into the same cluster share linguistic properties. This way, named entities, for instance, can be easily encapsulated within one class. Given this type of semantic generalization over sparse surface features, Rutherford and Xue (2014) demonstrate an additional gain in performance on the PDTB class-level predictions and show that Brown clusters represent the most predictive features in the supervised setting.[19] Beyond that, the results provide valuable interpretative insights: For instance, word pair interactions whose tokens are from the *same* cluster are indicative of COMPARISON relations, while semantically related words (potentially with a shift from general to specific) are more likely to inform CONTINGENCY senses.

On top of the semantic word classes, the authors experiment with a number of coreferential features between arguments. These are largely linguistically-informed by thorough observation of the sense relations and include, for instance, the number of coreferential pairs between arguments, similar nouns, subjects and verbal predicates assigned to the same Brown cluster. Rutherford and Xue (2014) show that temporal relations carry most of the coreferential information, which obviates the need for explicitly using a discourse connective along with these particular relations.

Rutherford and Xue (2015) improve upon their work in a follow-up publication with a new inspiring technique: The authors introduce a distant supervision approach to obtain additional training data for implicit discourse relations. Based on a large number of explicit relation pairs heuristically extracted from Gigaword (Graff and Cieri, 2003), the authors introduce two selection criteria (omission rate and context differential) to assess the *optionality* of an explicit discourse connective. Dropping it should not change the underlying orientation of the discourse relation. In that case the relation could serve as a distant supervision signal supplementing the scarce hand-annotated resources of the PDTB. In fact, discourse connectives such as *because* (having high omission rates, and low context differentials) tend to be highly suitable for generating additional training instances used to reinforce implicit relation classification. The idea of semi-supervised learning using a mixture of labeled and unlabeled data has also inspired subsequent research; cf. Fisher and Simmons (2015) for a more sophisticated approach involving spectral optimization for implicit discourse relations.

Braud and Denis (2015, 2016): A refinement of the work by Rutherford and Xue (2014) is presented in Braud and Denis (2015). Their work is one of the first

[19] It is noteworthy that both Rutherford and Xue (2014) and Pitler et al. (2009) report that Naive Bayes classification performed best among all settings.

fruitful attempts to employ only those types of resources which can be easily obtained in an unsupervised manner from large amounts of unannotated data, which makes their approach particularly attractive in terms of high flexibility and language-independence. The authors compare various unsupervised word representations suitable for implicit sense classification in the PDTB. In detail, Braud and Denis (2015) extend the idea of using only the Brown clusters as a proper substitute for sparse word (pair) tokens within arguments and inspect additional ways and combinations to derive argument pairs by using vectorial variants of, e.g., one-hot encodings, verbal head-word patterns, low-dimensional representations and (dense) word embeddings along with various composition functions. Their major findings are that dense representations perform better than raw tokens, that—contrary to prior conclusions—shallow lexical features are indeed helpful for the task at hand, and that adding traditional rich linguistic features from prior publications to the set of unsupervised word representations can further improve binary classifier performance for each of the four top-level classes by a small margin.

In a follow-up publication, Braud and Denis (2016) extend their work by a novel, semi-supervised approach to obtain unsupervised word representations. Their method learns statistics of word-connective co-occurrences in the two arguments of *explicit* discourse relations as found in large amounts of automatically annotated texts which are then applied to the recognition of implicit relations. These (word and argument-order-agnostic) low-dimensional vectors are particularly effective in a four-sense multi-class classification setting[20] and seem to gain predictive power by increasing the number of connectives encoded in them. This is particularly interesting as Braud and Denis (2016) point out that other textual cues—besides connectives—e.g., the phrase *one reason is*, could easily be integrated into their learned representation.

Chiarcos and Schenk (2015): In the context of the first edition of the CoNLL shared task on shallow discourse parsing (Xue et al., 2015), we have conducted experiments in a minimalist setting using a lightweight classifier for implicit discourse senses that we trained on the PDTB. The approach is described in Chiarcos and Schenk (2015a) and serves as a supplement to the various feature optimization techniques for the task by providing a closer view on the specific role of lexical features, in particular for the feature-intensive word pairs and their characteristic properties.

In this lightweight setup, we have restricted the label set to the six most frequent implicit discourse senses in order to obtain a reasonable degree of generalization, and excluded entity-based coherence relations. We trained several SVM models[21] (on argument pairs of implicit discourse relations only) and tested

[20] Note that this is different from four one-vs.-all binary classifiers as employed in Braud and Denis (2015).

[21] In all our experiments with SVMs (Cortes and Vapnik, 1995), we employed the *libsvm* implementation (Chang and Lin, 2011) with linear kernel and default parameters. Punctuation symbols were removed and all features were treated as boolean based on their presence (true) or absence

different word pair feature sets, as well as more abstract representations: As a substitute for the word forms we used stems, embeddings, word vectors, or a combination of them. The feature sets are summarized as follows:

1. Word pairs of `Arg1`–`Arg2` (*WP-tokens*):
   (i) normal-case (N) as encountered in the running text and
   (ii) after lower-case normalization (l), both with frequency thresholds applied.
2. Analogous to (1.) but using word stems (Porter, 1980) instead (*WP-stems*).
3. Analogous to (1.) but using a Brown cluster 3200 representation (Turian et al., 2010) as a substitute for each word form. In case it does not exist, we used the word form as feature (*WP-BC3200*).

In a subsequent experiment, we aimed at finding a more compact representation of an argument pair (denoted as *WordVecs*). To this end, for each `Arg1`–`Arg2` pair, we computed two real-valued 300-dimensional feature vectors (one for each argument). These vectors were obtained by summing over all skip-gram neural word embeddings (Mikolov et al., 2013a) present in each argument weighted by the number of embeddings found in each argument. This normalization makes it possible to compare sentences of different lengths. In a final experiment, we combined both Brown clusters and neural word embeddings into one feature set for each argument pair of an implicit discourse relation (*WP-BC3200+WordVecs*).

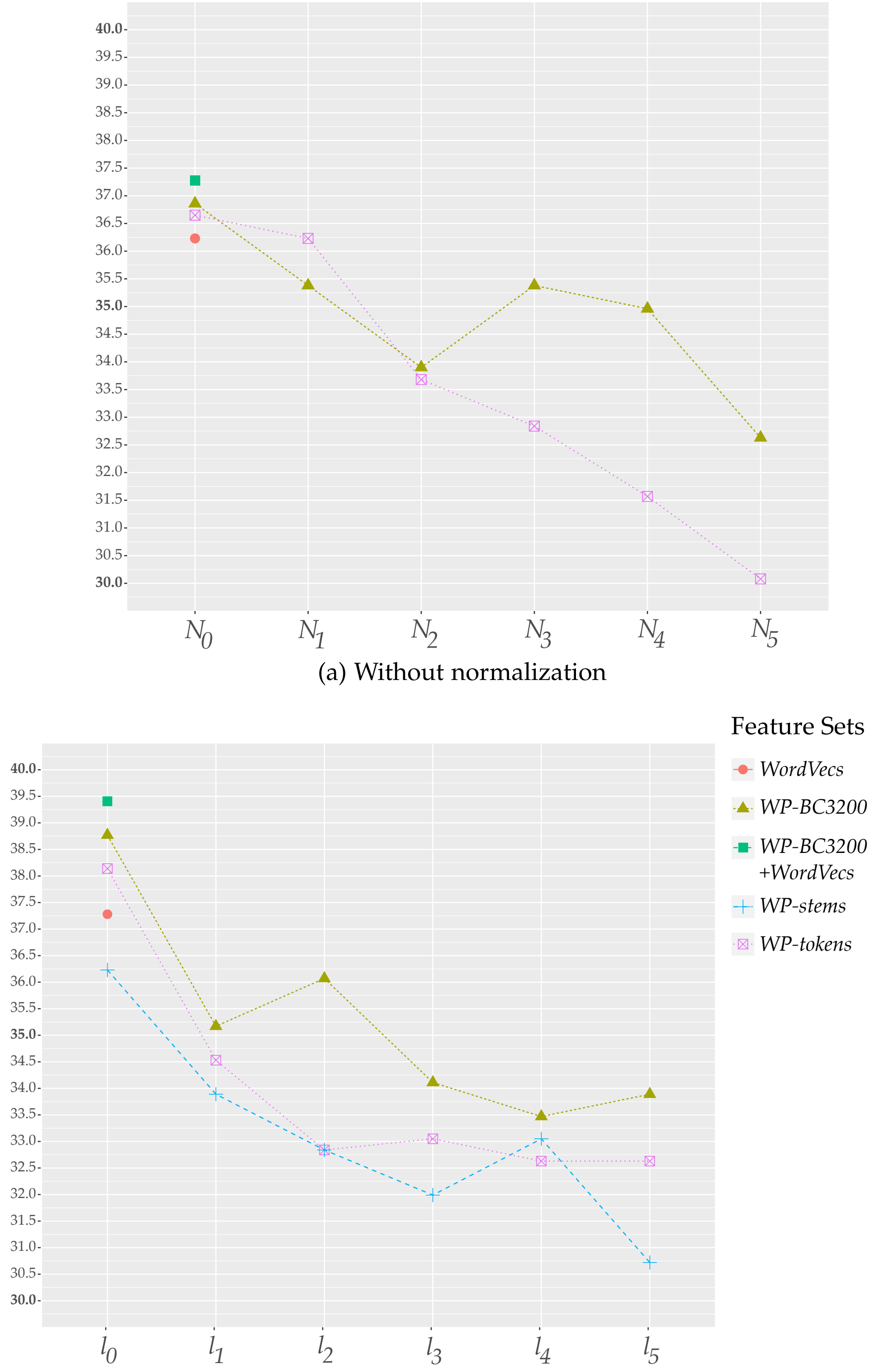
The results for implicit sense classification (472 implicit sense relations in total) based on the proposed feature sets are shown in Figure 2.9. There are several findings. First, we can observe that all models in general significantly outperform the majority class baseline (25.4% for EXPANSION:Conjunction).[22] Regarding preprocessing, we find that applying lower-case normalization to the input slightly improves the performance of the classifier (e.g., N_0 vs. l_0: 36.65% vs. 38.14% accuracy), however, a frequency threshold applied on the minimum number of occurrences of a feature does not seem to be helpful. Interestingly, this observation is not in line with the practices described in previous literature on implicit sense classification. Along with different feature selection criteria (e.g, mutual information with the classes), Lin et al. (2009) and Li and Nenkova (2014), for instance, use a minimum frequency cutoff of 5, ignoring features which occur less often in the training set. Also, we observe that stemming as another type of normalization seems not to be useful either and yields slightly worse accuracies as opposed to the raw tokens in each word pair.

Notably, when we replace the surface-level word pair features by the Brown Cluster 3200 representations, we obtain a slightly increased classifier performance. Even though this difference is not statistically significant, the improvement comes with a much smaller feature space ($\approx$ 1.4 million), which can be reduced by 23% as opposed to dealing with word pair features.[23] The feature sets that integrate the skip-gram neural word embeddings (*WordVecs*) have shown to outperform

(false).

[22] In all outlined experiments, the χ^2 test statistic was applied to assess significance.

[23] Lower accuracies were yielded with the other Brown Cluster representations (100, 320, and 1000).



(a) Without normalization

(b) With lower-case normalization

Figure 2.9: Performances (in % acc.) for 6-way classification on implicit discourse senses with different feature sets. N: normal-case, l: lower-case preprocessing; the indices refer to frequency thresholds for feature selection (0 = no threshold).

the Brown clusters. These features share similar contextual properties and at the same time they preserve the topology of the original feature space. They perform similarly well compared to the sparse, low-frequency word pair features and even significantly better than the configurations l_3, l_4, l_5. Their greatest benefit is attributable to the small number of real-valued features per argument-pair (600 dimensions only in our setting). Finally, when we combine the skip-gram representations and the Brown Clusters into one feature set, the best results can be obtained. We conjecture that this performance improvement over the embeddings alone may be due to non-linearities in the feature space that the Brown clusters can partly capture. Obviously, using only embeddings in combination with an SVM cannot account for this. It should be noted that all our results were obtained using linear kernels. The same experiments were conducted with polynomial and RBF kernels, but no improvements were yielded. However, we argue that non-linear (distribution-free) models offer a fruitful alternative to our proposed techniques and can be implemented with multi-layered neural networks. Since these classifiers can potentially yield better results by incorporation of word embedding features, such experiments will be described in the next two chapters.

Overall, the experiments outlined in Chiarcos and Schenk (2015a) demonstrate that frequency cutoffs for word pair feature selection do not seem to improve the classification scores on the task of implicit relation recognition. Whereas few prior approaches, most notably the one described in Lin et al. (2009), or the one outlined by Li and Nenkova (2014), make use of cutoffs in their systems, others do not. Yet, when a frequency filter is applied, most often the specific value of the threshold is either not motivated or it only appears as a side note. We argue that a potential explanation for the negative effect of cutoffs on the performance can be seen in the feature space that is extremely sparse. In fact, many word pair instances found in the training section of the Penn Discourse Treebank do not exist in the development set and vice versa. When frequency cutoffs are applied to the data sets, it is inarguably the case that the sparsity grows even further. In line with our observation are findings by (Blair-Goldensohn et al., 2007) which show that even a small list of stop words can have adverse effects on classifier performance. At first sight, this seems implausible but it supports our results presented here.[24]

In sum, all these observations again show that there is a difficult tradeoff between the quest for a more generic, less sparse representation, and the simultaneous desire to at the same time preserve the predictive power of each single feature in the large parameter space of the word pair model. We elaborate on alternatives in the final section of this chapter.

[24] It should be noted, however, that our setup differs greatly from the one in Blair-Goldensohn et al. (2007), e.g., in the label set and classification algorithms employed, the choice of the stemming algorithm, etc. Therefore, all findings are inherently specific to the parameterization of the experiments.

2.2.3.1 Sparse Features & Rich Linguistic Information: Towards More Abstract Representations

Word pairs are tremendously sparse, as has been proved, and amount to several millions on the PDTB data set. A great quantity of them are underrepresented and, contrary to previous claims, they do not bear any semantic content. This has also been pointed out by Braud and Denis (2015) and Pitler et al. (2009) who demonstrate that these are predominantly function word co-occurrences. In a systematic comparison by means of intensive feature optimization of previous works, Park and Cardie (2012) even conclude that—given the presence of other feature types—word pairs in fact only play a minor role in the classification task. Even though these observations contrast with the work of Biran and McKeown (2013), who propose to *aggregate* word pairs around semantically similar explicit discourse markers, striving for ways of abstraction, the appropriateness of word pairs in conjunction with synthetically acquired additional training data for the task of implicit relation modeling is still much-debated. The results from Chiarcos and Schenk (2015a) and all previously outlined related methodologies starting roughly in 2014 (Rutherford and Xue, 2014; Braud and Denis, 2015) have shown that it is beneficial to abstract from sparse surface-level information for at least two reasons. First, unsupervised word representations (*dense* and compact word embeddings or Brown clusters) seem to express a more general, semantic representation of the underlying relationship between two arguments in the discourse and, second, the number of features involved in a classification can be significantly reduced, which has a positive effect on computational efficiency.

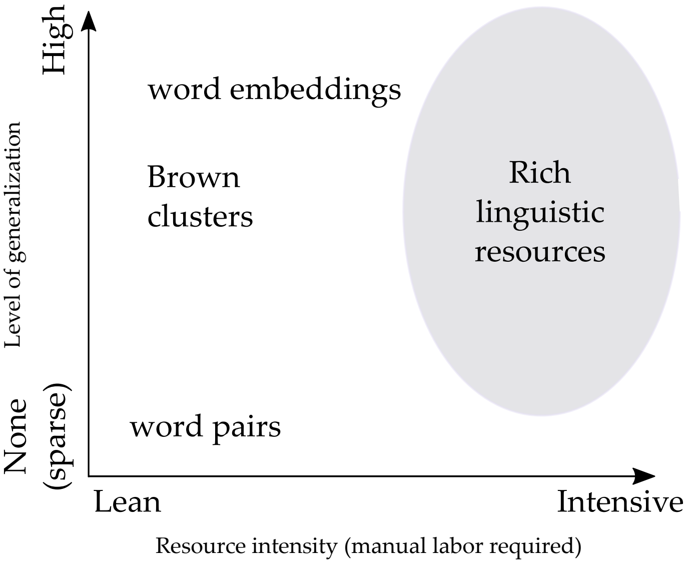


Figure 2.10: Different feature representations and their resource intensitivity vs. level of generalization for the task of modeling implicit discourse relations

In summary it can be said that the two main drawbacks common to all word pair-based and purely linguistically-informed methods—whose drawbacks we have already elaborated on in Section 2.2.3—are that they are either *sparse*, *expensive*, or both. Figure 2.10 illustrates this exact modeling tradeoff between sparse but easily obtained word pairs and more general but likewise more costly-to-produce rich linguistic information. Ultimately, cheaply acquired *unsupervised word representations* and their integration into advanced machine learning setups provide a solid basis for further research directions because they are more flexible, and thus offer various advantages. In the next section, we elaborate on closely related promising alternative strategies to overcome the feature engineering bottleneck in favor of *feature learning*, and introduce recent and more *generic* methods for implicit relation modeling.

2.2.4 Resource-Lean Implicit Discourse Parsing

2.2.4.1 Discourse Parsing by Deep Learning

As a reaction to the obvious drawbacks related to feature-sparse and resource-intensive discourse parsing, novel methods have led to the emergence of *resource-lean* modeling techniques in order to overcome the dependence on hand-crafted designs and specific scarce resources. After all, there has been a recent boom applying *neural networks* to NLP problems and the powerful "deep learning tsunami" (Manning, 2015a) has not only hit upon traditional NLP methods, such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2014) and speech recognition (Sak et al., 2015). It has also entered various other fields, such as abstractive text summarization (Rush et al., 2015; Lopyrev, 2015), multimodal sentiment analysis (Poria et al., 2015), the recognition of textual entailment (Lyu et al., 2015), natural language inference (Parikh et al., 2016; Mou et al., 2016), relation extraction (Zeng et al., 2014), semantic role labeling (Zhou and Xu, 2015), and finally has made its way into discourse processing, especially holding a lot of promise for implicit discourse relations. While, for example, in the first CoNLL shared task on shallow discourse parsing (Xue et al., 2015), only very few participants initially started to apply neural network-inspired architectures, in its second edition (Xue et al., 2016), one year later, the vast majority of submissions already focused on deep learning, improving upon the best results from the previous year, which shows the great importance and power of these methods.

In general, these techniques follow the paradigm of *representation learning*, due to Bengio et al. (2013), by which low-dimensional (dense) features are generated in a largely unsupervised manner. *Word embeddings* (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011b; Mikolov et al., 2013b), for instance, are a result of this process and have been shown to capture all the essential information of one-hot-encoded word representations from the originally sparse and high-dimensional input data. Any well-thought integration of the so-obtained embeddings into a neural classification framework can take advantage of their valuable syntactic and semantic properties and has been shown to significantly improve performance on downstream tasks.

In what follows, a general outline is given of structural concepts and methodologies based on deep learning which have successfully found their way into discourse unit identification and discourse sense classification. First, Section 2.2.4.2 introduces related work with a focus on general (both explicit and implicit) RST-style discourse parsing. Section 2.2.4.3 then highlights computational approaches based on the Penn Discourse Treebank, with special emphasis on the more challenging implicit discourse relations.

2.2.4.2 RST-Parsing

Ji and Eisenstein (2014): One of the first attempts to incorporate representation learning into RST-style discourse parsing, for both nuclearity detection and discourse relation identification, was made by Ji and Eisenstein (2014). The authors present an approach, called *DPLP* (*Discourse Parsing from Linear Projection*), which transforms a surface-representation of lexical features from gold-provided elementary discourse units (EDUs) into a *latent discourse space*, by discriminatively learning a linear projection function. The latent space is much lower-dimensional than the original bag-of-words feature representation. The approach is implemented as a multi-class shift-reduce parser in the style of previous work by Marcu (1999) and Sagae (2009). Specifically, their setup comprises a classification task in which features from EDUs on both a stack and a queue are used to decide on a suitable discourse relation. The task can be viewed as a large-margin transition-based structured prediction problem which jointly learns to project from surface features to the representative discourse vector-space. The main advantage of this approach can be seen in capturing the underlying meaning of EDUs and their relations without suffering from data sparsity of the originally high-dimensional input data. The projection matrix in fact successfully learns to group discourse-related words and connectives.

Li et al. (2014): Closely related, Li et al. (2014) introduced a *recursive neural network* for discourse parsing which jointly models distributed representations for sentences based on words and syntactic information from parse trees. The bottom-up approach is motivated by Socher et al. (2013) and models the discourse unit's root embedding (to represent the whole discourse unit) which is being obtained from its parts by an iterative convolution process. As originally proposed by duVerle and Prendinger (2009) and Hernault et al. (2010a), their system is made up of a binary structure classifier (in order to merge two text spans into a new subtree), and a multi-class relation classifier for discourse labeling. Both classifiers are three-layer neural network architectures and are trained jointly based on the discourse parse tree. The approach is novel in the sense that phrase-level distributed representations are extended via recurrent compositionality to an extended discourse level. Overall, the RST-style parsing performance described in Li et al. (2014) achieves competitive results compared to Ji and Eisenstein (2014).

2.2.4.3 PDTB-Parsing for Implicit Senses

Ji et al. (2016): A refinement of the work in Ji and Eisenstein (2014) is presented in Ji et al. (2016) who propose a hybrid architecture in the form of a *latent variable language model* and a *recurrent neural network language model* (*LVRNN*) for implicit shallow discourse relations from the PDTB. The authors argue that it can be employed both for discourse relation prediction and language modeling applications. Unlike in Ji and Eisenstein (2014), the model does not solely learn from the supervision signal from discourse annotations but also from the objective imposed by the language model, thus combining advantages of both probabilistic graphical models and neural networks. Practically, their model estimates the sequential flow of discourse information from one discourse unit to the next adjacent one, under the assumption of a latent variable which stands for the discourse relation. A comparison with accuracies reported in Rutherford and Xue (2015) shows that the approach beats the state-of-the-art in 4-way implicit sense classification on the PDTB.

Chen et al. (2016): Another architecturally sophisticated network model for implicit discourse relation classification is proposed by Chen et al. (2016). The motivation for their work is to replace sparse and hand-crafted features, e.g. word pairs, by dense, distributed representations for each word in both argument spans. Its purpose is to prevent the recognizer from potential ambiguity (e.g., for sentiment-contrasting word pair features which would be highly indicative of a wrong contrast instead of a correct causal relationship) and to overcome the lack of context in which it occurs. Specifically, Chen et al. (2016) first propose to encode all words in both arguments into an intermediate (positional) representation, which ties a word to its contextual information. Long-term dependencies between words are modeled via a bidirectional LSTM (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997). This is important in order to capture context about both the past and future for any discourse relation. In a second step, a relevance score is computed for each intermediate (word) pair representation, which essentially measures the (linear and non-linear) interaction between the two vectors (e.g., by means of cosine similarity). To this end, the authors introduce a *Gated Relevance Network*, whose essence consists of a gate mechanism which returns information about how the interactions should be combined. As a result, one obtains a semantic score matrix which indicates the strength of each word pair interaction. Finally, scores are fed into a pooling layer and multi-layer perceptron for discourse label classification. The complete framework is illustrated in Figure 2.11. Chen et al. (2016) demonstrate the usefulness of the gate property of their network architecture by showing that some words appear to obtain more predictive power for discriminating implicit relations depending on the specific contexts in which they appear. Their approach represents the state-of-the-art in 4-way (top-level class) implicit sense detection.

Zhang et al. (2015, 2016): Very recently, Zhang et al. (2015) have developed the idea of a *shallow convolutional neural network* (*SCNN*) to model implicit relation detection within the PDTB-framework. A prerequisite for the network construction is to replace each argument by a stacked word embedding matrix: each column in the matrix represents a single word by its distributed representation as initialized from large amounts of unlabeled data. Inspired by the work in Socher

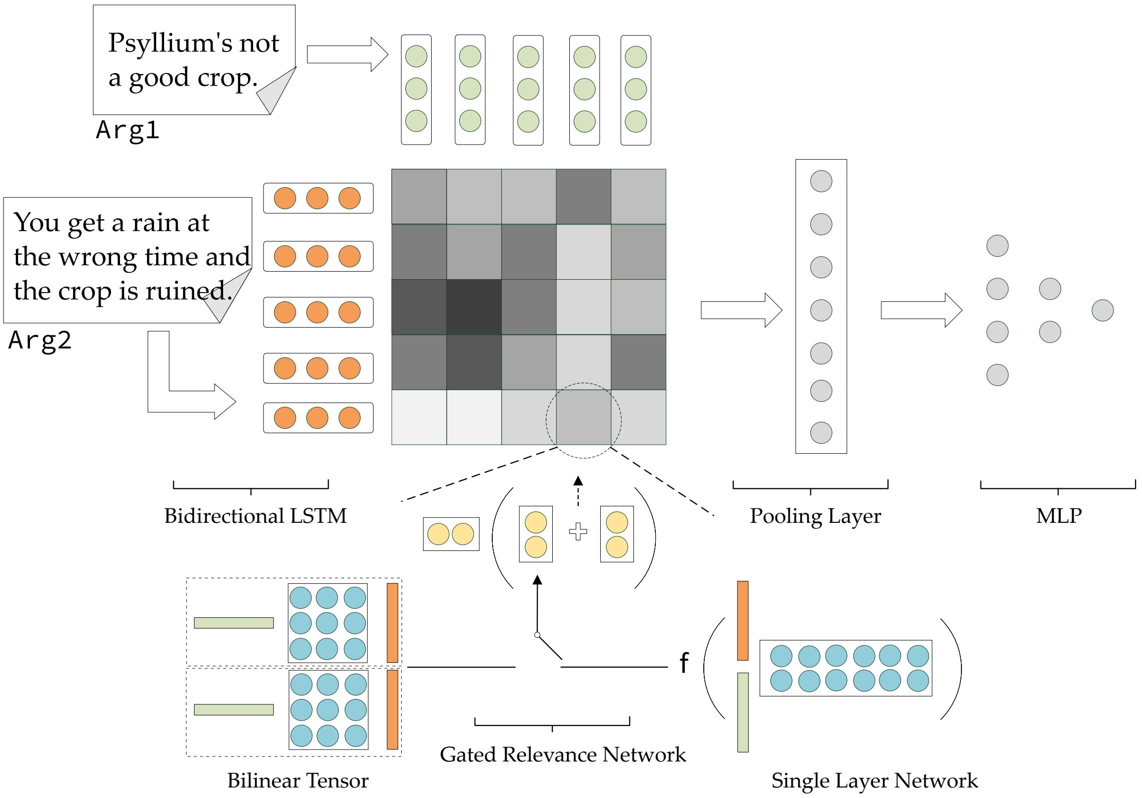


Figure 2.11: Bi-LSTM Gated Relevance Network Architecture as proposed by Chen et al. (2016). Tokens in arguments 1 and 2 are substituted by embeddings (top part). A bidirectional LSTM encodes their positional information. The gate mechanism (lower part) associates a relevance score with each word pair, expressed by the semantic score matrix (shaded gray square). Higher intensities represent a greater semantic interaction towards a certain discourse relation.

et al. (2011), three convolution operations are applied to the two word embedding matrices each, namely *average* vector calculation, *min* and *max* computations, in order to extract the structurally predominant information within both argument spans. All resulting six pieces of feature representations (three convolution operations per argument matrix) are concatenated to form one vector onto which a non-linear transformation and length normalization is applied. This, in fact, constitutes the only transformation on the words. The result is integrated into the neural network's single hidden layer. A softmax output layer is employed for relation classification. The overall network structure is depicted in Figure 2.12. Note that—unlike the previous methods—this network completely ignores sequential word order information within argument spans. The main advantage of this framework is expressed by its simplicity which alleviates issues related to overfitting on training data and better generalization on the test set. The authors present an evaluation setup involving traditional rich linguistic feature sets and demonstrate that their SCNN approach is superior in terms of performance when

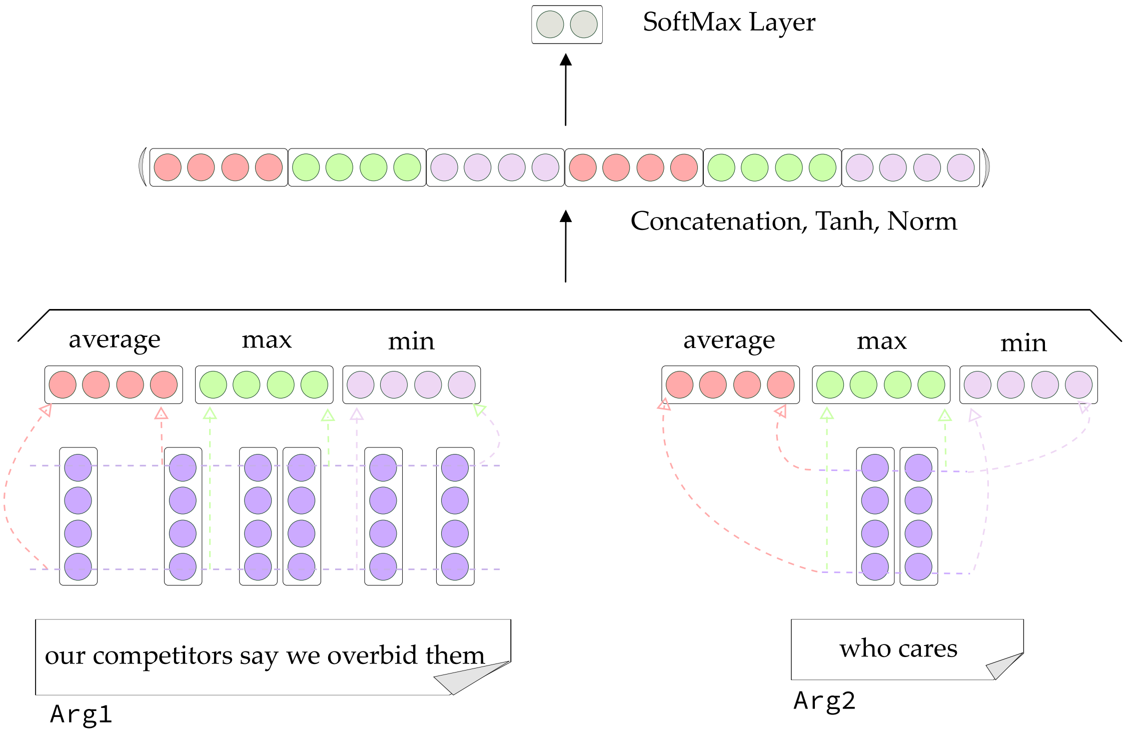

SoftMax Layer
Concatenation, Tanh, Norm
average
max
min
average
max
min
our competitors say we overbid them
Arg1
who cares
Arg2

Figure 2.12: The Shallow Convolutional Neural Network as proposed in Zhang et al. (2015) with arguments substituted by word embeddings (blue color), a hidden layer obtained by convolution operations, and a final softmax output layer

parameters are tuned accordingly.

A refinement of the work in Zhang et al. (2015) is presented in Zhang et al. (2016a). Here, the shallow convolutional neural network is extended with an additional component. The underlying idea is that the surface representations which are encoded in the shallow architecture could benefit from a *semantic memory* extension. This semantic memory emulates and stores general knowledge about factual data and concepts—similar to how cognitive processes for understanding and comprehension in the human brain function (Yee et al., 2013). Technically, the semantic memory is a matrix of distributed word representations which is obtained by an attention model (Mnih et al., 2014; Xu et al., 2015). Both surface representations and semantic memory information are fed into a semantic encoder which generates a deep, enhanced representation of the discourse arguments. Different attention weights for individual words in the semantic memory matrix reflect their importance with respect to a certain discourse relation. Zhang et al. (2016a) report slight performance improvements by inclusion of their novel component.

Liu and Li (2016): Similar to the previous work described above by Zhang et al. (2016a), the architecture presented in Liu and Li (2016) also benefits from an external memory component—this time even with multiple levels of attention. The authors introduce a system termed *neural networks with multi-level at-*

tention, *NNMA* for short. Their approach is motivated by analogy to the process of *repeated reading*, in which the two argument spans of a discourse relation are scanned multiple times for a deeper analysis of the text as opposed to a single-pass only. More precisely, this particular re-reading strategy has been shown to be advantageous by simulating human reading behavior. It is characterized by scanning a text repeatedly in order to better understand it by gradually pinpointing key features in the two arguments to finally infer an appropriate discourse relation. This dynamic process is emulated by the *NNMA* and its overall architecture of the attention-based model is illustrated in Figure 2.13.

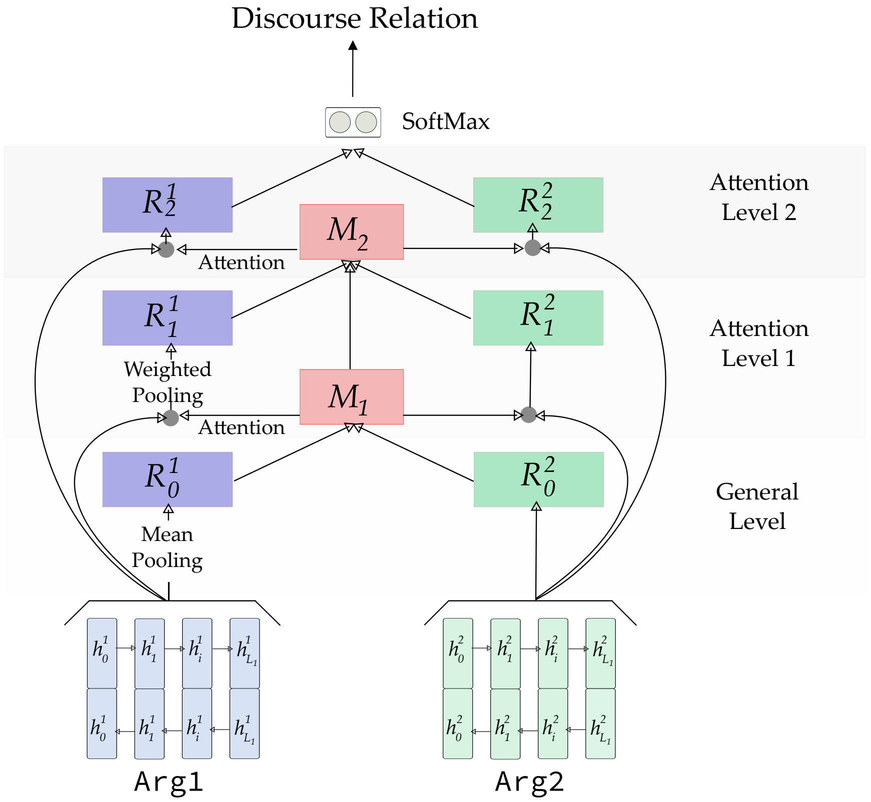


Figure 2.13: The multi-level attention system architecture introduced in Liu and Li (2016). The bottom part shows the two argument spans (`Arg1` and `Arg2`) which are first encoded by a bi-LSTM into the general level. Two subsequent attention layers are stacked on top to emulate the repeated reading strategy.

Technically, their model is a combination of a bidirectional LSTM for the general argument representation (i.e. for the simulation of just skimming the text which is represented in the lower part of the figure), and a variable number of stacked attention layers on top (simulating the thorough repeated reading process). First, pooling operations produce argument representations in the general level, while the memory vector of the first attention level is a combination of the two argument representations, which in turn is used to "re-calculate" the importance of each word. The overall process of tuning word weights through non-linear transformations by subsequent attention layers gradually infers more precise attention vectors. In order to classify an implicit discourse relation, the output of the topmost attention level is used. The architecture is flexible in the sense that a variable number of attention layers can be stacked on top of the general argument representation. The system described in Liu and Li (2016) achieves state-of-the-art performance in the PDTB, and the authors have shown that the repeated reading strategy improves upon a general (bi-LSTM only) argument representation. Also, a visualization of the attention activities provides useful insights into the inner workings of their model to progressively locate discourse relation-specific key features in either of the two argument spans.

Qin et al. (2017): The work of Qin et al. (2017) is inspired by a peculiarity of the PDTB by which—for each implicit discourse relation—annotators of the corpus were prompted to decide on a suitable connective which best describes the relation. Along the lines of deep generative modeling (Goodfellow et al., 2014), which has demonstrated recent success in image processing, Qin et al. (2017) make use of these additional connectives and introduce the first *adversarial* architecture applied to implicit discourse relations in the PDTB. The idea is to employ a feature-augmented strategy to learn better contextual discriminative features beyond the tokens in both arguments (which has been the standard approach in all previously outlined neural architectures up to now). To be more precise, this feature emulation process involves two counterpart networks, a standard one which has access to the tokens in both arguments, and a second one which makes also use of the additional implicit connective. The latter, the feature-augmented network, and a third component, a rival discriminator which tries do distinguish between the features from both networks, form an *adversarial pair*. During an interleaved training procedure, the implicit recognizer (the one without access to the connectives) is optimized on training data and, at the same time, the discriminator is to be confused by minimizing the chance of distinguishing between the input features obtained from the two networks. The result produces features which are close to the connective-augmented ones and at test time the model demonstrates superior performance over prior neural models—even without access to implicit connectives.

The system described in Qin et al. (2017) works end-to-end and successfully "mimics" a network which benefits from side information, yet the authors only evaluate their system on the PDTB class levels and leave a full discourse analysis as, e.g., in the context of the CoNLL shared tasks (Xue et al., 2016) for future work. The approach is orthogonal to the prior methods on data augmentation using explicit argument pairs (Rutherford and Xue, 2015; Braud and Denis, 2016) and complementary to the work by Zhou et al. (2010) who propose to learn a language model on unannotated data which can be used to predict discourse connectives in between arguments of implicit relations.

Further Related Work & Summary: Besides the works described in this section, there has been emerging research in related areas drawing on various aspects of implicit discourse parsing; cf. Zhang et al. (2016b), for instance, for a variational neural discourse relation recognizer or Liu et al. (2016) for a promising attempt to multi-task learning techniques across RST and PDTB corpora. Also, very recently, in the context of the CoNLL 2016 shared task on shallow discourse parsing (Xue et al., 2016), a whole series of neural variants have been proposed, e.g, by Pacheco et al. (2016) using event embeddings, Qin et al. (2016) and Wang and Lan (2016) presenting a filter-based approach using convolutional neural networks (the former including part-of-speech embeddings), and most notably by Weiss and Bajec (2016), who very successfully employ language-agnostic focused, deep RNNs in an end-to-end fashion without any external resources. Finally, Rutherford et al. (2017) provide a recent overview and a comparison among various neural network architectures for implicit discourse relation recognition and demonstrate that feedforward systems (as opposed to recurrent networks) are particularly powerful.

To summarize, the fundamental idea which is shared by all of the above described methods is that resource-lean learning can indeed improve upon traditional, resource-intensive methods—especially with regard to the multifaceted problem of dealing with implicit discourse relations. This can be attributed to the following major factors: Overall, task-specific setups have been replaced by the generic representation learning paradigm. Specifically, in the distributed frameworks outlined in this section, thoroughly choosing network setups and carefully learning hyperparameters has superseded extensive manual feature engineering for rich linguistic features. In this respect, compact word embeddings have replaced sparse lexical information (e.g., word pairs) and domain-specific word lexicons.

Based on these fruitful ideas, we follow the impulse of prior works and introduce a highly generic neural network architecture for PDTB-style implicit discourse relation classification, which we describe in the next chapter.

Chapter 3

Lightweight Parsing with a Feedforward Network

3.1 Motivation

On the basis of the great success of resource-lean deep learning methods in NLP, and specifically their application to discourse parsing, this chapter presents a novel neural network-based architecture for implicit sense classification. In the style of prior network architectures on the same task, our proposed system has to fulfill the requirements of being structurally simple in design, yet effective in performance. It is grounded on a *feedforward* neural network setup and does not rely on sparse surface word forms or any other type of handcrafted features. Instead, by following the representation learning paradigm (cf. Section 2.2.4.1), the network introduced in this chapter is largely language-independent and has proven to be effective for both English and Chinese discourse relation classification. Specifically, the design of its structurally lightweight components is inspired by the Shallow Convolutional Neural Network of Zhang et al. (2015) for implicit sense detection and its design principles are driven by the findings in Rutherford and Xue (2016), who demonstrate that simple feedforward neural network architectures—when thoroughly tuned—typically outperform more complex LSTMs on the same task. The proposed framework is a *shallow* discourse relation classifier. The classifier takes as input pairs of two argument spans. Solely based on the weighted distributed representations of the words in the arguments a final decision is made on a suitable label that best describes their relation. Within this framework, entity relations are treated equivalently as an additional class on top of the other implicit relations. The reason for that is that, first, they are semantically related, i.e. entity-based coherence is typically driven by pronominal reference or anaphoricity and thus their relationship is missing an explicit connective, too, and prior research on implicit sense detection has treated them in a similar fashion.

Apart from being structurally similar in design, the framework described in this section demonstrates several innovative improvements over prior work on feedforward neural network-based attempts to modeling implicit discourse relations. We demonstrate that i) **unsupervised pre-training** of out-of-the-box embeddings can contribute significantly towards the overall resolution of implicit senses, ii) an **incorporation of syntactic information** into argument representations can further improve results, and iii) an elaborate but even **simpler network architecture** compared to prior works can yet enhance classifier performance on the task. In the next section, we introduce design principles and the network architecture of our proposed model.

3.2 Design Principles & Network Architecture

3.2.1 Argument Representation

Both arguments of an implicit discourse relation, `Arg1` and `Arg2`, are essentially made up of the respective words in the spans including punctuation symbols. For further computations, it has been shown to be inconvenient to employ the words (i.e. their one-hot encoded, discrete vectors) directly. More favorable methods incorporate lower-dimensional (i.e. *distributed*, dense and real-valued) representations of a certain dimension *dim*. These embeddings encode latent syntactic and semantic properties of a word, a characteristic that one-hot vectors do not share. Specifically, as a prerequisite to representing arguments, we model the set of all words D (the dictionary) by a corresponding word embedding matrix $D \in \mathbb{R}^{dim \times |L|}$ (with $|L|$ being the lexicon size). Accordingly, each column in this matrix corresponds to a word in the vocabulary. The approach described here is word order agnostic. Therefore, we represent a single syntactic argument S by its set of word tokens $S = \{t_1, t_2, t_3, t_4, \ldots, t_n\}$, where n is the total number of words in the argument. Note that each word $i \in [1,n]$ maps to a corresponding entry in D and that by consulting the global dictionary, each word can be substituted by its dense embedding by simply retrieving the corresponding columns from D, i.e. $D_{t_i} = a_{t_i}$. This leaves us with a stacked embedding matrix $V = (a_{t_1}, a_{t_2}, a_{t_3}, \ldots, ...a_{t_n})$ for each argument, where $a_{t_i} \in \mathbb{R}^{dim}$ and $V \in \mathbb{R}^{dim \times n}$. We distinguish between $V(1)$ for the first argument and $V(2)$ for the second.

3.2.2 Fine-Tuning Word Embeddings

The word vectors for the construction of D are typically precomputed and come as external resources, for instance from the *word2vec* toolkit[1]. The *word2vec* library (Mikolov et al., 2013a) is a collection of models for computing word embeddings from raw, large amounts of unstructured (plain) texts. The models themselves are neural network-based algorithms and their underlying way of functioning is to model contexts of words as encountered in large corpora by minimizing

[1] `https://code.google.com/archive/p/word2vec/`

reconstruction error. There are two popular implementations, the continuous bag-of-words (CBOW) and the skip-gram method with negative sampling. The former essentially tries to predict the current word based on its context, the latter works the other way round by predicting the context with focus on the current word. For the induction of word embeddings from plain text, choices are given for two training algorithms—hierarchical softmax or negative sampling. The negative sampling algorithm seems to be superior when the vector dimensionality is low. Typically, the size *dim* for the precomputed vectors ranges between 50 and 300 and is also parameterizable during word vector training. The quality of the vectors, i.e. their syntactic and semantic properties, tend to improve with increasing corpus size. The popular online-available pre-trained vectors from the Google News data set underlie approximately 100 billion words. These embeddings are powerful and incorporate useful features which allow for word-to-word comparisons by simple measures of distributional similarity, e.g. cosine distance.

The out-of-the-box embeddings are sufficiently robust and qualitatively suitable for most NLP applications, and can be used directly to construct D. A better alternative would be to *improve* upon these word embeddings first for task-specific purposes. This means that in the context of discourse parsing, one would like to see the performance of any pre-computed embedding to be adapted towards the data at hand—in our case to the specific writing style of the Wall Street Journal genre from the Penn Discourse Treebank texts. The goal would be to, firstly, extend the embedding collection in terms of their coverage, e.g., by adding vectors for out-of-vocabulary words or punctuation symbols which might not be there yet in the data set of pre-trained embeddings. Secondly, one would like to enhance their quality with regards to task-specific semantic and syntactic properties, thus ultimately increasing their predictive power in discriminating between difference discourse senses. This process is termed *unsupervised pre-training* and refers to the task-specific adjustment of pre-trained word embeddings. We report on implementational details in Section 3.2.5.

The finally adjusted, task-tuned word vectors are then incorporated into an updated word embedding matrix D_{tuned} and—compared to standard out-of-the-box embeddings—represent an important extension to the work in Zhang et al. (2015) for implicit discourse parsing.

3.2.3 Incorporating Syntactic Information

Previously, we have contrasted feature-rich with resource-lean modeling techniques and have implied that both have their pros and cons independently of each other. Sometimes, however, it might be valuable to combine positive aspects from both worlds. For example, it could be beneficial to incorporate—to a certain extent—portions of linguistic information into the overall framework for analyzing implicit discourse relations. More precisely, another refinement on the tuned word vectors can be made by integration of *syntactic dependencies*. A syntactic dependency analysis of a sentence (Kubler et al., 2009) models the relationship between individual words in terms of a head-dependent structure, which gives

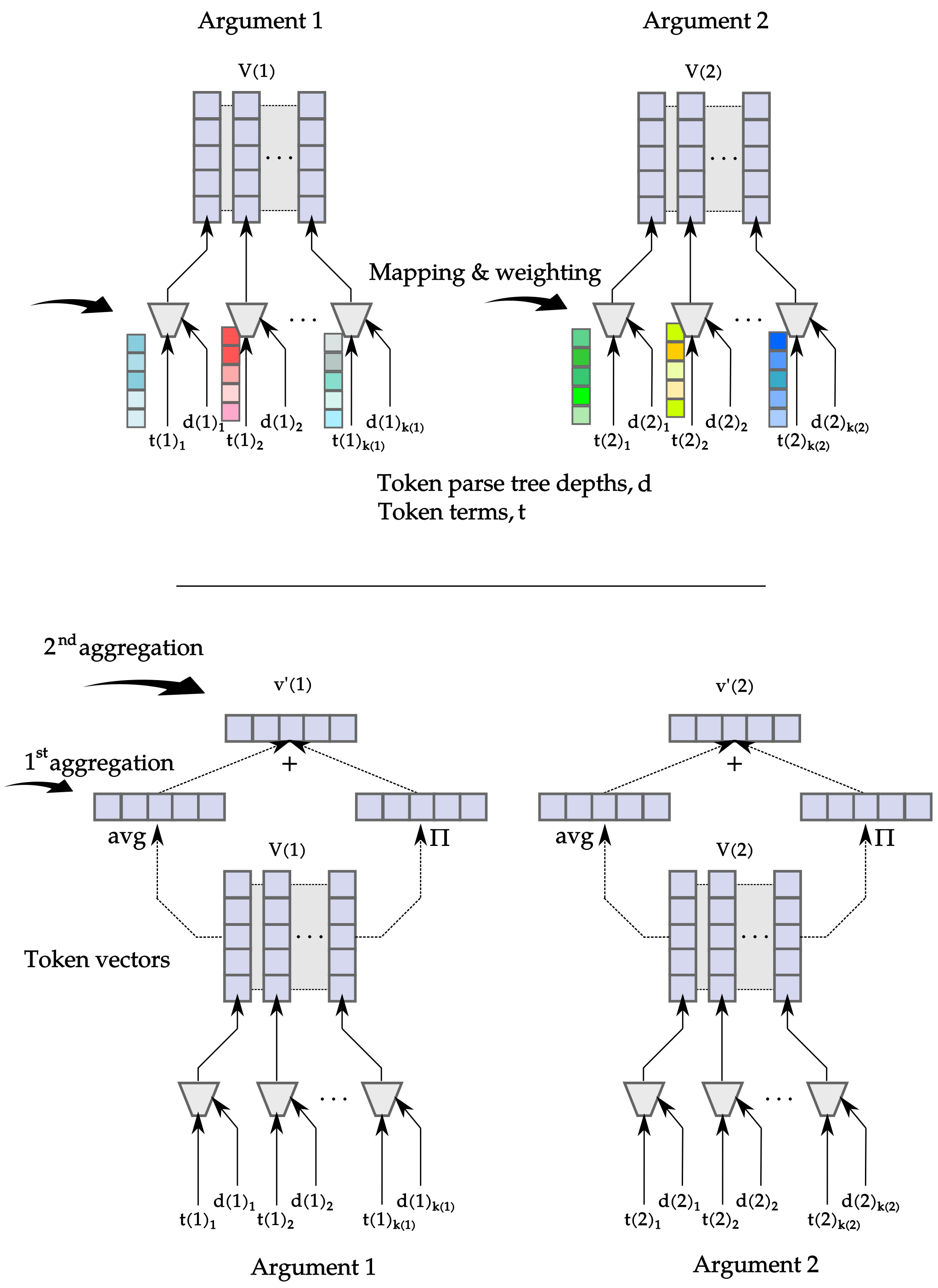


Figure 3.1: Argument representation and construction process. Tokens are substituted by fine-tuned embeddings, incorporating weighted syntactic dependencies, to build up argument matrices (top). Compositional operations (aggregations) on argument matrices produce two single-vector representations for each argument (bottom).

pointed insights into syntactic functions and relative contribution of certain words with respect to the global meaning of a sentence. For instance, words which are less deeply embedded within the dependency graph, and are thus more dominant, such as main verbs, carry potentially more relevant information as opposed to purely functional categories.

Returning to the argument construction for implicit discourse relations, specifically, for each token vector within an argument matrix, $V = (a_{t_1}, a_{t_2}, a_{t_3}, \ldots, ...a_{t_n})$, we calculate the token's depth d—relative to the root node—based on an automatically produced syntactic parse of that sentence.[2] The depth for each token is then fused with the tuned embedding a_{t_i} of the word by weighting the vector by the factor $\frac{1}{2^d}$.[3] This procedure shrinks a word's embedding value exponentially according to the depth of the current word in the parse tree. The motivation here is to scale the importance of words according to their syntactic dominance in the sentence: less deeply embedded words in the parse tree should be more important and overall more representative towards a certain discourse relation, while words whose embeddings have near-zero values (which make up the majority of words in very long sentences) should contribute less information. The top of Figure 3.1 illustrates this process, i.e. mapping tokens to their corresponding vectors in D_{tuned} based on the updated word vector model, as well as the integration of the token depth weighting. The result is a pair of two syntactically-informed argument matrices, $V(1)$ and $V(2)$.

3.2.4 Network Composition

Given the two argument matrices $V(1)$ and $V(2)$, Zhang et al. (2015) propose three convolution operations in terms of *average, minimum* and *maximum* computations over each row in the two word embedding matrices. With this method their model captures all the extreme and distinctly marked information with respect to the overall shape of the argument representations. These six pieces of information (three operations per argument) are concatenated to obtain a final vector for a single discourse unit. This is convenient, as this procedure always brings out a fixed-length representation of an input, even if arguments vary in the number of tokens.

Analogous to the well-performing model in Zhang et al. (2015), a slight variety in terms of a simpler network construction process involving only two aggrega-

[2] Automated dependency parsers, such as the Stanford Parser (Chen and Manning, 2014) or related phrase structure to dependency converters are freely available and come with pre-trained models for a variety of different languages. This makes them especially suitable for our purposes and enables large-scale, fully automated data acquisition.

[3] Even though this factor is a heuristic, it has been optimized on the development data (cf. Section 3.3). Note that some tokens might be missing in the parse tree, e.g., punctuation symbols. If these cases are encountered, experimental results have shown that an optimal default strategy weights them by a factor of 0.25.

tion operations is described in Equation 3.1:

$$\vec{v}'(j) = \frac{1}{n(j)} \sum_{i=1}^{n(j)} V(j)_i + \prod_{i=1}^{n(j)} V(j)_i \tag{3.1}$$

In the equation, $\vec{v}'(j)$ is a composed argument representation and is computed for both arguments $j \in \{1,2\}$ individually, where $n(j) = |S(j)|$ defines their lengths by the number of tokens for each span. The first component in Equation 3.1 computes a vector average (*avg*) of $V(j)$ and $\prod$ applies the pointwise product $\odot$ over the token vectors in $V(j)$. This process is illustrated at the bottom of Figure 3.1 (denoted by the first aggregation). Then, average and pointwise product are summed ($+$) to produce the final compositional vectors $v'(1) \in \mathbb{R}^{dim}$ and $v'(2) \in \mathbb{R}^{dim}$ for each single argument (denoted by the second aggregation step). A last step concerns an ultimate concatenation ($\oplus$) of both individual argument representations, $v'(1)$ and $v'(2)$, into a final neural input layer of dimensionality $\mathbb{R}^{2*dim}$ which serves as input to a feedforward network.

The network is set up with one hidden layer on top and a softmax output layer to classify among implicit senses in the discourse sense detection task. Our proposed global model architecture is depicted in Figure 3.2.

A Note on the Aggregation Functions: Note that, both aggregations—the average and the pointwise product of the vectors—produce simple argument representations that do not account for any type of word order variation or any other sentence structure information, yet they serve as decent features for the discourse parsing task and have been well-established in related experiments. Using pointwise multiplication over token vectors has the advantages that vector elements which stem from independent, latent semantic dimensions are not simply bundled up, but can scale according to their mutual relevance. We have tested in experiments, that Equation 3.1 performs better than simpler compositions of only either multiplication or average. This observation provides further evidence that it seems plausible to not completely suppress the dimensions that contain near-zero values for individual tokens.

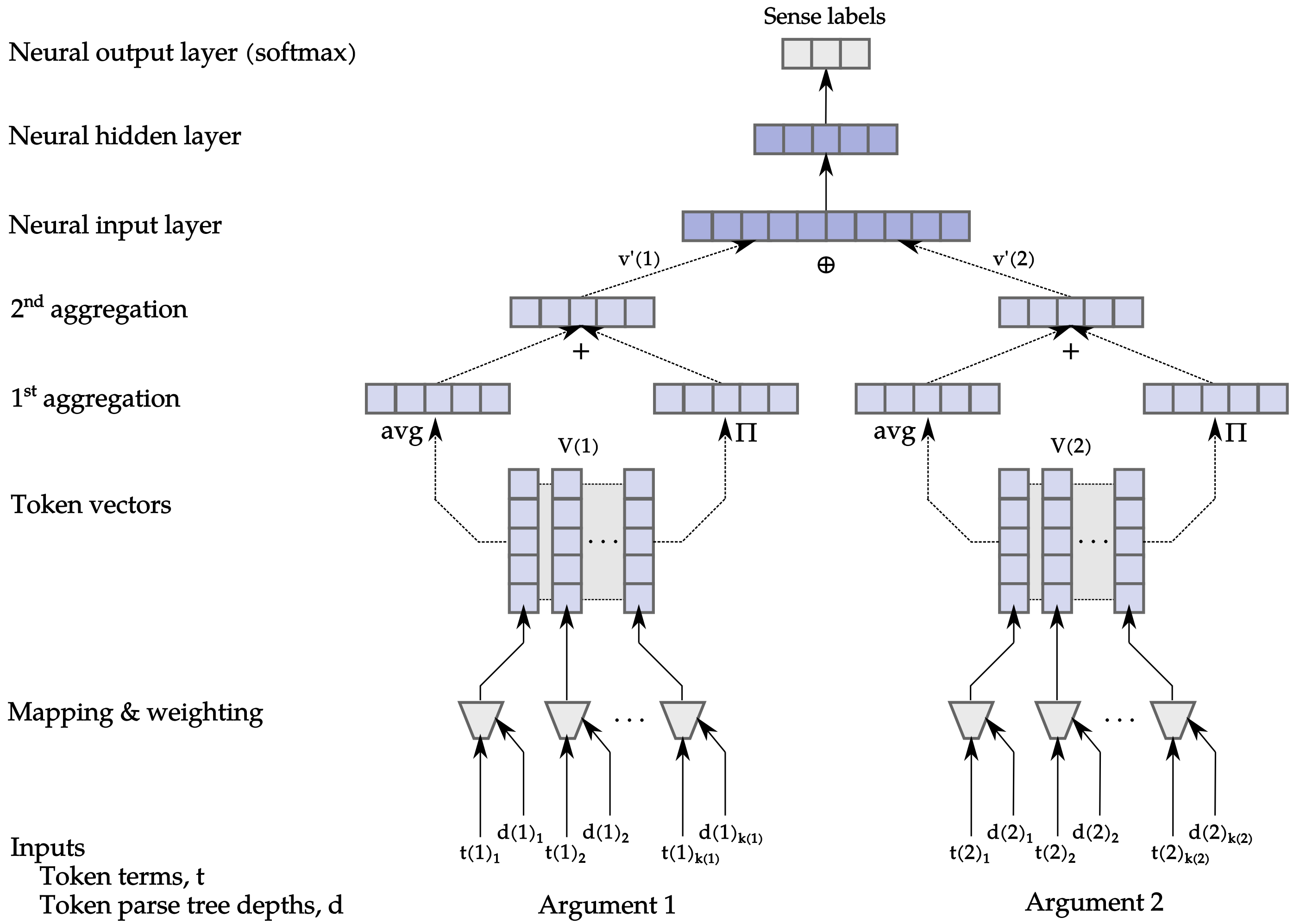


Figure 3.2: The global architecture for implicit sense classification, including entity relations. Argument span feature construction process (light blue) and neural architecture (dark blue). Dotted lines indicate pointwise vector operations.

3.2.5 Implementational Details on Training the Network

3.2.5.1 Training & Evaluation Data

The network described in this section has been designed and optimized in the context of the CoNLL 2016 shared task on shallow discourse parsing (Xue et al., 2016) for the Penn Discourse Treebank (Prasad et al., 2008, PDTB) and the Chinese Discourse Treebank (Zhou and Xue, 2012, CDTB). It is set up to distinguish between a classification among 20 implicit senses for English and 9 for Chinese, plus an additional `EntRel` (entity relation) label for each language. Other relations, such as `AltLex`, which are too infrequent to be of any statistical relevance, are not modeled.

The network for implicit sense detection has been trained on implicit relation pairs from the training section of the PDTB (sections 2-21 of the Wall Street Journal texts) and its parameters have been optimized on implicit relations from the development set (section 22). In the shared task, an official evaluation is performed on the test set from section 23. Similarly, for the Chinese data the CDTB training portions are taken from sections 0001-0270 and 0400-0803, respectively, section 0301-0325 constitutes the development set, section 0271-0300 has been applied for testing. In addition to the above mentioned test sets, the shared task organizers have provided an additional blind test set for each language, which consists of newswire texts, annotated in accordance with the PDTB annotation guidelines.

3.2.5.2 Vector Pre-training

In all experiments, the pre-trained Google News vectors[4] (for English) and the Gigaword-induced vectors[5] (Graff and Chen, 2005) (for Chinese) provided by the shared task organizers were used as initial resources for the argument construction process.[6] Unsupervised pre-training, i.e. vector adjustments have been performed on the raw Wall Street Journal texts, thus tuning the embeddings towards the specific genre, with the goal of considerably improving their predictive power and coverage in the sense classification task. We give pseudo code for this procedure in Figure 3.3. Specifically, the pre-trained Google News vectors of size 300 were updated by the skip-gram method due to Mikolov et al. (2013a) using a *word2vec* model. We found a window size of 8 and a minimum term count of 3 to be optimal during multiple passes over the newswire texts while *steadily decreasing the learning rate*. In detail, 20 iterations have been made on the task-specific PDTB texts and alpha parameters have been adjusted after every iteration. The idea is that, loosely speaking, the pre-trained vectors should only be slightly tuned and progressively adapted towards the PDTB data set and not undergo a completely new initialization during each iteration.

[4] `https://code.google.com/archive/p/word2vec/`

[5] `http://www.cs.brandeis.edu/~clp/conll16st/data/zh-Gigaword-300.txt`

[6] `http://www.cs.brandeis.edu/~clp/conll16st/dataset.html`

```
#Initalize word2vec model.
m = Word2VecModel(size=300, window=8, min_count=3)
#Read pre-trained Google News embeddings.
    m.intersect_word2vec_format(googlevectors)
    #Fine-tune the vectors...
    for iteration between 1 and 20
        m.alpha = 0.01/(2**iteration)
        m.min_alpha = 0.01/(2**(iteration+1))
        #Re-train on new iteration.
        m.train(PDTB_data)
```

Figure 3.3: Pseudo-code for unsupervised pre-training (task-specific adjustment) of word embeddings with pre-computed Google News vectors for the implicit discourse parsing task

3.2.5.3 Hyperparameters

The network was trained with Nesterov's Accelerated Gradient (*nag*) (Nesterov, 1983). The hyperparameters were optimized on the development set. The *rectified linear activation with learnable leak rate and gain (lgrelu)*[7] yielded the best results, in combination with 40-60 hidden nodes, weight decay, and hidden node regularization of 0.0001. The learning rate was set to 0.0001. Momentum of 0.35-0.6 and 60 hidden nodes performed well on the PDTB data for English, and momentum of 0.85 and 40 hidden nodes on the CDTB for Chinese (however, with fewer output nodes). Similar performances were obtained by *parametric rectified linear unit (prelu)* activation, as well as an increased hidden layer size combined with stronger regularization (e.g., L1 regularization of 0.1 on 100 nodes). An overview of the network-specific hyperparameter settings and optimal configurations for both English and Chinese are given in Table 3.1.[8]

3.3 Evaluation

3.3.1 A Note on the Evaluation

Despite the great success of regularly upcoming and inspiring implementations for implicit discourse parsing, it should be noted that it has been tremendously difficult to compare individual findings of novel and improved algorithms directly with previous attempts on the discourse relation detection task. The main issues can be seen in that most prior research has focused only on the general *class-level* comparison with only four senses, i.e. EXPANSION, COMPARISON, CONTINGENCY and TEMPORAL and has not attempted to model the complete

[7] `http://theanets.readthedocs.io/en/stable/api/reference.html`

[8] All neural networks were trained using the *gensim* package: `http://radimrehurek.com/gensim/`

Parameter	English (PDTB)	Chinese (CDTB)
method	*nag*	
learning rate	0.0001	
min_improvement	0.0001	
validate_every	5	
patience	5	
momentum	0.6	0.85
weight_l1	0.0001	
hidden_l2	0.0001	
hidden	60 *lgrelu*	40

Table 3.1: Optimal hyper-parameter configurations, number of nodes in the hidden layer (hidden) and activation functions of the feedforward neural network component for implicit sense labeling based on the PDTB & CDTB development sets (*lgrelu = rectified linear activation with learnable leak rate and gain*).

PDTB role inventory—or at least the 2nd-level class hierarchy. The fact that some researchers have preferred one-versus-all evaluation, i.e. by training 4 binary classifiers, whereas others have used the more challenging 4-way classification (with four output labels as in Ji et al., 2016; Rutherford and Xue, 2015), made an appropriate evaluation in terms of a direct comparison of results almost impossible. Beyond that, Zhang et al. (2015), for instance, do not compare their results directly with other state-of-the-art systems on exactly the same formulated task, as minor statistics on training and test data differ, e.g., from the setups described in Chen et al. (2016). Issues like these have emerged also due to different versions of the Penn Discourse Treebank (Miltsakaki et al., 2004; Prasad et al., 2008). Fortunately, the first and second edition of the CoNLL shared task on shallow discourse parsing (Xue et al., 2015, 2016) have set the stage for a first independent evaluation on that task. Within a unified framework, participating systems were evaluated on the more fine-grained 2nd-level hierarchy (instead of only 4 classes) and performances were measured in a fully automated server environment of the TIRA evaluation platform (Potthast et al., 2014).

With the framework for implicit relation detection which has been presented in this section, we have participated in the CoNLL 2016 Shared Task on shallow discourse parsing. Details can be referred to in the accompanying publication by Schenk et al. (2016). We elaborate on the setups and evaluation of the framework below.

3.3.2 Supplementary Task—Discourse Sense Classification

In the closed track of the competition's *supplemantary task*, gold arguments were provided for English and Chinese texts, and a participating system had to detect the correct sense label for each relation pair solely based on the tokens in the arguments and the pre-trained embeddings provided by the organizers. Participating systems were evaluated on three subtasks: *non-explicit*, i.e. implicit relation detection, including `EntRels` (without any connectives), *explicit* relation classification (with the presence of a connective) and their weighted combination (*all* parser performance)—all of them measured in F_1-score, the harmonic mean between precision and recall.[9]

3.3.2.1 Labeling Non-explicit Relations

For English and Chinese argument pairs, we have applied the framework presented in this section—a feedforward neural network-based algorithm for the classification of implicit discourse senses. All parameters have been optimized for both languages separately and are shown in Table 3.1.

3.3.2.2 Labeling Explicit Relation

For the detection of *English explicit* senses, we have made use of the system described in Stepanov et al. (2015), which has performed very successfully in the first edition of the shared task, especially on explicit relation pairs. The system makes use of surface-level token features.

For *Chinese explicit* relation pairs, we followed Occam's razor with the minimalist approach described in Chiarcos and Schenk (2015a), and trained a linear-kernel SVM classifier using a single feature—the connective token.

3.3.3 The Performance in the CoNLL 2016 Shared Task—Official Evaluation

The official shared task results for the supplementary task are split up into non-explicit parser performance (i.e. for implicit senses and `EntRels`), explicit, and a combined (all) parser performance.

3.3.3.1 Non-Explicit Relations

The accuracies for our model on the standard English (PDTD) dataset are illustrated in Figure 3.4, while the Chinese (CDTB) classifier accuracies are reported in Figure 3.5.[10] Throughout all evaluation scenarios, it can be seen that non-explicit

[9] Besides the implicit parser performance, we also report the explicit and overall results for reasons of completeness.

[10] Our system had participated under the name of *Goethe* and its performance is graphically indicated within the charts.

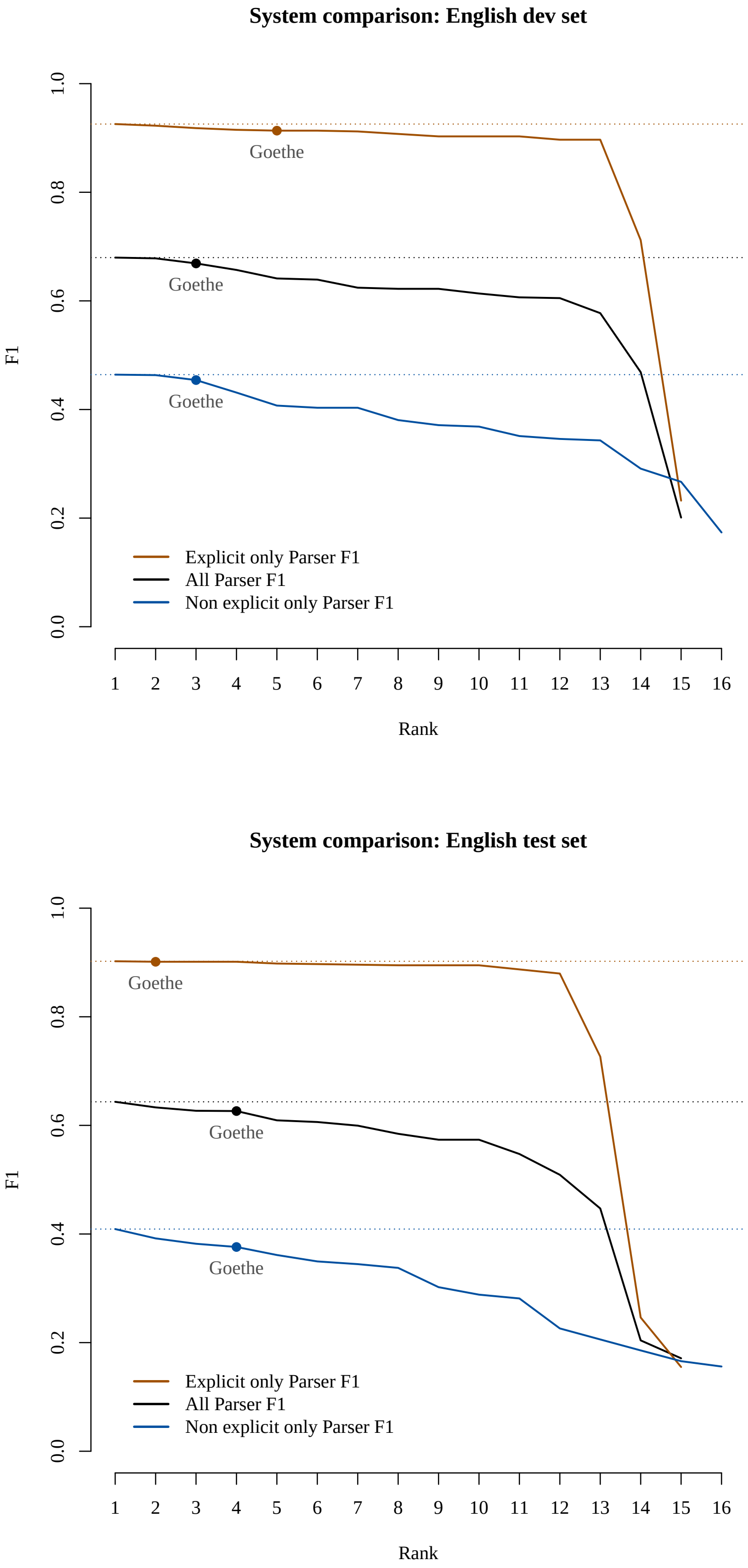
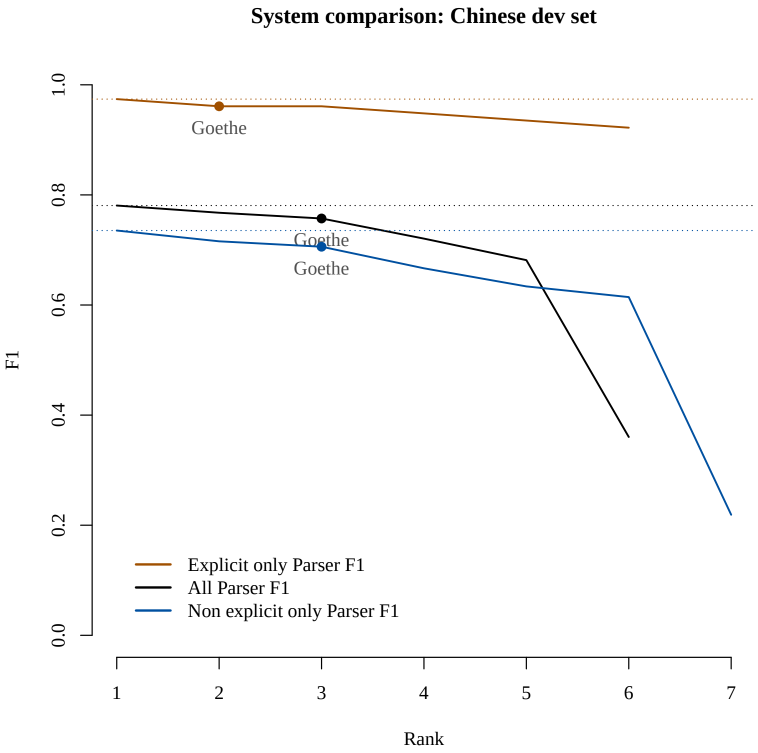


Figure 3.4: English parser performances on the PDTB dev and test sets



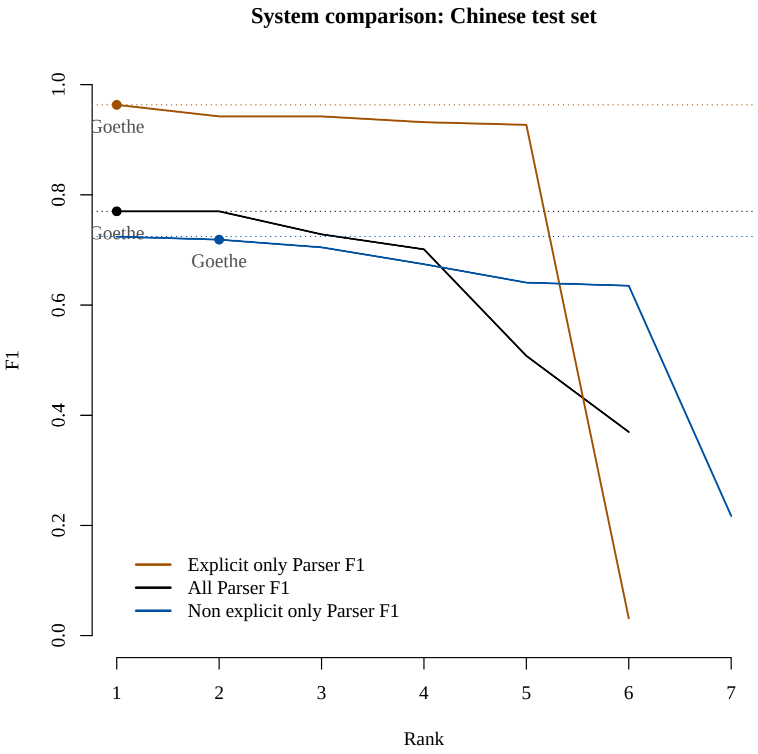


Figure 3.5: Chinese parser performances on the CDTB dev and test sets

sense classification (blue lines) is obviously much harder than the explicit counterpart (red lines). More precisely, the discrepancies between the best test set performances on explicit vs. implicit relations range between up to 47% for English and 25% for Chinese. The combined parser scores usually range in between the two.

The major findings can be summarized as follows: our proposed framework for implicit relations works well on both data sets and has proven to be language-independent. It ranks 3/16 and 3/7 on the development sets, and 4/16 and 2/7 on the test sets, respectively. This indicates that the implicit sense parser is able to generalize well over unseen data of the same genre, is highly competitive with other systems and can even demonstrate an improved performance on the independent Chinese test data. More importantly, according to a χ^2 test, there is no statistically significant difference between the performance of the best system on the CDTB's test set (Wang and Lan, 2016) with 72.4% accuracy, and our proposed method with 71.9% correctly classified test instances, χ^2 (1, N = 455) = 0.024, $p > .05$. However, there is a performance improvement over the third best system with only 67.4%, thus making our proposed framework a powerful, state-of-the-art implicit sense classifier. An overview of the performances of the individual systems for the non-explicit sense classification subtask are given in Tables 3.2 and 3.3. Note that not all systems attempted both supplementary tasks for the two languages.[11].

Penn Discourse Treebank Test Set

Rank	System	% accuracy
1	Wang and Lan (2016)	40.91
2	Mihaylov and Frank (2016a)	39.19
3	Qin et al. (2016)	38.20
4	**Schenk et al. (2016)**	**37.61**
5	Georgia Tech (no pub)	34.95
6	Pacheco et al. (2016)	34.45

Table 3.2: Non-explicit parser scores of the six best-performing participating systems on the official PDTB test set

Interestingly, the systems that outperformed our approach on the English test set all used some sort of convolutional neural network architecture (Wang and Lan, 2016; Qin et al., 2016), which again emphasizes the great importance of upcoming, resource-lean neural methods. One of the best systems on that subtask by Mihaylov and Frank (2016a) is inspired by a promising linguistic add-on to the general CNN architectures and demonstrates small further improvements.

[11] A detailed list of all results is given in Xue et al. (2016) and available from `http://www.cs.brandeis.edu/~clp/conll16st/results.html`

Chinese Discourse Treebank Test Set

Rank	System	% accuracy
1	Wang and Lan (2016)	72.42
2	**Schenk et al. (2016)**	**71.87**
3	Qin et al. (2016)	67.41
4	Weiss and Bajec (2016)	64.07
5	Weiss and Bajec (2016)	63.51
6	Jian et al. (2016)	21.73

Table 3.3: Non-explicit parser scores of the six best-performing participating systems on the official CDTB test set

3.3.3.2 Explicit Relations

Concerning the explicit parser scores, there is clearly an upper bound reached, especially on the English PDTB test set (cf. Figure 3.4), whose performances among all systems range around 90% on the test set. Our system performs well here and ranks 5/16 and 2/16, respectively, with 90.1% accuracy, and is the 4th best system overall on the PDBT.

A similar trend can be observed for the Chinese Discourse Treebank (cf. Figure 3.5), in which our proposed explicit classifier represents the state-of-the-art on this dataset (96.3% vs. 94.2% for the two second-ranked systems). Here, our system is able to outperform all other approaches, which is in fact also true for the overall parser performance. Out of all relation pairs, 77.0% can be correctly labeled by our system with respect to the gold standard.

3.3.3.3 Blind Evaluation Data

Despite the good performance of our method and its ability to generalize well on the same textual genre of Wall Street Journal texts—for both PDTB and CDTB—it should be noted that accuracy scores by all systems (including ours) significantly dropped on the blind evaluation data (cf. Figure 3.6). The blind test set has been designed specifically for the shared task and has been annotated in accordance with the PDTB guidelines, but represents—to a certain extent—out-of-domain texts. The reason for this performance drop can be seen in the less homogeneous (newswire) data, as opposed to the PDTB texts and similar issues supporting this claim have been brought up by Wang and Lan (2016), as well. Moreover, out-of-vocabulary words could play a potential role here, which, as a result, has a negative effect on vector pre-training. Still, this blind set represents a useful resource for further optimization of the existent parsers and should provide helpful insights into domain-adaptation procedures for (implicit) discourse parsing.

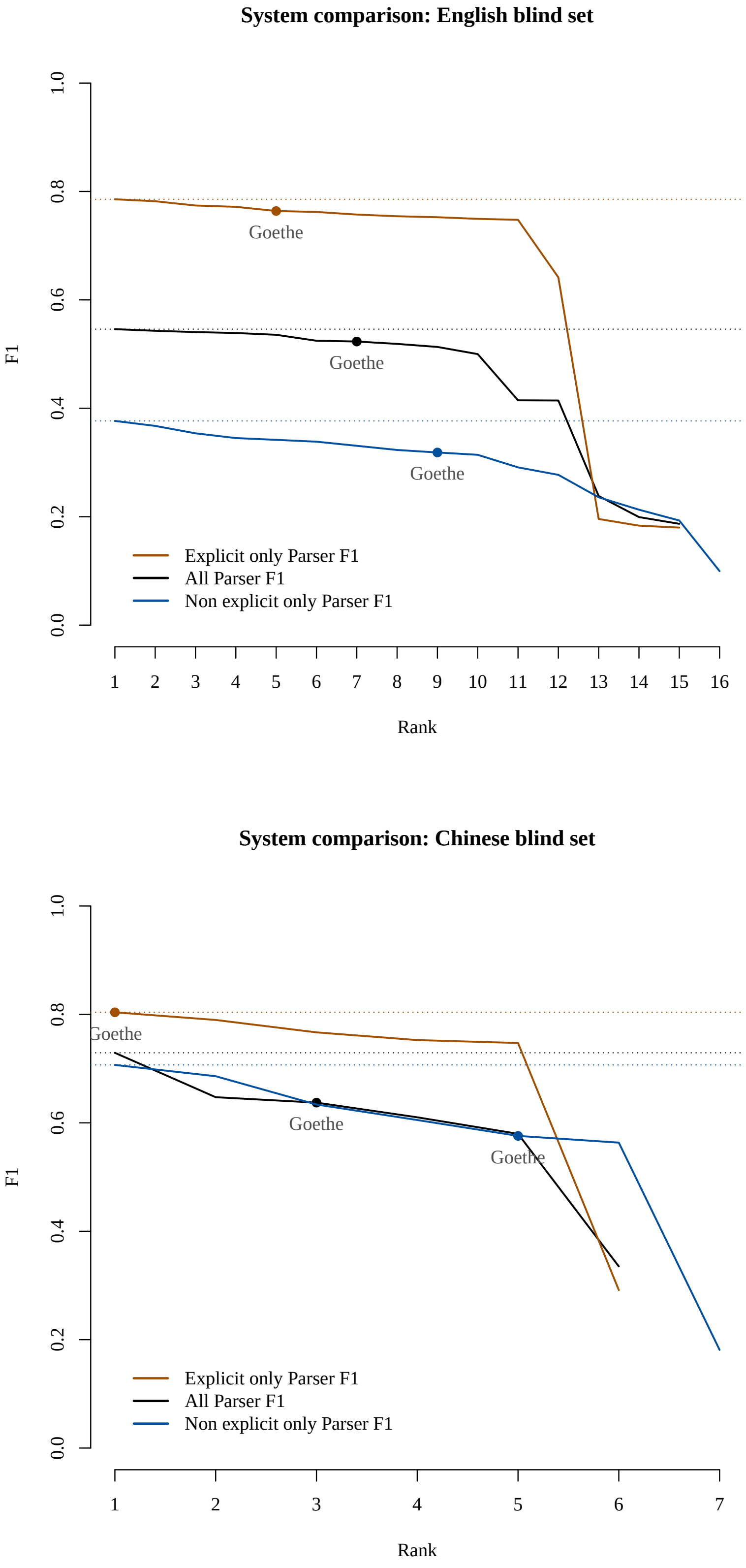


Figure 3.6: Parser performances on the English/Chinese blind (newswire) test sets

3.3.4 Contribution of Network-Specific Factors

In this evaluation, it remains to be discussed to what extent the network-specific parameters can contribute towards the overall parser performance of our suggested model. On the one hand, the max-pooling strategies work well for the convolutional neural networks proposed by Wang and Lan (2016) or Qin et al. (2016). On the other hand, in the context of our proposed framework, for example, we found that both *average* and *pointwise multiplication* of argument matrices performed better compared to either of the two in isolation. Using the shallow convolution operations of Zhang et al. (2015) (*min*, *max* and *average* which worked reasonably well for 4-way classification) had a negative effect on the parser performance for the more fine-grained label set of the shared task. Also, we could verify that pre-training word embeddings for English and Chinese texts effectively improved their predictive power in the sense labeling task. An overview of network-specific factors and their impact on the performance is given in Table 3.4. The table shows their effect (when hyperparameters are kept fix) compared to a baseline in which one of the specific features is disabled. Interestingly, unsupervised pre-training turns out to be the key concept of our proposed method in that it reaches accuracy improvements on the development sets for English and Chinese between 3.2 up to 4.0%. Also, the network composition operations, as well as the integration of syntactic dependency information into argument representations are fruitful add-ons to the global architecture proposed in our work.

Network Feature	Performance contribution of feature (in % acc.)
unsupervised pre-training	+ 4.0
avg + mult vector-composition	+ 2.9
dependency depth-weighting	+ 1.5

Table 3.4: Contribution of individual network-specific factors towards the classifier performance on implicit discourse relation classification (measured in % accuracy on all implicit relations in the development sets of the PDTB and CDTB).

3.4 Summary

This chapter has presented a lightweight and resource-lean architecture based on a feedforward neural network for implicit sense classification in the style of the Penn Discourse Treebank and Chinese Discourse Treebank. The network takes as input two argument spans and builds upon the distributed representations of their containing words. The decision on a suitable label for the relation that holds between the two arguments is solely based on these embeddings. The novelty of the proposed method can be seen in the way argument representations are composed to form a compact, standalone representation of their underlying semantic structure. Compared to prior attempts to the same task of classifying shallow implicit discourse relations using a subtype of convolutional neural networks, our proposed feedforward architecture is generic and structurally simpler in design, which makes it easy to tune and optimize, and thus demonstrates innovative improvements over previous works. It has been shown that our neural model can further benefit from small, carefully chosen add-ons in the form of linguistically inspired information, which is by no means a contradiction to the claim of being a resource-lean architecture in the first place, cf. Mihaylov and Frank (2016a) who have explored similar strategies along these lines. Overall, the main features of the system can be summarized as follows:

1. Word embeddings are used as the only source of information.
   - They are fine-tuned towards the data at hand through unsupervised pre-training.
   - Weighted syntactic dependencies are incorporated into argument representations.
2. Compositional argument representations are structurally simpler than in prior work.
3. The proposed model can be successfully applied to the fine-grained PDTB/ CDTB label inventory and performs excellent beyond class level with more than only four class labels.

The presented system has been evaluated in the context of the CoNLL 2016 Shared Task on Shallow Discourse Parsing (Xue et al., 2016). Overall, the parser performance is highly competitive with the best-performing systems for English and Chinese implicit relations (and one of its subcomponents represents the state-of-the-art for Chinese explicit senses). The framework has been shown to be language-independent and its setup does not require the use of handcrafted (rich) linguistic features.

Overall, it can be observed that, for implicit discourse sense classification, neural architectures, including our own proposed method, have superseded traditional, feature-rich approaches, resulting in improved performance among the majority of systems on standard evaluation sets. Despite their good performance on that task, one principle drawback, however—which is shared by all feedforward variants—is that they are largely *word order agnostic,* i.e. i.) the information within the two single arguments is first treated *separately* as if arguments were unrelated and ii.) their final composition is then simply rendered as a "bag-of-embeddings". Arguably, the obvious advantage of having a highly generic feedforward model whose core basis is a single condensed and pooled representation of arguments raises the question of whether this concept might not be an over-simplification in the global modeling procedure: in fact, human cognitive processes which happen during discourse processing would suggest that reading and progressively understanding text (by scanning it from left to right or right to left) is rather a linear and *sequential* procedure, instead of a static snap-shot of the content of the involved discourse units. It has been shown that this sequential process evokes certain expectations on discourse structure and affects our interpretation as we encounter words one at a time, cf. Croft and Cruse (2004); Kehler and Rohde (2017).

In the following, we specifically address this question and draw upon the few emerging works on *recurrent* neural networks for sequential discourse relation modeling which we have introduced in Section 2.2.4.3. These recurrent variants have been proposed to account for this type of word order information in argument spans, by relaxing the assumption of having fixed argument representations ahead of the learning step, in which words and their order of appearance are completely unrelated. In the next chapter, we introduce a refined version over the feedforward topology described here. We further pursue the idea of a generic, language-independent and resource-lean architecture and demonstrate that the improved model can be successfully applied to the recognition of Chinese implicit discourse relations. The resulting network demonstrates further performance gains on the CDTB data set and will be elaborated on in the next chapter.

Software: An implementation of the *Frankfurt Shallow Discourse Parser* outlined in this chapter is available online from this URL: `https://github.com/acoli-repo/shallow-discourse-parser`.

Chapter 4

Sequential Parsing with a Recurrent Network

4.1 Motivation

The previous chapter has presented a resource-lean feedforward neural network model for the recognition of implicit discourse relations in the style of the Penn Discourse Treebank and the Chinese Discourse Treebank. The model performed very successfully in the CoNLL 2016 shared task on Shallow Discourse Parsing (Xue et al., 2016). Strikingly, the majority of participating teams, with only one notable exception by Weiss and Bajec (2016), came up with a variant of feedforward topology for implicit sense classification. In fact, the best systems were all of the type of convolutional neural network or were vanilla feedforward nets (Wang and Lan, 2016; Schenk et al., 2016; Qin et al., 2016). Two issues, however, are shared by all of these architectures: First, during the feature construction process, these feedforward networks initially analyze discourse units *separately*, i.e. the first argument is treated in a completely unrelated manner from the second argument (see also Section 3.2). Second, after the composition into a joint discourse unit, the resulting component is completely word order agnostic, i.e. the original sequential appearance of the individual words in both arguments (and the order of the two arguments themselves) is irrevocably overwritten.

Interestingly, there has been emerging research on promising approaches with *recurrent* neural networks which also enable *sequential* modeling of textual information. These networks are much harder to train than feedforward nets (Pascanu et al., 2013), but have yielded promising results in domains such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2014), speech recognition (Sak et al., 2015) or word/character prediction.[1] Certainly, the feedforward systems in the shared task on discourse parsing were developed under certain time limitations which made it almost impossible for participants to come up with a thoroughly-tuned recurrent network which at the same time performs among the state-of-

[1] Cf. Andrej Karpathy's blog on *The Unreasonable Effectiveness of Recurrent Neural Networks* `http://karpathy.github.io/2015/05/21/rnn-effectiveness/`, accessed May 2017.

the-art. On a related note, Rutherford et al. (2016) have shown in an experiment that feedforward topologies generally outperform their recurrent counterparts on the discourse modeling task. Contrary to these assumptions, we do, however, believe that recurrent neural networks can indeed hold a lot of promise for the recognition of implicit discourse relations, among others because their mode of operation can be motivated from a cognitive perspective: it seems plausible that during human reading comprehension semantic content is processed or scanned in a left-to-right manner (say, for example, for English). In order to infer an appropriate discourse relation for any given pair of sentences, psycholinguistic as well as computational experimental studies have suggested that it seems more likely for text understanding to be explained as a progressive, *sequential* modeling procedure (directly relating both arguments) instead of a single and static one-shot inspection of the content of the involved discourse units, cf. Van Dijk (1997); Croft and Cruse (2004) or the repeated reading experiments by Liu and Li (2016) described in Section 2.2.4.3. We thus conjecture that feedforward networks and their "bag-of-embeddings" representation are an oversimplification of the inherent complexity of modeling discourse structure.

Building on the motivation for a sequential assessment of discourse structure, in this chapter, we introduce a recurrent neural network as a refinement over the feedforward system from the previous chapter, which we apply to the recognition of Chinese implicit discourse relations. The reason for this is that most systems have initially been designed for the English Penn Discourse Treebank and involve complex, task-specific architectures (Liu and Li, 2016), while discourse modeling techniques for Mandarin Chinese have received very little attention in the literature and are still seriously underrepresented in terms of publicly available systems. What is more, over 80% of all words in Chinese discourse relations are implicit—compared to only 52% in English (Zhou and Xue, 2012). Aside from that, the organizers of the CoNLL 2016 Shared Task on Shallow Discourse Parsing have released the Chinese Discourse Treebank relations (Zhou and Xue, 2012, CDTB) for the first time, which has recently made the data set particularly attractive to work on with regards to novel experimentation, and thus for potential room for improvement on the classification scores.

In the following sections, the first attention-based *recurrent* neural sense classifier, specifically developed for Chinese implicit discourse relations is presented. Inspired by Zhou et al. (2016), the system is a practical adaptation of the recent advances in relation modeling extended by a novel sampling scheme (which will be described in Section 4.2). Contrary to previous assertions by Rutherford et al. (2016), the introduced model demonstrates superior performance over traditional bag-of-words approaches with feedforward networks by treating discourse arguments as a *joint sequence*. Following the paradigm of previous evaluations, this novel method is assessed within an independent framework performing very well beyond standard class-level predictions, achieving state-of-the-art accuracy on the CDTB test set. The model comes with a special *attention mechanism* (which is inherent to the nature of our recurrent network and which is outlined in Section 4.2), providing means to highlight those specific parts of an input sequence that are relevant for the classification decision, and thus, may enable a better understanding of the implicit discourse parsing problem. In this context, we revisit entity-based coherence relations and point out structural idiosyncrasies as well as distinct semantic differences with regards to standard conjoining relations. This is particularly relevant as some of the related research on implicit discourse classification has per default simply consolidated both entity relations as well as expansion senses and merged them into the same class (Cianflone and Kosseim, 2018).

Compared to other resource-lean methods in that domain, the proposed network architecture is flexible and largely language-independent, as well, as it operates only on word embeddings. It stands out due to its structural simplicity and provides a solid ground for further development towards other textual domains. In what follows, we introduce design principles and the network architecture.

4.2 Design Principles & Network Architecture

In this work, we propose the use of an attention-based bidirectional Long Short-Term Memory (Hochreiter and Schmidhuber, 1997, LSTM) network to predict senses of discourse relations for a given argument pair. The model draws upon previous work on LSTM, in particular its bidirectional mode of operation (Graves and Schmidhuber, 2005), attention mechanisms for recurrent models (Bahdanau et al., 2014; Hermann et al., 2015), and the combined use of these techniques for entity relation recognition in annotated sequences (Zhou et al., 2016). More specifically, our model is a flexible recurrent neural network with capabilities to *sequentially* inspect tokens and to highlight which parts of the input sequence are most informative for the discourse relation recognition task, using the weighting provided by the attention mechanism. Unlike in most previous works in which discourse arguments are typically treated as unrelated, pairs of discourse arguments are modeled as a *joint* sequence in our setting. Moreover, our model benefits from a novel sampling scheme which is inspired from the classical literature on discourse analysis, whose details are reported in Section 4.2.1. The system is learned in an end-to-end manner and consists of multiple layers, which are illustrated in Figure 4.1. We describe the individual components hereafter.

First, token sequences are taken as input and special boundary markers (<Arg1>, </Arg1>, <Arg2>, </Arg2>) are inserted into the corresponding positions to inform the model on the start and end points of argument spans. This particular type of modeling allows for a certain flexibility because it will in principle become straightforward to extend the span boundaries to include additional context, for instance, to the left or right of a current span. However, in our experiments on implicit arguments, we restrict the analysis to the gold-annotated spans and the respective tokens. Note again that, unlike classical approaches to implicit discourse parsing especially with feedforward neural networks, our approach treats Arg1-Arg2 pairs as a *joint* sequence. In our novel setting, there is no need to first compute an intermediate representation of both arguments separately.

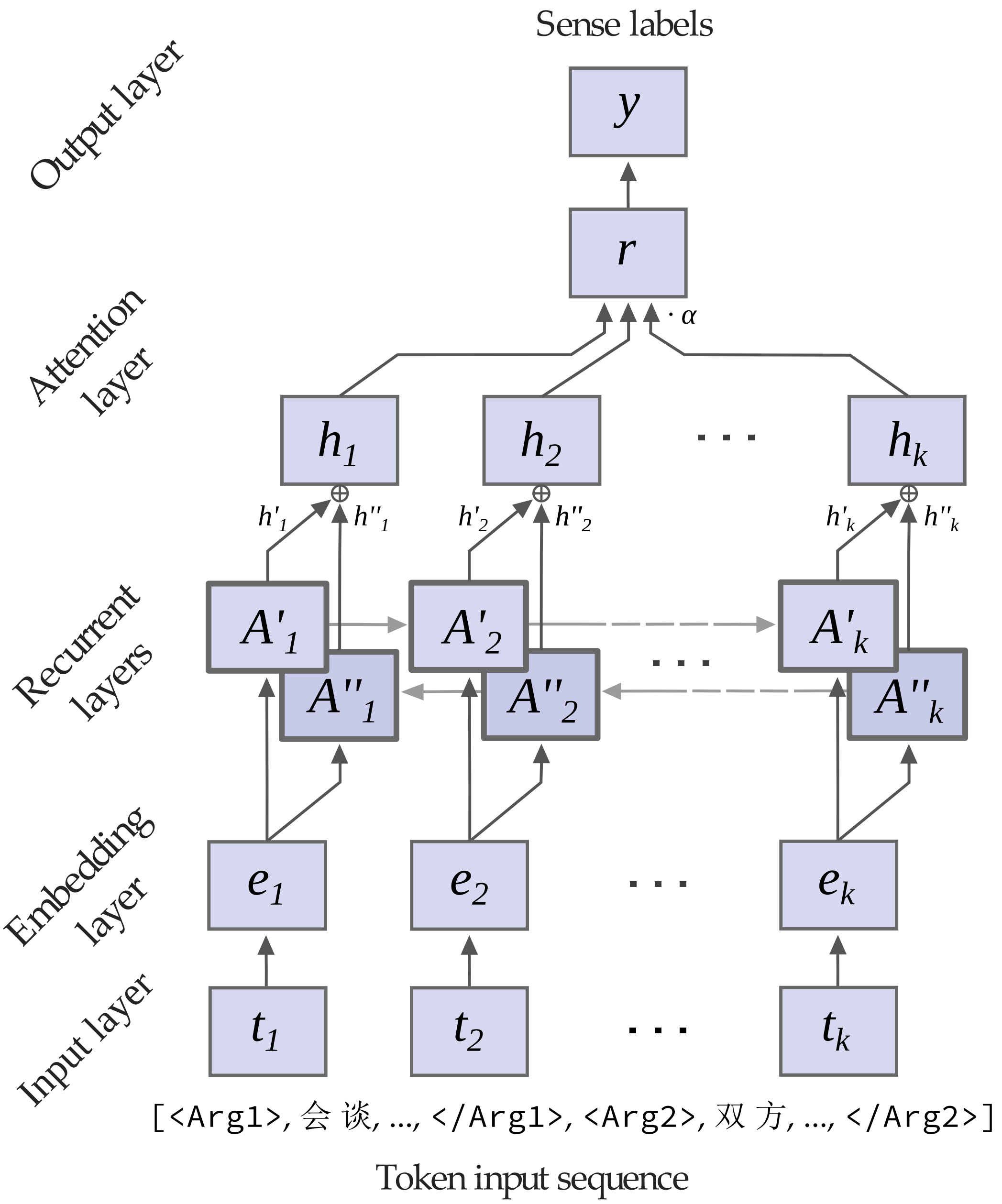


Figure 4.1: Modeling Chinese implicit sense relations with an attention-based bidirectional LSTM network. The bottom part of the figure illustrates the input layer which encodes the tokens using special markers for argument boundaries, which are substituted by distributed word representations in the first place (embedding layer). The network is completed by three modular components in the form of a stacked layer for the bidirectional modeling (recurrent), a pre-final attention layer and a final softmax output layer for sense classification.

Second, an input layer encodes tokens using one-hot vector representations (t_i for tokens at positions $i \in [1, k]$), and a subsequent embedding layer provides a dense representation (e_i) to serve as input for the recurrent layers. The embedding layer is initialized using pre-trained word vectors, in our case 300-dimensional Chinese Gigaword vectors (Graff and Chen, 2005).[2] These embeddings are further tuned as the network is trained towards the prediction task. Embeddings for unknown tokens, e.g., markers, are trained by backpropagation only. Note that, tokens, markers and the pre-trained vectors represent the only source of information for the prediction task.

For the recurrent setup, we use a layer of LSTM networks in a bidirectional manner, in order to better capture dependencies between parts of the input sequence by inspection of both left and right-hand-side contexts at each time step. The LSTM holds a state representation as a continuous vector passed to the subsequent time step, and it is capable of modeling long-range dependencies due to its gated memory. The forward (A') and backward (A'') LSTMs traverse the sequence e_i, producing sequences of vectors h'_i and h''_i respectively, which are then summed together (indicated by $\oplus$ in Figure 4.1).

The resulting sequence of vectors h_i is reduced into a single vector and fed to the final softmax output layer in order to classify the sense label y of the discourse relation. This vector may be obtained either as the final vector h produced by an LSTM, or through pooling of all h_i, or by using attention, i.e. as a weighted sum over h_i. While the model may be somewhat more difficult to optimize using attention, it provides the added benefit of interpretability, as the weights highlight to what extent the classifier considers the LSTM state vectors at each token during modeling. This is particularly interesting for discourse parsing, as most previous approaches have provided little support for pinpointing the driving features in each argument span.

Finally, the attention layer contains the trainable vector w (of the same dimensionality as vectors h_i) which is used to dynamically produce a weight vector α over time steps i by:

$$\alpha = softmax(w^T tanh(H))$$

where H is a matrix consisting of vectors h_i. The output layer r is the weighted sum of vectors in H:

$$r = H\alpha^T$$

4.2.1 Partial Argument Sampling

In what follows, we describe a novel sampling technique that involves partial discourse arguments. It serves two purposes: First, it should enlarge the instance space of the typically sparse amount of manually annotated training instances in the CDTB, and second, we aim at improving the predictive performance of our model with the help of this technique. In this particular setting of *partial*

[2] http://www.cs.brandeis.edu/~clp/conll16st/dataset.html

sampling, our recurrent neural network model is trained and validated on input sequences containing both arguments, as well as *single* arguments in isolation. Specifically, a single training instance (a_1, a_2, y), where a_i is the sequence of argument tokens in either the first or second discourse argument and y is the sense label, is expanded to produce a set containing four training instances: $\{(a_1, a_2, y), (a_1, a_2, y), (a_1, y), (a_2, y)\}$. Note that we keep a duplicate of the standard argument pair (a_1, a_2, y) (only in training and development sets) in order to counter-balance their frequencies with regards to the newly introduced single-argument samples. Note again that this proposed partial sampling technique only applies to the training and the validation/development phases. In the assessment of the official evaluation scores on the test set of the CDTB, we refer to the standard argument pairs provided.

The proposed method to include single discourse arguments is inspired and particularly motivated by two lines of research that have their roots in machine learning and theoretical linguistics, respectively. On the one hand, the proposed procedure can be considered a direct encouragement for model to come up with better representations of single arguments in support of deriving representations for arguments in composition, cf. LeCun et al. (2015). This type of data augmentation technique for our model can be an effective supplement in reinforcing its overall robustness. On the other hand, theoretical evidence is provided by the fact that single utterances and expressions can elicit a strong expectation towards a particular type of implicit discourse relation in the following context, for instance, local cues may evoke a cause or an explanation, cf. Asr and Demberg (2015); Rohde and Horton (2010) and the related psycholinguistic study on *implicit causality verbs* (Garvey and Caramazza, 1974). We make direct use of single arguments, because (in their standalone form) they can provide valuable training instances.[3]

4.2.2 Implementational Details

We train the model using fixed-length sequences of 256 tokens with zero padding at the beginning of shorter sequences and truncate longer ones. Each LSTM has a vector dimensionality of 300, matching the embedding size. The model is regularized by 0.5 dropout rate between the layers and weight decay ($2.5e^{-6}$) on the LSTM inputs. We employ Adam optimization (Kingma and Ba, 2014) using the cross-entropy loss function with mini batch size of 80 during 15 epochs.[4]

4.3 Evaluation

In line with the previous evaluations on standard data sets, the recurrent model described in this chapter is evaluated on the CoNLL 2016 shared task data (Xue

[3] We outline a related experiment which draws on the particular features of single discourse units in Chapter 8 of this thesis.

[4] The model is implemented in *Keras* `https://keras.io/`.

Sense Label	Training	Development	Test
CONJUNCTION	5,174	189	228
majority class	(66.3%)	(62.8%)	(64.8%)
EXPANSION	1,188	48	40
`EntRel`	1,099	50	71
CAUSATION	187	10	8
CONTRAST	66	3	1
PURPOSE	56	1	3
CONDITIONAL	26	0	1
TEMPORAL	26	0	0
PROGRESSION	7	0	0
# implicit relations	7,804	301	352

Table 4.1: Distribution of implicit sense labels in the Chinese Discourse Treebank

et al., 2016).[5] It includes the official training, development and test splits of the CDTB. An overview of the label distribution of the implicit discourse senses is provided in Table 4.1. It is noteworthy that, in the Chinese Discourse Treebank, implicit discourse relations are almost *three times as frequent* as their explicit counterparts. Out of these implicit relations, 65% have their arguments within the same sentence, and 25 argument pairs (in the training set) have two sense labels.

In order to comply with previous setups as reported in Rutherford et al. (2016), entity-based coherence relations, `EntRel`, are considered a distinct implicit discourse relation (of same sense), and `AltLex` relations are omitted in the analysis, because they simply appeared too infrequent in order to be statistically relevant. In the evaluation, we focused on the *sense-only* track. In this subtask the discourse gold arguments are already provided and a participating system is supposed to predict the sense label for a given argument pair. We report on the results in Tables 4.2 and 4.3, respectively.

As the figures in Table 4.3 suggest, with our proposed method it is possible to obtain state-of-the-art results evaluated on the test set of the Chinese Discourse Treebank. Our recurrent network correctly labels 257 out of 352 relations which results in a parser performance of 73.01% accuracy. Furthermore, our approach beats the best feedforward system from the shared task in 2016 by Wang and Lan (2016) as well as all other word order-agnostic network approaches. A comparison of the development and test set performance figures in Tables 4.2 and 4.3 suggest that our approach is robust enough for the data at hand and demonstrates its ability to generalize as well to unseen examples.

[5] `http://www.cs.brandeis.edu/~clp/conll16st/`

CDTB Development Set

Rank	System	% accuracy
1	Wang and Lan (2016)	73.53
2	Qin et al. (2016)	71.57
3	Schenk et al. (2016)	70.59
4	Rutherford and Xue (2016)	68.30
5	Weiss and Bajec (2016)	66.67
6	Weiss and Bajec (2016)	61.44
7	Jian et al. (2016)	21.90
	Rönnqvist et al. (2017)	**93.52***

Table 4.2: Parser scores for implicit CDTB relations on the official CoNLL 2016 development set. (*Scores are due to the partial sampling technique and are thus not directly comparable.)

CDTB Test Set

Rank	System	% accuracy
1	Wang and Lan (2016)	72.42
2	Schenk et al. (2016)	71.87
3	Rutherford and Xue (2016)	70.47
4	Qin et al. (2016)	67.41
5	Weiss and Bajec (2016)	64.07
6	Weiss and Bajec (2016)	63.51
7	Jian et al. (2016)	21.73
	Rönnqvist et al. (2017)	**73.01**

Table 4.3: Parser scores for implicit CDTB relations on the official CoNLL 2016 test set

4.3.1 Ablation Study

In order to assess the usefulness of the two key features of our proposed model—the attention mechanism and the partial sampling technique—we performed an ablation study in order to quantitatively assess their contribution towards the overall parser performance. In one experiment, we first constructed a simpler alternative in which we directly forwarded the final LSTM hidden vectors (both h'_k and h''_1) to the output layer. When, like in this case, attention is not at work, this results in an absolute performance decrease of approximately 2.70% on the CDTB test set. The drop is significant according to a Welch two-sample t-test ($p < .001$). Similarly, in a second experiment, we simply trained our recurrent neural network on the standard argument pairs. Without the partial labels the performance again drops, this time by 5.74%. The accompanying test indicates a significant decrease in performance ($p < .001$). Indeed, both methods outlined in this chapter, attention and the partial arguments, represent a substantial contribution towards the overall model performance in order to achieve competitive results.

4.3.2 Performance on the PDTB

In order to test the model's ability to perform as well on other domains and languages, we conducted another side experiment. More precisely, we applied it to the implicit discourse arguments from the English Penn Discourse Treebank. Since this experiment should only serve illustration purposes, the hyperparameters of the model were not optimized, but instead we reused the identical setting that we derived from training on the CDTB. Of course, tuning a model on Chinese data and re-applying the parameters to English is expected to result in suboptimal performance. Nonetheless, an accuracy of 27.09% on the PDTB test set can be measured, which shows that our model has potential to generalize across implicit discourse relations in a cross-domain setting.

4.3.3 A Visualization of the Attention Activity

Part of the motivation of the attention mechanism was related to the fact that the neural model benefits from increased interpretability. In what follows, we demonstrate that attention weights can indeed highlight and pinpoint plausible subcomponents within an implicit discourse relation. Figure 4.2 is a graphical illustration of the model's learned attention weights applied to two particular example relations from the Chinese Discourse Treebank, a CONJUNCTION sense and an entity-based coherence relation, i.e. `EntRel`. The color boxes relate to the individual Chinese characters. High intensities are represented by dark blue color, low intensity by light blue. English (translated) phrases whose structures are closely associated between two arguments are underlined.

In the implicit CONJUNCTION relation in the upper part of Figure 4.2, for example, the weights have a maximum peak roughly at the transition between the argument boundary of the discourse units. This transition indirectly establishes a connection between the semantically related words *understandings* and *agree*.

The majority of `EntRels` in the data set show an opposite trend: towards the end of the relation (typically in the second discourse argument), they exhibit much larger intensities than in `Arg1`. This is an interesting observation and related to the characteristic writing style of newspaper texts (and the newswire domain from which the examples are drawn). Here, entity relations typically establish local discourse coherence by *adding additional information* by reference to the same entity.

Although it should be noted that not all discourse relations (and individual tokens) can be straightforwardly interpreted as in the example outlined here, some general semantically-informed trends are visible, especially when the attention weights are aggregated over the spans separately. We elaborate on details hereafter.

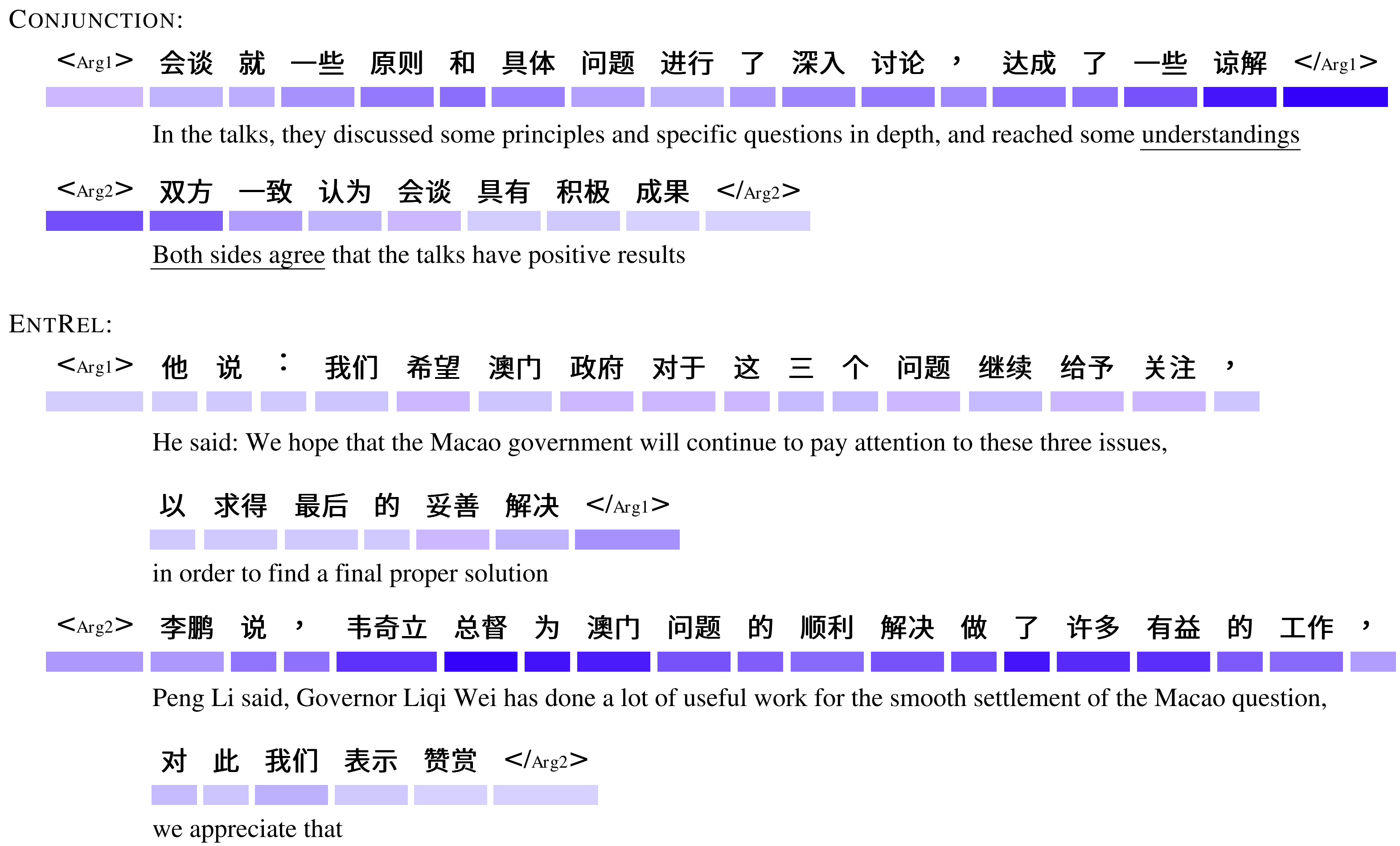


Figure 4.2: Attention weights visualization for two Chinese implicit discourse relations.

Sense	Relation ID	Posterior Prob.	**Max. Attention**
EntRel	287	0.796	0.018
CONJUNCTION	187	0.839	0.014
EntRel	322	0.590	0.013
EntRel	348	0.619	0.011
CONJUNCTION	201	0.667	0.011
EntRel	108	0.753	0.011
CONJUNCTION	176	0.506	0.011
CONJUNCTION	206	0.962	0.010
CONJUNCTION	97	0.865	0.010
EntRel	173	0.625	0.010

Table 4.4: Top 10 relations of maximum attention-weighted argument tokens from the CDTB test set

A closer inspection of the discourse relations of type CONJUNCTION and EntRel (which together make up the vast majority of implicit discourse relations in the CDTB with $\approx$85% of all relations) reveal some interesting contrastive properties. For visualization purposes, we concentrate on those relations in the test set which the recurrent system has classified correctly. These relations are first ranked according to their maximum attention activity by considering the contribution of each token's individual attention weight. Table 4.4 shows an overview of the resulting ten most "active" discourse relations of type CONJUNCTION and of type EntRel, along with their posterior probabilities for the classification of the correct sense and their relation IDs in the test set.

Figure 4.3 (schematically) and Figures 4.4, 4.5 (with details) visualize the attention activities for each of the ten discourse relations. The first and second arguments are graphically distinguished—Arg1 in yellow, Arg2 in gray. The first column contains CONJUNCTION senses, the second entity relations. The y-axis indicates the strength of attention activity. Each bar represents one token and its associated attention weight. Higher bars represent larger attention weights. Argument transitions are indicated by the marker boundary </ARG1><ARG2> which are highlighted in red.

Regarding the visualizations of attention activities, a few general observations can be made: All attention weights are steadily increasing from the beginning of the first argument until a first peak is reached which is always localized within the same first argument. Then, towards the argument boundary in both relation types attention activities are constantly decreasing. Interestingly, within the second argument there is a different pattern observable between the two relation types: For the CONJUNCTION senses, the attention weights remain roughly the same (or show only a marginal increase), while second arguments of EntRels behave completely different. Here, for all five relations, Arg2 exhibits larger boosts in attention

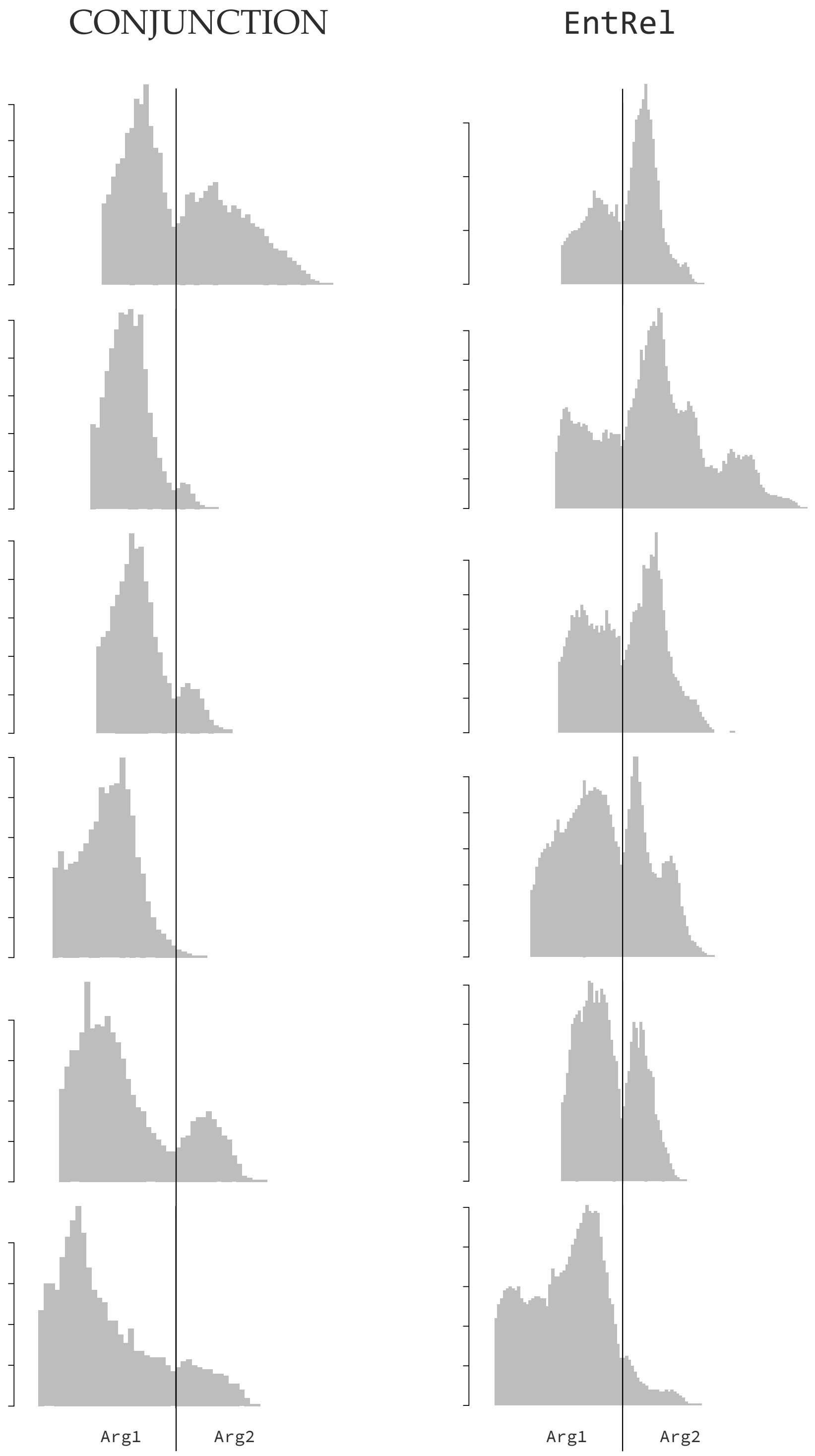
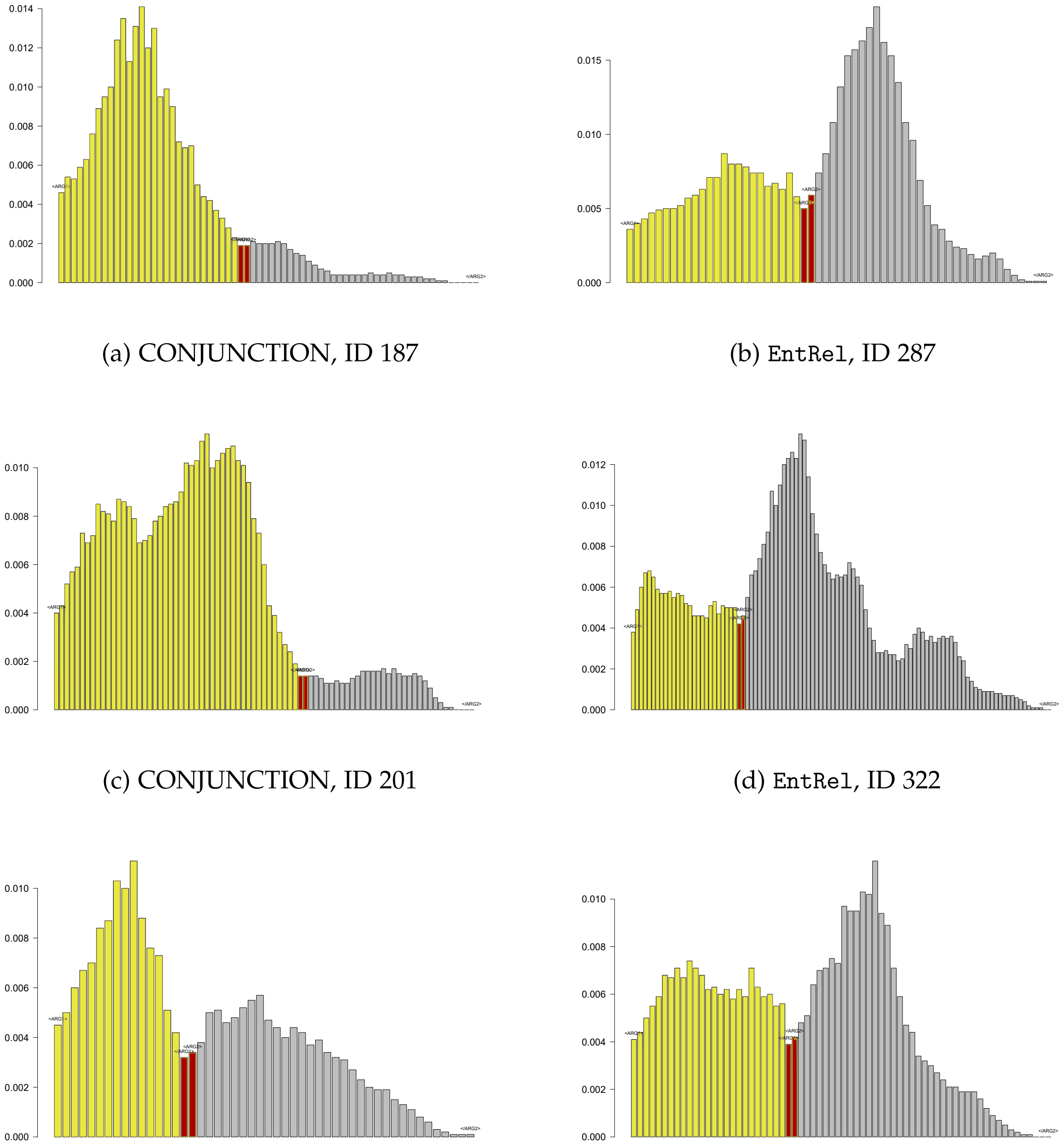


Figure 4.3: Schematic visualization of attention activity for ten most active CDTB discourse relations



(a) CONJUNCTION, ID 187

(b) `EntRel`, ID 287

(c) CONJUNCTION, ID 201

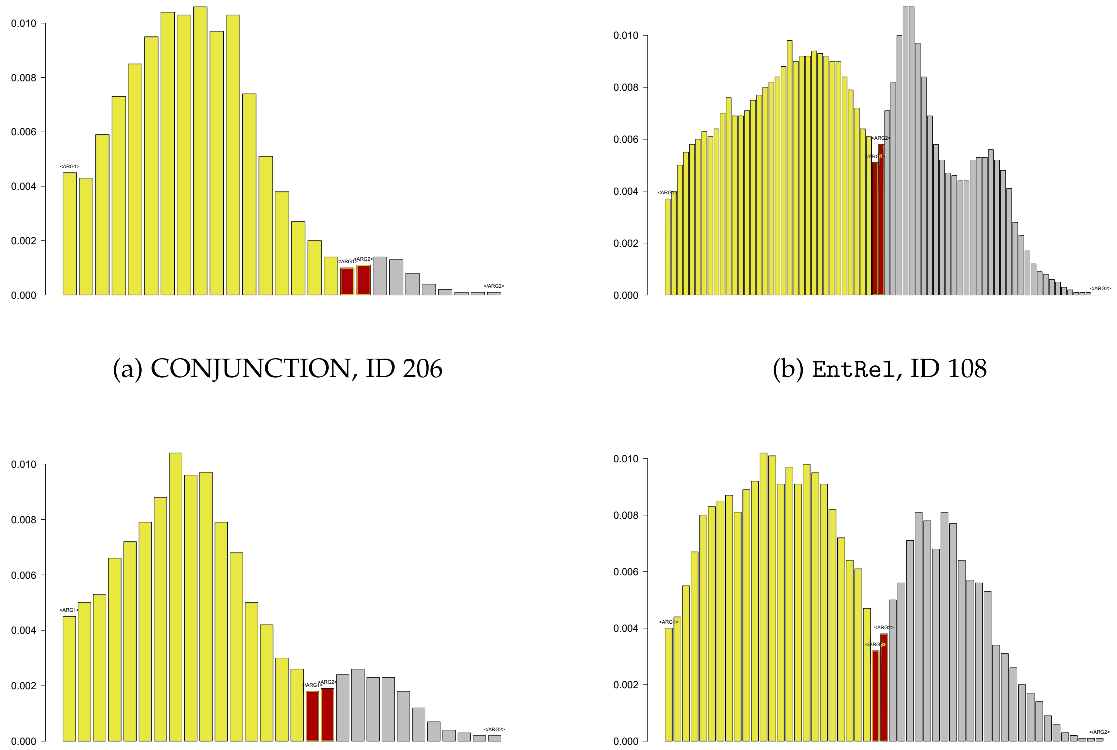
(d) `EntRel`, ID 322

(e) CONJUNCTION, ID 176

(f) `EntRel`, ID 348

Figure 4.4: Attention activity of discourse relations (`Arg1` in yellow, `Arg2` in gray) for CONJUNCTION relations (left column) and `EntRels` (right column)

activity—in four of five cases larger than in the respective `Arg1`s. We believe that a possible explanation for this trend can be attributed to the characteristic properties of these relations which—by reference to the same entity—elaborate more thoroughly on further aspects by introducing semantically important *additional* content on top of the present information given the first argument (responsible for the "boost"). In contrast, CONJUNCTION senses can represent arbitrary in-



(a) CONJUNCTION, ID 206

(b) `EntRel`, ID 108

(c) CONJUNCTION, ID 97

(d) `EntRel`, ID 173

Figure 4.5: Attention activity of discourse relations (`Arg1` in yellow, `Arg2` in gray) for CONJUNCTION relations (left column) and `EntRels` (right column)

formation, oftentimes but not necessarily, mentioning a (named) entity in either of the two arguments. Typical CONJUNCTION relations add "minor" information along with the second argument as, for instance, in the following example (8):

(8) `Arg1`: 中国红十字会根据新的实践要求，开展了卓有成效的人道主义工作 *The Chinese Red Cross has carried out fruitful humanitarian work according to the new practice requirements*

`Arg2`: 在社会上产生了广泛而良好的影响 *which had broad and positive impacts on society.*

Implicit discourse relation[6] / sense: CONJUNCTION

To summarize, entity-based coherence exhibits two impulses in both arguments, while conjoining relations have characteristic peaks only in argument one. Note that this trend for boosts in second arguments for entity relations, and rather decreasing activity for CONJUNCTION is shared by all ten relations, except for one

[6] CDTB Document ID `chtb_0279`

with ID 176. However, note also that this is the argument pair with lowest posterior probability (only 50.6%) which indicates that the structure of this discourse relation is hard to unambiguously classify into one unique class. In fact, a manual inspection of the content of the second argument reveals that two named entities in the form of *Macao* and *China* are indirectly related to `Arg1`, even though the overall relation is of type CONJUNCTION.

4.4 Summary

This chapter has presented the first attention-based recurrent neural sense labeler specifically developed for Chinese implicit discourse relations. Its characteristic network topology contrasts with the feedforward architecture described in the previous chapter by one important aspect: the recurrent system is primarily motivated in terms of a *sequential and joint* modeling of discourse units, as opposed to operating on static and fixed argument representations. The theoretical foundations of the computational model are supported by experiments in cognitive linguistics, which suggest that discourse phenomena can best be explained by a linear scanning process of textual information. This assumption has been realized by the bidirectional mode of the LSTM components in the recurrent network. Also, from a classification perspective, the ability of the recurrent model to associate discourse units sequentially and jointly has been shown to be profitable, resulting in state-of-the-art accuracies on the Chinese Discourse Treebank. In fact, the proposed network model has demonstrated superior performance over the traditional word order-agnostic approaches, beating all feedforward topologies from the first edition of the CoNLL shared task on Shallow Discourse Parsing. These results contrast with previous observations by Rutherford et al. (2016), who claimed that feedforward systems typically perform relatively better than their recurrent counterparts on the task of implicit sense classification.

Two additional features make the model distinct from previous works on implicit discourse parsing: First, attention is incorporated, which has demonstrated an additional boost in performance over a standard bidirectional LSTM encoding of the input sequence. Moreover, the attention mechanism and the induced attention weights come with the added benefit of insightful observations into the inner workings of the model. This has been particularly useful in pinpointing those words in an argument span which the model considers important to arrive at the final classification decision on a sense label and also for a deeper modeling of distinct discourse relations. In this context, we have revisited two closely related types of discourse relations—`EntRel` and CONJUNCTION—for which the previous literature on implicit discourse parsing has typically not made a distinction. We have shown that the learned attention weights can reflect interesting, idiosyncratic properties between local entity-based discourse coherence and ordinary CONJUNCTION (EXPANSION) relations. To summarize, even though deep recurrent neural networks are often considered opaque "black box" architectures, an associated attention mechanism makes these models more transparent and valuable for real application scenarios.

On top of these properties which are specific to the network structure, this chapter has introduced a novel training technique, termed *partial argument sampling*. To the best of our knowledge, this particular method has not been used before in the context of discourse relation recognition. Specifically, during partial sampling, argument pairs—as well as *single* arguments in isolation—are fed into the model for parameter estimation. The reason to use single argument training instances is linguistically inspired, as previous research has demonstrated tendencies towards the expectation of a certain discourse relation, given only one discourse unit (cf. Section 4.2.1 and Chapter 8 for a correlation study investigating this effect in closer detail). Along with an overall increased space of training instances, this technique has yielded a slight increase in performance of the presented model.

Overall, it can be summarized that the network outlined in this section is structurally self-contained and could be easily extended in various ways, for instance, by increasing the span size for arguments (in either direction) to see if additional augmented context could support implicit relation recognition[7], or by unsupervised data acquisition from external resources. Also, regardless of a specific classification algorithm (be it a feedforward or a recurrent system), a practical tool for discourse relation recognition seems viable and many downstream applications benefit from the proper extraction of discourse relations, which has been demonstrated already very successfully by numerous previous works.

Software: The code for the recurrent neural network parser outlined in this chapter is publicly available at `https://github.com/sronnqvist/discourse-ablstm` and from `http://www.acoli.informatik.uni-frankfurt.de/resources.html`.

[7] Personal communication with Fatemeh Torabi Asr during her visit in the Frankfurt ACoLi lab.

Summary

This part of the dissertation has been primarily concerned with automated methods to recognize senses in implicit discourse relations.

Chapter 2 has laid the theoretical foundations for computational discourse processing in general. In particular, four discourse frameworks were outlined, which draw on distinct aspects for modeling the coherent structure in texts. For example, we have introduced Centering and the associated transition types as a theory to model local text coherence (Section 2.1.1). We have briefly sketched the framework of Rhetorical Structure Theory (Section 2.1.2), which explains the interaction of discourse units in a hierarchical, tree-structured fashion. The concept behind (Segmented) Discourse Representation Theory (Section 2.1.3) makes it possible to address fine-grained and more elaborate linguistic phenomena and is grounded in formal semantics. Finally, the Penn and the Chinese Discourse Treebank have been introduced in Section 2.1.4. These two resources have provided the basis for all ensuing methods presented in this part of the thesis. From a computational perspective, a main benefit can be seen in the convenient, shallow manner in which discourse units are shaped within the framework, which in turn has prompted a wide range of prior works on automated discourse parsing to focus their systems on these two corpora.

Following a quantitative exposition on the distribution of relations and sense labels in the Penn and Chinese Discourse Treebank, Section 2.2 has motivated the need for automated approaches to handle implicit discourse relations, i.e. those specific argument pairs which lack the presence of an informative connective. We have seen that, in particular in Chinese texts, implicit discourse relations account for almost two thirds of all relations which makes them especially relevant in the context of NLP systems. Any practical downstream application, e.g., a QA system or a text summarizer, needs to rely on the accurate discourse analysis of these implicit relations.

Concerning this matter, we have provided a literature review of prior methods on the challenging task of implicit discourse processing in the PDTB framework, with a focus on sense label classification. First approaches have initially integrated word-pair features (Section 2.2.1), which have proven to serve as a solid baseline but which typically suffer from sparsity issues. Section 2.2.2 has introduced resource-intensive classifiers. These methods naturally rely on rich linguistic features, e.g., in the form of hand-crafted lexicons or external knowledge sources. These perform well for a specific domain but implicate cost and flexibility issues.

Moreover, mildly resource-intensive methods (Section 2.2.3) substitute the sparse and rich features by abstract representations, most notably by word embeddings, and denote a promising shift towards greater flexibility. Finally, distributed word representations in combination with neural network-based classifiers were introduced under the notion of resource-lean parsing (Section 2.2.4). These methods constitute a novel learning paradigm and represent the current state of the art in parsing implicit discourse relations (cf. Section 2.2.4.3).

Based on the recent success of deep learning for implicit relation recognition, we have outlined two resource-learn approaches: Chapter 3 has demonstrated the applicability of a feedforward neural network to implicit sense classification. The system treats argument pairs as a bag of embeddings and by expedient composition derives at a representation which is suitable to capture the latent syntactic and semantic properties of the discourse relation. The structures benefit from syntactic dependency information and unsupervised pre-training of the involved embeddings. The discourse parser demonstrated substantial performance improvements over participating systems in a shared task. As a core feature of our proposed system, we highlighted a particular weighing scheme which is based on the syntactic path from the root node to a specific word in a discourse unit. We want to point out that other types of normalization and aggregations are conceivable in this regard, and elaborate on potential alternatives in the final chapter of this thesis. Moreover, it should be noted that, from a structural point of view, the presented architecture is a bag-of-embeddings system, i.e. the tokens (and crucially the discourse arguments themselves) are treated irrespectively of their order of appearance in the texts, which seems counterintuitive from a human cognition perspective, and which has prompted us to further refine our work in the ensuing chapter.

An improvement upon the feedforward system was sketched in Chapter 4. Since from a cognitive perspective, it seems plausible that processing discourse units along with their interpretation is a rather sequential process, we have presented a recurrent neural network model which is able to encode this additional linearity constraint. Technically, the system is an attention-based bidirectional LSTM and comes with the added benefit of interpretability of the derived classification result as the attention weights can be used as indicators to pinpoint important word interactions in a specific discourse relation. The architecture benefits from a novel method inspired from psycholinguistics in which single argument pairs (which can themselves be highly indicative of a discourse sense) are fed into the system as training instances. Overall, our proposed model demonstrates state-of-the art performance on the Chinese Discourse Treebank. An additional remark should concerns the sequential assessment of tokens in the discourse units, and the shallow argument pairs that we obtained from the Chinese Discourse Treebank. Our method has proven to be particularly flexible because the start end end markers of the discourse units, in fact, explicitly signal the argument boundaries in terms of beginning and end positions, but these need not necessarily be restricted to the positions imposed by the instances from the CDTB. In fact, it might seem plausible that additional context could be helpful in determining the sense relation that holds between to extended argument spans, whose experimental realization is straightforward with our proposed architecture. Our future research is dedicated a relaxation of this fixed span assumption.

As opposed to the global implicit relations between sentences and clauses that we have presented in this part, the next part of this thesis is dedicated to local implicit information which is naturally evoked within sentences with links typically associated between single words or phrases. Sometimes, local elements, such as the subject or the object of a sentence, can be locally unexpressed but it is clear from the context that they can be related to antecedents or postcedents in the discourse. One example of this type are zero anaphora in Mandarin Chinese (Tao, 1996). It has been shown that anaphora resolution in general is constrained by text coherence and discourse relations (Polanyi, 1988; Grosz et al., 1995; Knott et al., 2001; Kehler and Rohde, 2017, inter alia). We elaborate on details on the next part of this thesis, which is dedicated to *implicit semantic role labeling*, i.e. the detection of local implicit information, and the linking to appropriate antecedents in the discourse.

Part III

Implicit Semantic Role Labeling

Chapter 5

Theories, Frameworks & Computational Approaches

As we have seen in the previous part of this thesis, an implicit discourse relation can link two textual spans even when no explicit cue (i.e. a discourse connective) is present. This property holds for many sentence or clause pairs within a document and can be attributed to the coherent structure of a text. For instance, a causal relation holds between two subsequent sentences in which the latter describes the result of the former. However, implicit links cannot only be observed cross-sententially between long and extended descriptions. Crucially, they are as well evoked on an atomic level, for instance, *by words or phrases within clauses or sentences*. Notably, the implicit counterparts of such a relation can be other words or phrases—sometimes found in the immediate periphery (of the same sentence), sometimes also in faraway parts of the document. As a general basis for investigating word or phrase-level implicit information current research typically follows the traditional literature on the meaning of predicates in terms of *argument structure* and *semantic roles* (Levin, 2013, 2014).

The notion of argument structure is closely related to the valency of predicates and the concept of subcategorization (Chomsky, 1965). More precisely, the argument structure of a lexical item specifies its syntactic realization, for instance, a verbal predicate together with its associated complements as realized by (phrasal) constituents in a text. Typically, arguments are considered core elements of a lexical item which are distinguished from co-occurring adjuncts, whose function is to only act as a modifier in the environment of the predicate. The predicate-argument structure also determines the number of arguments which varies across individual lexical items. The argument structure is mostly applicable to verbal predicates but also other parts-of-speech such as nouns, adjectives and even prepositions can be modeled in this fashion.

Moreover, the thematic *relationship* between lexical items and (predominantly) nominal arguments is expressed by semantic roles (Bruce and Moser, 1992; Levin and Hovav, 2005).[1] The concept behind semantic roles is best illustrated with an

[1] The interested reader is referred to the early work on *theta roles* described in Fillmore (1968)

example.

(9) My friends asked Sarah about her new job.

The argument structure and thematic relationship between elements in the sentence in (9) is graphically depicted in Figure 5.1. Arguments in the form of three syntactic constituents (two noun phrases and one prepositional phrase) are related to the verbal predicate *ask*, which is sometimes also referred to as the predicative *head* of the sentence. The associated participants involved in the event (and their function) are directly dependent on the head and are equipped with semantic role labels (drawn with arcs). These labels are generally theory-dependent but for the sake of the example, they could best be termed *agent* (who is asking), *recipient* (who is being asked) and *topic* (what is being asked about). A key aspect of semantic roles can be seen in the fact that they generalize across variants of grammatical voice. In fact, the syntactic realization of argument structure is different in Figures 5.1 and the passive version of 5.2, where *Sarah*, for instance, is the direct object in 5.1 but the subject in 5.2. Note, however, that in both sentences the *same* set of semantic roles is employed, i.e. the semantic role representation preserves the sentence meaning on a higher-order level of abstraction even when a sentence is realized in terms of different surface form, i.e. by different constituent parts-of-speech or word order, cf. Jurafsky and Martin (2017, Chapter 22). Put simply, in both Figures 5.1 and 5.2, *Sarah* has the same semantic role label *recipient*.

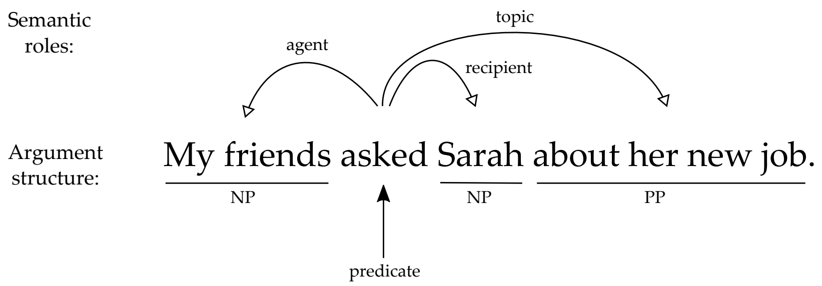


Figure 5.1: Sentence (9) with argument structure and semantic roles

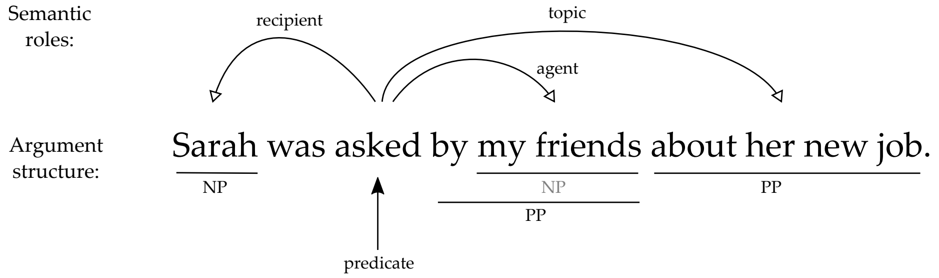


Figure 5.2: Passive version of the sentence in Figure 5.1 with a different syntactic argument realization but the same set of semantic roles

or a formal definition including role hierarchies introduced in Dowty (1991).

Semantic Roles & Implicit Information

Argument structure defines patterns and theoretical slots for lexical units which are to be filled by complements in running text, as we have seen in the previous two examples. We could, for instance, presume that *ask* requires at least a subject and two elements in object position—an agent, a recipient and an optional topic to ask about. However, a problem arises when arguments are assumed to be present, but when they are *not* encountered in the text. Consider the following example in Figure 5.3.

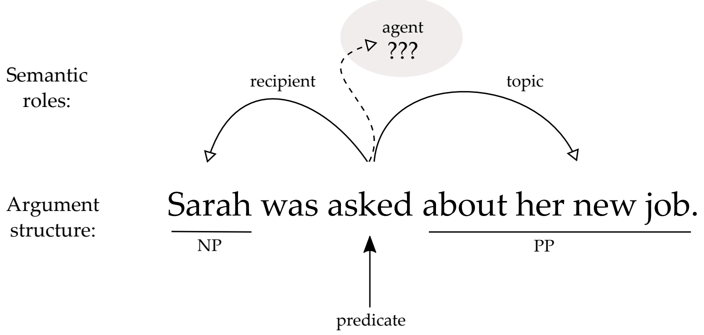


Figure 5.3: Variant of sentence (9) with no explicit constituent for the agent role

Here, although the sentence is grammatically flawless, arguably sufficiently meaningful and also interpretable in some specific context, only two semantic roles are present—recipient and topic, however, the agent role is not overtly filled by a constituent in the text. One could assume that, when the interaction of argument structure and semantic roles requires a specific component to be overtly realized, but when this is not the case in a given sentence, loosely speaking, the intended listener must nonetheless be able to pragmatically infer any related kind of information—either through world knowledge, or by means of recovery from the context. Such a non-explicitly realized (covert) role is called *implicit semantic role* or *null instantiation/NI* (Fillmore, 1986; Ruppenhofer, 2005). In the ensuing descriptions, we sometimes also refer to the notions of *null complement* or pragmatically controlled *zero anaphora* from the classical literature due to Fillmore (1986). Null instantiations are distinguished among a taxonomy of three types by *Definite* (*DNI*), *Indefinite* (*INI*) and *Constructional* (*CNI*) *Null Instantiations*, respectively. DNIs are typically anaphoric and the frame elements are resolvable from the context as, e.g., the *friends* constituent in Figure 5.3. DNIs are closely related to zero anaphora in Mandarin Chinese (Tao, 1996) because they can be typically resolved to an antecedent in the (prior) discourse. INIs are existential and do *not* need to be resolved by a constituent in the context (e.g., *Sarah is eating* with missing object) which can be understood by general principles of interpretation. CNIs are a result of the grammatical form of sentences in which the frame elements are omitted, for instance, due to a passive construction.

Note that in the given example in Figure 5.3, it could be the case that the *friends* (or whoever did in fact ask) have been mentioned already in the prior discourse and thus—due to stylistic reasons or due to a certain efficiency in text and language production—would make any additional mention of *friends* redundant in that particular sentence. Thus, any overt realization of the "friend" constituent (potentially found in any other sentence in the discourse) could serve as a *filler* for the missing implicit agent role of the verbal predicate *ask*. Note again that this process of detecting and linking implicit information is evoked locally on the word level by applying the search for an appropriate filler beyond the sentence boundary to the non-local context of the predicate.

From a practical perspective, the recognition of implicit semantic roles is highly beneficial because it provides information retrieval systems with additional, valuable information. For example, a question answering system could profit from the recovery of fillers for implicit agent roles, being capable of responding to questions such as "*Who asked Sarah?*". Most of the experiments in this thesis are thus concerned with anaphorically recoverable DNIs, but we will see that our methods are also generally enough to be applicable to existential and constructionally licensed interpretations of implicit roles.

Before we elaborate on efforts for supervised learning and gold-annotation of implicit roles and fillers in text (Section 5.2.1) as well as automated approaches to handling implicit roles (Section 5.2.2), the first part of this chapter, Section 5.1, lays out the theoretical foundations for their systematic detection by giving an overview of existent computational frameworks of predicate-argument structure and semantic roles. Specifically, we review three resources *FrameNet*, *PropBank* and *NomBank* and briefly describe how lexical units are modeled by elaborating on the specific semantic role label inventories and their idiosyncratic properties.

5.1 Frameworks of Argument Structure & Semantic Roles

5.1.1 FrameNet

The *FrameNet* project[2] (Baker et al., 1998) features an electronic data base of fine-grained schematic patterns in the form of an English dictionary with sample annotations for individual sentences. Its foundations are based on the theory of frame semantics (Fillmore, 1976). FrameNet distinguishes different high-level conceptual representations of situations along with their definitions, called *semantic frames*, each of which describes a specific interaction between events, participants involved in the event, and their associated role in terms of spatiotemporal relations. Lexical units are typically said to *evoke* a specific frame, for instance, the verbal predicate *ask* is captured by the QUESTIONING frame.[3] Intuitively, this means that the word *ask* activates a semantically connected concept representa-

[2] `https://framenet.icsi.berkeley.edu/fndrupal/`

[3] `https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Questioning`

tion which contains all the relevant pieces of information in order to understand the process behind asking—among others, the speakers involved, the message, topic, medium (e.g., over the phone) or the manner of asking. A frame instance is typically made up of a lexical unit and accompanying *frame elements* (the semantic roles) which can function either as *core roles* (obligatory parts in the interaction with a lexical unit) or as *non-core roles* (additional, mostly modifying content, also known as extra-thematic or peripheral roles). An example of such a frame instance for the text in sentence (9) is given in Figure 5.4. According to definitions in FrameNet, all three frame elements, *speaker*, *addressee* and *topic* are of type core role, i.e. they are obligatory for the interaction with the predicate. Other lexical units which could evoke the same frame are, for example, *interrogate* or *question*.

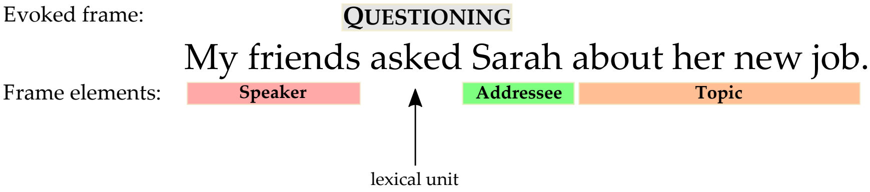


Figure 5.4: FrameNet analysis of sentence (9).

FrameNet distinguishes itself from other semantic role inventories by a very fine-grained label set and frame-to-frame relations, for example, by inheritance or other types of connected structure. In the illustration 5.4 above, the QUESTIONING frame *uses* the COMMUNICATION frame and *is used* by the COURT_EXAMINATION frame. Moreover, FrameNet comes with special constraints on relations between some frame elements, e.g., by *CoreSets* which state that given a set of core frame elements, the presence of only one is sufficient for a sentence to be "complete".[4] Similarly, the occurrence of a frame element can *Exclude* or *Require* the presence of another.

It should be noted that for a few of the manually collected example sentences by FrameNet it has not been possible to mark all core roles explicitly by human annotators. Therefore, in its current version, FrameNet provides a small set of annotations for implicit semantic roles, cf. the annotation description on null instantiations in Ruppenhofer et al. (2006). Finally, successful efforts have been made to extend FrameNet to other languages, e.g., to French[5] (Candito et al., 2014), Chinese[6] (Liu, 2011), or German.[7]

[4] Cf. http://www.cs.mu.oz.au/research/lt/nlp06/materials/Baker/CFBcourse6.pdf, accessed August 2017.

[5] https://sites.google.com/site/anrasfalda/

[6] http://sccfn.sxu.edu.cn

[7] http://www.laits.utexas.edu/gframenet/, all accessed July 2017.

5.1.2 PropBank

The Proposition Bank (Palmer et al., 2005; Kingsbury et al., 2002, PropBank[8]) is a corpus annotated with semantic propositions and verbal predicate-argument structure. Semantic roles are *contextualized*, i.e. they are annotated in running text of news articles of the Penn Treebank (Marcus et al., 1993). While FrameNet follows a rather semantically-driven approach, PropBank annotations are for the most part syntactically-oriented. Also, the semantic role types in PropBank are defined on a general "verb-by-verb" basis and are less fine-grained than the individual frame elements in FrameNet, which makes the shallow PropBank label inventory particularly suitable for practical systems and applications. More precisely, the label set comprises argument roles (which rougly correspond to FrameNet's core roles) enumerated from `A0` to `A5`. Argument `A0` has the property of a prototypical agent according to Dowty (1991); `A1` is a prototypical theme or patient and less clear-cut distinctions are made for the remaining roles `A2`-`A5` among the vast number of different predicates. As an illustration, the verb *ask.01*, for example, in its specific usage has the meaning of a direct question[9] and it can be described by the following role set:[10]

ask.01 (ask a question)	
`A0` (agent)	asker
`A1` (patient)	question
`A2` (recipient/goal)	hearer

Table 5.1: PropBank role set for the verbal predicate *ask*.

Returning to the motivating example of sentence (9), Figure 5.5 illustrates what a PropBank analysis would look like.

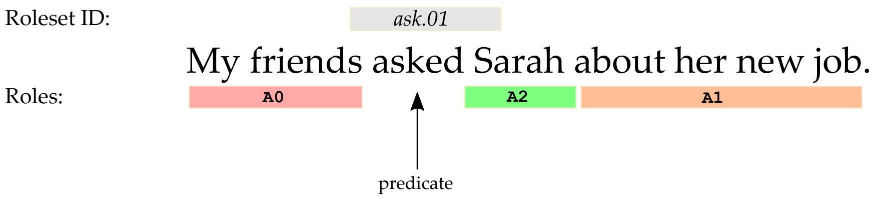


Figure 5.5: PropBank analysis of sentence (9).

Besides core argument roles the ProbBank label set comprises tags for non-core or adjunct roles, as well. For example, `AM-CAU`, `AM-PNC`, `AM-LOC`, `AM-TMP`, `AM-MNR` or `AM-NEG` represent modification for cause, purpose, location, time, manner or negation, respectively. The interested reader is referred to Palmer et al. (2005, Table 1) for a complete overview of the subtypes of modifier roles and to Jurafsky and

[8] `https://propbank.github.io/`

[9] It is distinguished from *ask.02* with similar role set, for example, which stands for *asking a request*.

[10] See also: `http://verbs.colorado.edu/propbank/framesets-english/ask-v.html`.

Martin (2017, Chapter 22) for a more precise description of the predicate-specific argument roles beyond A0 and A1. Also, for an extensive statistical analysis on individual labels and their distribution in mass data we elaborate on an experiment in the ensuing Chapter 8 of this thesis.

Finally, it is worth noting that there exist mapping mechanisms between PropBank and, e.g., FrameNet as described in Loper et al. (2007) as well as other lexical resources such as VerbNet (Schuler, 2005). Futhermore, PropBank annotations have found their way into other languages, e.g., for Chinese (Xue and Palmer, 2009) or Hindi/Urdu (Bhat et al., 2017), however, no implicit semantic roles are annotated in either the original English version or any other PropBank-based resource.

5.1.3 NomBank

The NomBank resource (Meyers et al., 2004)[11] has been developed in the style of the PropBank data set and employs the same label inventory (PropBank role set), yet focuses exclusively on the analysis on *nominal predicates*. The example of a PropBank predicate *reduce.01* is given in Table 5.2.[12] Its application to a nominal instance *reduction* is shown in Figure 5.6.

reduce.01 (make less)	
A0 (agent)	reducer
A1 (logical subject, patient)	thing falling
A2 (extent)	amount fallen
A3 (direction)	start point
A4 (goal)	end point
AM (location)	medium

Table 5.2: PropBank role set for the verbal predicate *reduce*.

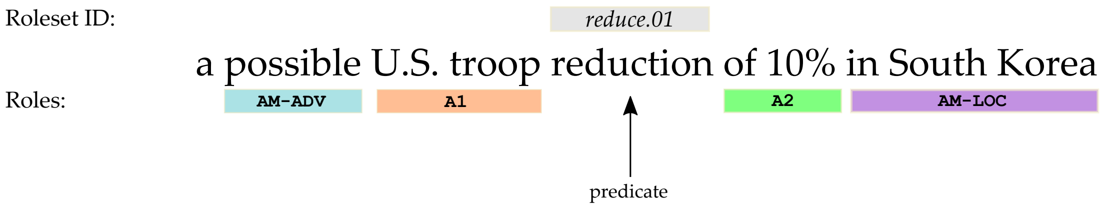


Figure 5.6: Sample NomBank analysis adapted from Meyers et al. (2004).

[11] http://nlp.cs.nyu.edu/meyers/NomBank.html

[12] http://verbs.colorado.edu/propbank/framesets-english/reduce-v.html

5.2 Automatic SRL & The Challenge of Implicit Roles

The previous section has briefly outlined three computational frameworks of predicate-argument structures which incorporate manual annotations for semantic roles. These resources have formed the basis for automated approaches to semantic role labeling (Gildea and Jurafsky, 2002; Carreras and Màrquez, 2005, *SRL*), and have resulted in a number of practical systems and applications, for example *SEMAFOR*[13] (Das et al., 2014) or *mate-tools*[14] (Björkelund et al., 2009), as large-scale annotations can be fruitfully exploited to train statistical models in a supervised setting for either FrameNet, PropBank or NomBank-style parsing (Giuglea and Moschitti, 2006; Jiang and Ng, 2006). SRL systems have been shown to heavily benefit various downstream NLP tasks such as question answering (Shen and Lapata, 2007; Moreda et al., 2011), recognizing textual entailment Sammons et al. (2012), (abstractive) text summarization (Trandabăţ, 2011; Khan et al., 2015) or as features for machine translation as proposed in Liu and Gildea (2010).

However, one issue related to the automated recognition of verbal, nominal or adjectival predicates along with their corresponding semantic roles is that the systems in the SRL analysis restrict their search for roles to *the local syntactic context of the predicate* or its *maximal projection*. That implies that a number of potential roles cannot be detected when these occur, for instance, in relative clauses, nested subordinations (as observed by Roth and Lapata (2016)), related constructions or even in the wider discourse context of other sentences. As an illustration, consider Example (10) from Roth and Frank (2013).

(10) El Salvador is now the only Latin American country which still has troops in [Iraq$_{\texttt{impl-A2}}$]. [Nicaragua$_{\texttt{A0}}$], [Honduras$_{\texttt{A0}}$] and the [Dominican Republic$_{\texttt{A0}}$] have [withdrawn$_{pred}$] their [troops$_{\texttt{A1}}$] [$\emptyset_{\texttt{A2}}$].

Ideally, a standard semantic role labeling system would identify *withdraw* as the main verbal predicate in the second sentence. In its thematic relation to the other words within the same sentence, all countries serve as the overtly expressed (explicit) agents, and are thus labeled as arguments `A0`. Semantically, they are the action performers, whereas *troops* would carry the patient role `A1` as the entity which undergoes the action of being withdrawn. However, given these explicit `A0` and `A1` annotations in the second sentence, the standard system would definitely fail to infer the underlying, linguistically unexpressed, i.e. non-overt realization of an implicit core argument of *withdraw* (denoted by [$\emptyset$]) about source information. Its corresponding realization is associated with *Iraq* in the preceding sentence, which is outside of the scope of any standard SRL parser. The resulting implicit role has the label `A2`.

As another example which involves a nominal predicate, consider (11).

[13] `http://demo.ark.cs.cmu.edu/parse`

[14] `https://code.google.com/archive/p/mate-tools/`

(11) "The answer isn't [price$_{pred}$] reductions, [. . .]", he said.[15]
([$\emptyset_{\texttt{A0}}$], [$\emptyset_{\texttt{A1}}$], [$\emptyset_{\texttt{A2}}$], [$\emptyset_{\texttt{A3}}$])

Given the nominal predicate *price.01*, the NomBank resource typically associates different semantic roles to it: a *seller* (A0), the *commodity* or *goods* / the price for what? (A1), the amount of the price / money (A2), a potential buyer (A3).[16] However, as denoted by the empty sets in the example sentence, note that *none* of these core argument roles are explicit (overt) in that sentence. To be more specific, they are not present within the immediate syntactic surroundings of the predicate instance, hence they cannot be detected by a traditional SRL system.

Interestingly, however, when we inspect the surrounding context of the predicate in previous sentences we can observe some words or phrases in the non-local context that could serve as candidate fillers for the semantic roles that are unexpressed in (11). In particular, this applies to the implicit goods role (impl-A1), which can be bound to two antecedents (*mainframe* and *machines*) that are realized in the extra-sentential prior context of the *price* target predicate. The sentence preceding (11) in the Wall Street Journal corpus from which the examples is drawn is shown hereafter in (12).

(12) "He questions whether that will be enough to stop Tandem's first [mainframe$_{\texttt{impl-A1}}$] from taking on some of the functions that large organizations previously sought from Big Blue's [machines$_{\texttt{impl-A1}}$]."

The example shows that the prices, that ought to be reduced as described in (11), refer to computer hardware. The establishment of these links is beyond the scope of a traditional SRL parser, however, the detection of this association would provide added value to any information retrieval system.

Traditional SRL has also been successfully employed in the biomedical domain. Here, the typical sentence structures oftentimes lack the presence of overt arguments. As an illustration, consider another example in (13), taken from Ruppenhofer et al. (2010).

(13) [Twenty-two month old$_{\texttt{impl-A0}}$] with history of recurrent right middle lobe infiltrate. Increased [$\emptyset_{\texttt{A0}}$] [cough$_{pred}$], [$\emptyset_{\texttt{A0}}$] [tachypnea$_{pred}$], and [$\emptyset_{\texttt{A0}}$] work of [breathing$_{pred}$].

In the second sentence, a standard SRL system would detect *cough, tachypnea* and *breathing* as nominal predicates. However, the A0 (experiencer/agent) role of these predicates is only explicitly realized in the preceding sentence and thus its identification is again beyond the scope of the traditional SRL parser. The agent role of *cough* is implicit, i.e. locally unexpressed and can only be resolved in the context. Its unique identification would be essential to downstream IR systems, especially for biomedical NLP which heavily relies on high-precision performance.

[15] PDTB Document ID wsj_2396.

[16] http://verbs.colorado.edu/propbank/framesets-english/price-v.html

5.2.1 Implicit Semantic Role Labeling: Corpora & Resources

We refer to the automated identification of implicit arguments and the linking process to appropriate antecedents or postcedents in the discourse as **implicit semantic role labeling** (iSRL). In order to allow for supervised machine learning and training of parsers for the challenging task of iSRL, it has become evident that—in the same line as for the traditional SRL systems—hand-crafted resources have had to be set up. In this regard, we briefly sketch three manual annotation efforts and resulting corpora which have initially built the sole foundation for these purposes. These resources have proven useful for a quantitative and thorough investigation of the involved phenomena behind implicit roles and have shed light on their idiosyncratic properties and how these should best be handled by automated systems.

5.2.1.1 SemEval Task 10

Ruppenhofer et al. (2010) were the first to release a data set[17] of fiction texts (≈17k tokens) annotated with null instantiated elements in running text in both a FrameNet (primary source) and a PropBank style format (converted annotations). The data contains a large number of distinct predicates with only a few (≈ 4) annotated instances for each type.[18] As opposed to news data, for example, the creators of the data set argue that the narrative characteristic of fiction novels would add to a larger number of resolvable implicit semantic roles.

The data set is accompanied by a shared task *"Linking Events and Their Participants in Discourse"* for i.) detecting and ii.) resolving null instantiated expressions and their fillers in the context. More specifically, participants are supposed to make their systems first identify whether a predicate's core role is missing, secondly whether the null complementation is *constructionally* or *lexically* specific and—for the latter case where the frame element is context-resolvable, i.e. in case it is a DNI—it should be linked to the correct antecedent. Note that if a frame element is omitted, potential fillers in the context can be single words, phrases, and sometimes also complete sentences, for example, for an implicit MESSAGE role.

5.2.1.2 Augmented NomBank

In the work of Gerber and Chai (2010) and Gerber and Chai (2012), the authors motivate the need for an additional layer of annotation on top of the primal NomBank data which (similar to PropBank) adds contextualized annotations to the Penn Treebank.[19] In particular, ten pervasive nominal predicates with unambiguous role sets (e.g., *price*, *sale*, *investment*) have been annotated for within-sentence and extra-sentential implicit arguments in addition to the already

[17] `https://bitbucket.org/josefkr/semeval2010-10/`, `http://semeval2.fbk.eu/semeval2.php`

[18] The average annotator agreement F-score for the frame assignments is about .75.

[19] `http://lair.cse.msu.edu/projects/semanticrole.html`

present roles. This way, the coverage of explicit roles could have been extended through implicit ones by 65% relative increase over roughly 1,200 predicate instances. An example of additional implicit argument annotations taken from the descriptions in Gerber and Chai (2010) is shown in (14). Note that the annotated phrases represent logical arguments of the *investment* predicate but none of these are part of NomBank or can be inferred by standard SRL.

(14) [Participants$_{\texttt{impl-A0}}$] will be able to transfer [money$_{\texttt{impl-A1}}$] to [other investment funds$_{\texttt{impl-A2}}$]. The [investment$_{pred}$] choices are limited to [a stock fund and a money-market fund$_{\texttt{impl-A2}}$].

On the one hand, compared to the SemEval Task 10 data, the augmented NomBank predicates cover only ten types but, on the other hand, more than 100 instances per type are annotated, which makes the resource more suitable to obtain better statistical generalizations for the detection of implicit arguments in text. Along with the resource, Gerber and Chai (2010) introduce a statistical model for the successful identification of implicit arguments, which will be elaborated on in closer detail in the next section. Importantly, for the construction of their resource, Gerber and Chai (2010) assume that implicit arguments are only core arguments.

5.2.1.3 Further Annotation Efforts

The severe lack of gold annotated resources of implicit roles in running text has motivated Moor et al. (2013) to pursue another annotation effort for a small set of predicates in order to obtain more valid generalizations across the multifaceted problem of implicit role linking. In particular, five predicates with varying number of frame elements were selected for annotation in OntoNotes (Hovy et al., 2006). To this end, their original PropBank semantic role labels were first mapped to FrameNet via SemLink. Approximately 2,000 instances were judged according to whether a missing null instantiated role is *resolvable*, i.e. whether the referent is anaphorically bound within a context window, or not. Moor et al. (2013) found that when NIs can be resolved this is typically the case for only one or two core roles of a predicate. For the predicate *leave* of the Departing frame, approximately 60% of the fillers for the Source frame element are resolvable, while 30% are not. In contrast, no filler for Theme can be found in their data experiment. Overall, the corpus study has shown that a majority of implicit roles is non-resolvable (≈60%). When they are resolvable, a realization of roughly 80% can be observed within a context window of 3 sentences around the predicate under investigation. Moor et al. (2013) relate their findings to Gerber and Chai (2010) (90% resolvable in the same window) and to the data of Ruppenhofer et al. (2010) (70% resolvable) and explain the discrepancies by the different text genres. Interestingly, Moor et al. (2013) observe a quantity of predicate instances whose frame elements are in conflict with the definition of their property descriptions in FrameNet, e.g., expected "physical" roles which are filled by abstract entities. The authors conclude their work by demonstration of the usefulness of hand-annotated implicit roles in a supervised classification task for NI role linking.

To summarize, manually augmented data with implicit semantic roles are scarce. In this section we have described three hand-crafted resources based on either FrameNet or PropBank-style annotations. In what follows, we review (roughly chronologically) how distinct efforts have been made to try to yield useful generalizations by exploiting these infrequent—yet valuable—annotations for supervised machine learning in order to tackle the challenging task of iSRL. The next section also describes rule-based systems as well as empirical observations to guide the proper resolution of fillers for implicit roles in running text.

5.2.2 Related Work on iSRL

An automated iSRL parser integrates at least two components: The first one *identifies* whether a given predicate instance has locally unfilled arguments, i.e. implicit roles. The second one optionally determines the orientation, i.e. whether an implicit role is resolvable (DNI) or not (INI), cf. Ruppenhofer et al. (2006), and then computes a set of candidate constituents in the (mostly prior) context *linking* the implicit role to its correct antecedent(s). Implicit role identification and implicit role resolution are two separate tasks, and most of the literature on automated iSRL has focused on the latter—which is typically more challenging—under the assumption that the gold annotated data (cf. Sect. 5.2.1) provides a system already with an indication on which DNIs need to be linked in the discourse. This section reviews a number of FrameNet (Sect. 5.2.2.1) and PropBank-style iSRL approaches (Sect. 5.2.2.2), as well as related techniques on the task (Sect. 5.2.2.3).

5.2.2.1 FrameNet-style iSRL

Chen et al. (2010): One of the first attempts to implicit argument resolution in the independent evaluation framework of the SemEval Task 10 on *"Linking Events and Their Participants in Discourse"* (cf. Sect. 5.2.1.1) was made by Chen et al. (2010). To this end, the authors have adapted the first version of *SEMAFOR* (Das et al., 2010), a frame-semantic parser for overt argument identification based on the FrameNet paradigm by extending the system to cope with non-overt arguments. In fact, the search space for implicit roles has been widened to the discourse context beyond the predicate's target sentence according to a two-stage pipeline: First, a rule-based target selection determines (local) frame-evoking lexical units and a probabilistic procedure is applied to disambiguate the associated frame. Secondly, in the *argument identification* step, the search for implicit candidate fillers is now extended to the discourse as sometimes not all arguments are realized locally.[20] Here, Chen et al. (2010) restrict the search for DNI ref-

[20] In the NI-only task, Chen et al. (2010) rely on gold annotations for overt arguments and consult FrameNet for non-overt core roles for a given predicate instance. Special care is taken by the authors for frame elements which belong to the same *CoreSet*—a FrameNet-specific aspect which informally states that the presence of a particular frame element renders the overt realization of another redundant in that particular context so that Chen et al. (2010) do not need to search for implicit fillers.

erents to the previous three sentences and limit fillers to nouns, pronouns and noun phrases. In order to deal with non-local argument structure, the orginal *SEMAFOR* parser was equipped with a slightly adapted feature set: In particular, the authors make use of FrameNet's manually drafted lexicographer files (which exemplify the prototypical usage of a particular frame). In the enhanced version of *SEMAFOR* candidate NIs (or their head words) are compared to these examples—either directly or by distributional similarity. Another modification in the feature set takes care of the distance between a potential filler and the target predicate.

Overall, the performance of their model is adequate, however, DNI detection has a low overall recall. Chen et al. (2010) assert that this is mainly due to the highly skewed distribution of DNIs in the gold annotated data, as well as the overall scarcity of gold annotated NIs which makes it tremendously hard to obtain useful generalizations in a supervised setting.

Tonelli and Delmonte (2010, 2011): In the same style as Chen et al. (2010), Tonelli and Delmonte (2010) adapt an existing system (originally used for entailment recognition) to the task of implicit role linking evaluated on SemEval Task 10. Specifically, the authors describe *VENSES*++, a deep symbolic processing pipeline which consists of three modules for syntactic and lexico-semantic processing, anaphora resolution and a semantic component. The approach by Tonelli and Delmonte (2010) does not make use of supervised learning, instead it is highly linguistically inspired: It involves logical form representations and aspects from LFG (Bresnan and Kaplan, 1982) and Centering Theory (Grosz et al., 1995) being applied to anaphoric expressions. Crucially, *VENSES*++ computes predicate-argument structures for both nominal and verbal predicates separately and first maps them to the predefined valence patterns in FrameNet in order to derive information on missing core roles. Here, Tonelli and Delmonte (2010) assume that null instantiated arguments are always core frame elements. In order to find candidate fillers as referential expressions for the implicit roles, Tonelli and Delmonte (2010) compute semantically related phrases using WordNet (Miller, 1995), for instance. Even though their approach is promising, it is not reasonably effective in resolving NIs which is mainly due to the convoluted system architecture.

A follow-up to the aforementioned work is presented in Tonelli and Delmonte (2011) in which the authors extend their linguistically motivated approach by making use of frequency information directly obtained from the training sections of the SemEval data, for instance, in order to classify arguments as DNI vs. INI based on majority vote. Moreover, Tonelli and Delmonte (2011) define a relevance score for candidate NIs and a distance penalty in order to assess the adequacy of an antecedent to serve as a filler for an implicit role. Their heuristic procedure is driven by FramNet-specific modeling assumptions, e.g., *incorporated* frame elements are not treated as NIs, as their contribution towards the core status of a frame element remains questionable according to Tonelli and Delmonte (2011). The general idea of deriving frequency-based generalizations in the form of a background knowledge base is highly promising, however their approach is restricted to only a small set of predicates in the SemEval data. In an evaluation, Tonelli and Delmonte (2011) report slight performance improvements over the work in Chen et al. (2010).

Silberer and Frank (2012): All previously described methods have treated iSRL as a special case of SRL. However, one of the earliest implementations for rendering implicit information explicit made use of (co-)reference as described in the system of Palmer et al. (1986). Along similar lines, Silberer and Frank (2012) tackle the problem of linking NIs in the discourse as a special case of (zero) anaphora resolution, in which the anaphoric role, i.e. the DNI to be linked, is bound to a discourse antecedent.[21] Silberer and Frank (2012) motivate their approach through the fact that in iSRL—unlike in classical SRL—relationships between semantic roles and their fillers *cannot* be modeled by local syntactic paths, which indeed suggests an approach to the problem from a higher perspective. In their approach, Silberer and Frank (2012) promote the idea of an *entity-mention model* in which an implicit role is connected to an entity as part of a singleton or non-singleton coreference chain. Semantic class features (Rahman and Ng, 2011) are computed between a DNI and an entity, as well as salience properties in order to derive evidence for its appropriateness as a filler. A discriminative model is trained in a supervised setting with the additional guidance of heuristically acquired iSRL data and its performance is tested in the framework of SemEval 2010. According to an ablation study and an exploration of coreference and traditional SRL features, it turns out that the most distinctive features are coreference-related, e.g., the *prominence* of an entity. Silberer and Frank (2012) achieve state-of-the-art performance with their approach over the best system heretofore by Tonelli and Delmonte (2011).

In a related work, Roth and Frank (2013) re-address the iSRL data bottleneck and further pursue the idea of heuristic data acquisition. The authors propose to align predicate-argument structures of comparable texts and demonstrate that the so-inferred, automatically induced implicit arguments can help the identification of DNIs and explain local text coherence.

Laparra and Rigau (2012, 2013): A novel technique for resolving fillers for implicit roles on the SemEval data is presented in Laparra and Rigau (2012). Their method exploits lexical units and co-occurring *explicit*, i.e. overt frame elements in order to learn a useful semantic representation of a candidate filler for a specific frame element. For instance, Laparra and Rigau (2012) demonstrate that words such as *hotel*, *dwelling* and *house* can be aggregated by the semantic type *Building* with the help of word sense disambiguation and a suitable ontology (Álvez et al., 2008). This implies that (the head word of) any appropriate candidate filler for an implicit, say, Location frame element would need to share the same semantic properties, thus yielding a higher degree of flexibility and generalization in their

[21] Note that, as opposed to classical coreference resolution, the local entity to be linked (here: the DNI) is non-overt and not necessarily a noun phrase.

resolution. Besides having a probability distribution for frame elements and associated semantic types, their model also associates part-of-speech information to all explicit arguments of a lexical unit. When maximizing the joint probability of both among all terminals in a three sentence candidate window, their approach outperforms the one by Silberer and Frank (2012).

A refinement of their work adapted to the same data set is made in Laparra and Rigau (2013b). In addition to the semantic type agreement, this study is a more thorough inspection of various linguistically-inspired features from the domain of co-reference and anaphora resolution and it investigates how these relate to the task of resolving implicit arguments. To this end, Laparra and Rigau (2013b) experiment with a whole range of "sources of evidence", e.g., syntactic phrase-structure features in the form of c-command and distance information (termed *nearness*). The idea is that early work on pronoun resolution has established a close connection between the referenced entities and their syntactic relation (Hobbs, 1977). Laparra and Rigau (2013b) find that most DNIs appear in close proximity to the target frame-evoking predicate (*recency*). A peculiarity of the SemEval data can be seen in the mixture of narrative content. Interestingly, Laparra and Rigau (2013b) observe that in the large majority of cases the target lexical unit and the corresponding filler for the implicit role belong to the *same* level of discourse structure, for example, it rarely happens that the filler for the null complement appears in a monologue and the predicate in a dialogue. Closely related, another finding relates to the fact that candidate constituents which are part of a singleton coreference chain (i.e. they are mentioned only once and are not referred to again) have a low probability of filling an implicit role, as opposed to entities which are salient in terms of focus and whose mentions are connected throughout the text. Laparra and Rigau (2013b) report a slight performance improvement with their host of co-reference features in a supervised setting on the SemEval data over the prior state-of-the-art, however, crucially leave the question unanswered of how to deal with non-nominal entities and referents belonging to distinct sentences.

Gorinski et al. (2013): Recently, Gorinski et al. (2013) have explored alternative strategies to the supervised systems and introduce a semi-supervised method for NI resolution on the SemEval data. Their approach is largely linguistically motivated and consists of an explorative investigation of four different types of resolvers: The first module follows an exact match for the semantic (frame element or WordNet-derived) type between an uninstantiated role and an active left-context member mention in all coreference chains of some active window. The second string-based resolver checks the prior discourse for an overt frame element which fills the current DNI of the same type. The motivation here is very similar to the one in Laparra and Rigau (2012). A third component for NI resolution makes use of frame element patterns and computes an overlap of roles between target and candidate predicates. A final resolver applies distributional similarity between an NI and a candidate's head word. To this end, context vectors are estimated from external corpora. Gorinski et al. (2013) demonstrate that the semantics-based resolvers perform best individually and that the highest scores can be achieved by majority vote among all four. In line with the conclusions made by Gerber and Chai (2010) (cf. next section), the qualitative evaluation in the explorative study by Gorinski et al. (2013) suggest that NI resolution should be more predicate-specific as for some frames, some frame elements are never resolved in the context, while for others already a narrow window context is adequate.

5.2.2.2 PropBank/NomBank-style iSRL

Gerber and Chai (2010, 2012): A main contribution of the work by Gerber and Chai (2010) can be seen in the additional layer of implicit role annotations which has been added to the primal NomBank data (cf. the description in Section 5.2.1.2). As opposed to the null instantiations from SemEval 2010, here only ten predicates have been selected for annotation (yet in a larger per-instance volume) and Gerber and Chai (2010) were the first to investigate the phenomenon of *nominal* PropBank-style iSRL on a computational scale. Resolving fillers for implicit arguments is cast as a two-step approach: First, a simple lexicon lookup in NomBank determines the missing role(s) of a predicate. Second, Gerber and Chai (2010) train a binary classifier to decide whether a candidate constituent (as part of a coreference chain) fills a missing implicit role. This process is guided by a set of hand-crafted features. This set contains, among others, semantic type generalizations between candidates and missing arguments based on VerbNet, part-of-speech, head word and discourse relation information, sentence distance, or special features of pointwise mutual information as inspired from the literature on narrative events (Chambers and Jurafsky, 2008) adapted to (implicit) semantic roles. Gerber and Chai (2010) demonstrate that the most powerful and expressive features can be attributed to the semantic types which is in line with the findings in Laparra and Rigau (2012). Also, similar to Silberer and Frank (2012), Gerber and Chai (2010) limit the search for NIs to the current and previous two sentences of the target predicate, however, note that this simplified assumption accounts for a non-neglectable error. Interestingly also, some implicit arguments were filled by *non-core*, i.e. adjunct roles which raises a principled question of whether a simple lexicon lookup is sufficient to initiate a search for an implicit role.[22] Gerber and Chai (2010) also note that iSRL might need more predicate-specific strategies (e.g., predicate-specific window sizes to search for fillers) instead of having one global model for all distinct predicate types. Finally, the authors demonstrate that, in their setup, discourse features do not help the resolution of implicit roles—a claim which we will take up again in the ensuing Chapter 8.

In a follow-up publication, Gerber and Chai (2012) present a more principled and extended evaluation scenario (using an extended feature set and cross-validation instead of fixed training and test splits) for the discriminative classification model and can report slight performance improvements.

[22] We will elaborate on an alternative strategy in the next chapter.

Laparra and Rigau (2013): In a follow-up work, Laparra and Rigau (2013a) approach iSRL from a different perspective. This time, Laparra and Rigau (2013a) propose a completely deterministic algorithm (*ImpAr*) which does not rely on supervised machine learning and iSRL-specific training data. The idea is to exploit the discourse coherence properties of predicates between explicit and implicit roles and is inspired by the early work of Dahl et al. (1987). More precisely, it is assumed that subsequent (anaphoric) mentions of the same (nominal) predicate can be linked to the same arguments, even if these arguments appeared only once in the prior context. For example, a single mention of the nominal predicate *loss* can have unfilled local arguments but it is highly likely that previous mentions of the same predicate have already explicitly filled those. A special issue to consider is how *changes* in the shared argument realization should be handled, e.g., when for a coherent chain of the same predicate an argument role is suddenly realized by a different filler (e.g., by a change in the `A0` role for a company name). To this end, Laparra and Rigau (2013a) introduce a damping factor on the salience for a candidate filler which is sensitive to the occurrence of a new explicit filler. Candidate phrases with highest salience scores in a windows of two sentences prior to the target predicate are selected as fillers. In addition to these metrics, their proposed algorithm makes use of a semantic tagger for named-entities and WordNet senses (Ciaramita and Altun, 2006) and a custom defined category system to obtain information on the appropriateness of a candidate filler. Overall, the results reported in Laparra and Rigau (2013a) are highly competitive with supervised systems and can even outperform them on some of the predicates of the iSRL NomBank data from Gerber and Chai (2010, 2012), however, as the authors note themselves, some room for improvement in their system is still left for ways to deal with more accurate span detection for the fillers.

5.2.2.3 Further Approaches to iSRL & Related Techniques

Besides the individual core efforts outlined above, there have recently emerged various related techniques which draw on the challenging task of iSRL. For example, it has been pointed out that the existing iSRL data sets of SemEval and NomBank do in fact differ heavily with respect to genre (novels vs. news), the predicate types (different parts-of-speech vs. nouns), frameworks of argument structure (FrameNet vs. Prop/NomBank) or seem to come with a tradeoff regarding the quantity of annotations (many types but few instances and vice versa). Therefore, many systems which have been developed were specifically task-tailored to one or the other data set. Contrary to these observations, however, Feizabadi and Padó (2015) have come up with a promising technique to exploit the complementary properties of both resources by *combination* of corpora in a domain adaptation setting. In particular, they describe a discriminative model trained on iSRL annotations from both corpora, including a set of syntactic, semantic and discourse features. Importantly, their model incorporates a balance mechanism for features which undergo changes across different domains. Feizabadi and Padó (2015) demonstrate that even small proportions of out-of-domain data can lead to an increased performance on the task. It should be noted, however, that a major improvement stems from the feature augmentation method they employed, cf. Daume III (2007).

On a related note, Li et al. (2015) extend the search for implicit role fillers to Chinese. A main contribution of their work has been the annotation of a sample of approx. 1,600 sentences and 420 null instantiations. Their annotation workflow follows the theoretical foundations and terminology of the English FrameNet in terms of DNIs and INIs, and frame element-specific constraints such as *CoreSet*, *Excludes* or *Requires* properties (cf. Section 5.1.1). In their work, Li et al. (2015) are the first to exploit the frame-to-frame relations in the Chinese FrameNet (Liu, 2011) for the resolution of implicit roles in their data. Here, the authors rely on the assumption that explicit frame elements can serve as fillers for other predicates' DNIs. A maximum entropy model is trained for NI type identification (DNI vs INI) and DNI resolution is guided by path and inheritance properties between frames and frame elements. In agreement with the vast majority of the related literature on automated iSRL, Li et al. (2015) report a window size of two sentences prior to the target for DNI resolution to be optimal in their setting and state-of-the-art performance is achieved over one related attempt to Chinese iSRL (Lei et al., 2013).

5.2.3 Current Issues in iSRL

Most of the techniques outlined in this section have cast iSRL as a two-step pipeline for i.) implicit role *identification* and ii.) implicit role *resolution*. However, the specific ways of how these two individual steps have been addressed come with related issues. In the following, we elaborate on these aspects, and in the ensuing two chapters, we propose suggestions for improvement.

5.2.3.1 Implicit Role Identification

Detecting null instantiated arguments of any given predicate instance relies heavily on predefined lexicons. For a predicate and its overt arguments in a FrameNet setting, for example, standard iSRL parsers directly consult the FrameNet resource via an API to check which *core* arguments are not overtly realized in that context (Silberer and Frank, 2012). Similarly, for the PropBank or NomBank, all roles which are part of a predicate's role set of `A0-A5` but which are not locally present in running text are considered implicit. This assumption follows a lexicon lookup and requires the presence of frameworks of argument structure and semantic roles, which we outlined in Section 5.1.

This heuristic approach seems straightforward at first glance, but it comes with additional complications, in particular, since the number of core roles is *predicate-specific* and there is no complete consensus among linguists and the NLP community on the exact number of argument slots for any given predicate type. Only very recently, Jauhar and Hovy (2017) have proposed to learn predicates and (the number of) associated semantic roles (slots) automatically in an unsupervised manner from text—which is a promising attempt and relaxes the assumption of pre-defined static lexicon entries.

Yet, the particular pattern-based approach to identify null complements in text has been reported in the literature as the default strategy in most iSRL parsers and it is still employed in current systems. The main disadvantages of this method do not only relate to the costly hand-crafted and idiosyncratic rules that have to be designed for that purpose (Ruppenhofer et al., 2011), but also to the lexicon-specific heuristics (Chen et al., 2010), the static dictionary lookups (Gerber and Chai, 2012), and the required rich background knowledge in the form of the lexical resources. The design of such language and domain-specific heuristics is expensive and their realization is not possible to the same extent for other languages for which no such resources are available.

Even though prior research on iSRL has (almost exclusively) restricted the analysis of implicit roles to core roles (Tonelli and Delmonte, 2010), or excluded "conceptually redundant" roles without further elaboration (Silberer and Frank, 2012), some researchers have pointed out that *non-core* roles could be implicit, too. In fact, Gerber and Chai (2010) relate this observation to an error rate in their system—an issue not to be neglected. Furthermore, pattern-based methods to detect null complements offer little flexibility in the resolution process in that they assume that all candidate NIs are equally likely to be missing, but such a scenario is unrealistic and obviously a simplification given the linguistic variety of different contexts in which predicates co-occur with semantic roles, and crucially, some prior approaches are ignorant of INIs altogether, i.e. they do not account for the non-resolvable null instantiations (Laparra and Rigau, 2013a). Moreover, it is unclear how different predicate senses are handled in this setting, e.g., in the system by Tonelli and Delmonte (2011), or the detection of NIs is artificially limited to only a small set of predicates with only one unambiguous sense, as in Gerber and Chai (2012).

In particular for the FrameNet frames, implicit role identification is extremely convoluted—especially when considering the additional interacting constraints of, e.g., the *CoreSet*, *Requires* or *Excludes* relationships between frame elements. Closely related, in the linguistically-motivated exploratory work of Ruppenhofer et al. (2011) the authors find among other things that statistics obtained from FrameNet are not necessarily helpful in deciding on the correct *interpretation* of an implicit role (DNI vs. INI, i.e. whether it is resolvable or not in a specific context). Crucially, the exemplar annotations in FrameNet only depict a single, majority use (sentence instances are not collected randomly from corpora) and do not account for context variation, which is, however, typical of real/authentic texts.

In this context, we argue that overall, the identification of whether a predicate instance is missing an implicit role in the discourse context should *not* be a binary decision based on lexicons but rather be driven by tendencies whose foundations are *statistical evidence* as obtained from usages in real corpora. We elaborate on a specific approach in the next Chapter 6.

5.2.3.2 Implicit Role Resolution

Similar issues as for iSRL role detection relate to the identification and linking of a correct antecedent.

In general, the state-of-the-art in iSRL relies on costly gold-annotated training data and hand-crafted features (Gerber and Chai, 2012; Silberer and Frank, 2012; Li et al., 2015) for resolving fillers for implicit roles. Moreover, it has been pointed out, that the available amount of iSRL training data is tremendously sparse (cf. Section 5.2.1) which represents a serious bottleneck for training systems in a supervised setting and restricts iSRL unproductively to only specific domains and languages.

As a consequence some efforts have been made to combine the scarce iSRL training resources in the framework of domain adaptation (Feizabadi and Padó, 2015) which have been fruitful but the serious problems regarding data sparsity on corpora with suitable implicit role annotations still remain.

To circumvent these issues, deterministic methods for iSRL have been proposed (Laparra and Rigau, 2013a) which can do without iSRL specific training data but crucially these systems require language-specific tools and resources, e.g., (syntactic) dependency parsers, semantic taggers for named-entities and *super senses* or ontologies and custom taxonomies. A combination of rich semantic information obtained from these specific tools works reasonably well for iSRL in a restricted domain, however yields worse performance on out-of-domain texts. Moreover, these semantic annotations are not available to the same extent in other languages which again restricts the resolution of fillers for implicit roles to a subset of available texts.

As pointed out by Gerber and Chai (2010), we argue that iSRL and, in particular, the search for an adequate antecedent for a locally unfilled role should be *predicate-specific*. Here, instead of having one global model, we further argue that role resolution should be based on *distributional properties* of candidate fillers and the predicate at hand. Gorinski et al. (2013) have made a first promising attempt in that direction in that they estimate prototypical features of a particular frame element to fill a certain role. This method works largely unsupervised and does not rely on gold annotated iSRL training data, however, can be modified in various ways. We report on an improvement of that method in the ensuing Chapter 7. We present a knowledge-lean extension of their idea which obviates the need for gold iSRL training data and manual feature engineering.

To summarize, we relate our methodology described in the next two chapters to the previous work on iSRL by introduction of two novel methods for implicit role identification (Chapter 6), and implicit role resolution (Chapter 7).

Chapter 6

Role Identification by Pattern-based Learning

6.1 Motivation

In the last chapter, we have introduced implicit semantic role labeling as a promising extension of traditional SRL, to widen the search for arguments to the non-local discourse context. This chapter introduces a novel statistical approach for **implicit role identification**, i.e. to determine for any given predicate and its overtly realized explicit semantic roles which roles are locally uninstantiated.

Section 5.2.3 has described current issues related to iSRL. We have seen that corpora with manually annotated implicit roles are extremely rare and that the available annotations are tremendously scarce. Annotation efforts and techniques to bundle these rare resources have been initiated only recently (Ruppenhofer et al., 2010; Gerber and Chai, 2012; Feizabadi and Padó, 2015). Therefore, most state-of-the-art iSRL systems cannot rely on supervised machine learning for that specific task and, as a consequence, employ rule-based procedures to detect null complements in text. Here, the overtly realized roles of a predicate are compared against a predefined template. Generally speaking, all roles that are part of the template but that are not present in the text are regarded as null complements. We have described a whole range of drawbacks related to these heuristic template-based methods. In general, they require the establishment of expensive, language- and domain-dependent frameworks of argument structure, they suffer from coverage issues, they are affected by inflexible modeling assumptions among individual predicates, resolvable (DNI) and non-resolvable (INI) roles, insofar as the templates are either too complex, too narrowly defined, or ignorant of particular roles.

As an alternative, in this chapter, we propose an improvement upon traditional methods to detect null instantiations (NIs) in the form of a novel and more generic approach. Crucially, this approach obviates the need for predicate-specific templates and direct reference to manually annotated gold data. The technique only considers a predicate's overtly expressed role(s) as encountered in running text, and solely based on that information, determines which roles are implicit. It is specific to a predicate instance and its predicate sense, however, not bound to a particular release of a hand-crafted lexicon. It is agnostic with regards to the role inventory, i.e. in principle, our approach applies to both FrameNet as well as PropBank/NomBank-style frame elements/roles, and it can detect both core roles, as well as non-core roles. Instead of a rule-based lookup, in this work, we primarily argue for a *probabilistic detection of NIs* as a more flexible, context-sensitive mechanism that conditions on the presence of other co-occurring roles in a given predicate environment. In particular, implicit arguments are predicted using probabilistic information directly derived from large corpora which have been automatically equipped with SRL structure. Based on these annotations, we construct a background knowledge base of predicates and their corresponding co-occuring (overtly realized) roles. This approach is conceptually similar to memory-based learning (Daelemans and van den Bosch, 2009) and can be considered a mildly-supervised learning technique to infer evidence for implicit roles, grounded on a large-scale collection and generalization of explicit role patterns. It should be noted that our method is structurally similar to the one previously described by Laparra and Rigau (2012), but ours does not only estimate the *raw* frequencies from a very limited training corpus. Moreover, as we have pointed out already, we do not omit all the valuable less frequent patterns in limiting NI detection to only a subtype of NI instances that are resolvable from the context. Our derived models are in large measure domain-independent, and the methodology can straightforwardly be ported to other languages as well, given as only requirement the availability of an (explicit) SRL parser. Unlike the approaches described in earlier studies, we exemplify our method using a generic role set that is based on PropBank/NomBank-style parsing, rather than FrameNet, as the small PropBank role inventory allows for better statistical generalizations as opposed to the fine-grained frame elements. The approach described in this chapter is also partly motivated by the recent development of AMR parsing (Banarescu et al., 2013, Abstract Meaning Representation). The purpose of AMR representations is to represent the semantics of a text by abstraction from idiosyncratic syntactic properties. In this setting, extensive use of PropBank roles is made. We believe that AMR parsing in general would greatly profit from the recovery of implicit roles, as well.

In this chapter, we focus on the identification of implicit roles first, and provide a direct connection point to NI role resolution that we describe in the next chapter. The reason why we treat both identification and resolution as two separate modules is partly also due to the fact that not all NIs are resolvable from the discourse context (cf. the distinction between DNIs and INIs in the introductory chapter). The chapter is structured as follows: Section 6.2 presents our novel approach to probabilistic NI detection. In Section 6.3, we introduce a series of three experiments along with their evaluation. We compare our results to those from the related literature, and finally, in Section 6.4 conclude our work presented here.

6.2 Induction of Predicate-Specific Role Patterns

6.2.1 Memory-Based Learning

Memory-based learning in the context of NLP (Daelemans and van den Bosch, 2009) is a lazy learning technique that stores a collection of training instances as represented by a *background knowledge base* (BKB). In a classification scenario, new instances are directly compared to the items stored in the BKB by means of a distance metric. Applications related to background knowledge and memory-based learning have been proposed in the literature, e.g., Peñas and Hovy (2010) for semantic enrichment of text, or Chiarcos (2012) in order to derive implicit markers for discourse relations. In this section, we describe how we adopt its general concepts to identify null complements, i.e. implicit semantic roles in running text. In order to realize this goal, we first construct a BKB that stores probabilistic information on explicit predicate-role co-occurrences. In a subsequent step, we introduce constellation-specific thresholds that need to be optimized on iSRL training data. These thresholds serve as a trigger for the prediction of a null complement. On a general level, they can be considered a slight modification of the distance metric. We elaborate on details in Section 6.3.

6.2.2 Data & Preprocessing

As a textual basis for our model, we make use of a subset of the *WaCkypedia_EN*[1] corpus (Baroni et al., 2009). The corpus represents a Wikipedia dump from 2008 from which we retrieved the pre-split sentences. We have further subdivided the text collection cumulatively into smaller pieces of growing sizes (100 sentences each) and employed an SRL labeler to the texts that produces SRL annotations in PropBank format, *mate-tools*[2] (Björkelund et al., 2009). For each sentence of the dump, *mate* detects all predicate instances, distinguishes different predicate senses for verbal and nominal predicates, and identifies all associated core and non-core arguments of a predicate.[3] Note that the same system has been used in previous research on implicit semantic roles already, e.g., by Roth and Frank (2013).

6.2.3 Model Generation

We generate our probabilistic model from the annotated SRL/predicate-role patterns as follows:

1. For every sentence, we keep a record of all distinct predicates and their associated roles.

[1] `http://wacky.sslmit.unibo.it/doku.php?id=corpora`

[2] `http://code.google.com/p/mate-tools/`

[3] Unrealistically long sentences from the data dump (> 90 tokens) were re-split using the Stanford Core NLP module (Manning et al., 2014) in order to remove noisy content. We ultimately rejected all splits > 70 tokens.

2. For every predicate instance, the role labels are sorted in lexicographic order, and duplicate role labels are collapsed. This way, we disregard their original sequential appearance and obtain a normalized template of role co-occurrences for each frame instantiation.

3. We compute the frequencies for all distinct patterns of the same predicate.

4. We derive all following conditional probabilities by relative frequency estimation:
$$P(r|R, \mathit{predicate})$$
where $\mathcal{R}$ is the complete role inventory of the SRL parser, $R \subseteq \mathcal{R}$ is a subset of overt semantic roles in a given predicate context, and $r \in \mathcal{R} \setminus R$ an arbitrary semantic role. When we try to assess whether a particular role is implicit, r is typically the unrealized role. The *predicate* consists of the lemma of the corresponding verb or noun, optionally followed by sense number (if predicates are sense-disambiguated) and its part of speech (V/N), e.g., *play.01.N*.

We build models from SRL data in PropBank format, both manually and automatically annotated. We experiment with models for two different styles of predicates: *Sense-ignorant* or ***SI model***s represent predicates by lemma and part of speech (*play.n*), *sense-disambiguated* or ***SD model***s represent predicates by lemma, sense number and part of speech (*play.01.N*, *play.02.N*, etc.). To illustrate this mechanism, let us return to the introductory example (10) from the previous chapter that we reproduce here in its abbreviated form, where an explicit core role `A2` was missing in the sentence.

(15) [Nicaragua$_{\mathtt{A0}}$], [Honduras$_{\mathtt{A0}}$] and the [Dominican Republic$_{\mathtt{A0}}$] have [withdrawn$_{\mathit{pred}}$] their [troops$_{\mathtt{A1}}$] [$\varnothing_{\mathtt{A2}}$].

As an application of our model to Example (15), we would consult the BKB to look up the probability of an implicit `A2`, given the predicate *withdraw*[4] and its overt role constellation consisting of (collapsed occurrences of `A0`) and `A1`. Note again that this information can be solely derived from automatically annotated SRL data:
$$P(\mathtt{A2}|\{\mathtt{A0}, \mathtt{A1}\}, \mathit{withdraw.V}) = \mathit{some\ probability}$$

6.2.4 Annotated Data—FrameNet and PropBank Formats

We evaluate our proposed model on the SemEval data from Ruppenhofer et al. (2010), in line with previous iSRL approaches. However, for the reasons concerning generalization that we outlined before, our focus is on the evaluation based on the PropBank format of this data set. To the best of our knowledge,

[4] `http://verbs.colorado.edu/propbank/framesets-english/withdraw-v.html`

Paradigm		**#Roles**			**#Explicit**
		Explicit	DNI	INI	**#DNI+#INI**
Training	FrameNet	2,526	303	277	4.36
	PropBank	1,027	125	101	4.52
Test	FrameNet	3,141	349	361	4.42
	PropBank	1,332	167	85	5.28

Table 6.1: Explicit and implicit roles in the SemEval 2010 data set in both the FrameNet (all roles and parts of speech) and the PropBank version (only core roles for nouns and verbs).

this is the first such study to do so. In fact, the PropBank version has been derived semi-automatically from an original FrameNet base format with the help of hand-crafted conversion rules (which are part of the data set) both for verbal and nominal predicates. As an illustration, consider a mapping for the *fear* predicate from FrameNet's Experiencer_focus frame. It corresponds to the PropBank predicate *fear.01* (the first predicate sense) and the associated roles Experiencer and Content are ported to the PropBank labels A0 and A1, respectively. As an indirect effect of the mapping patterns, the resulting distribution of null instantiated arguments varies slightly between the base format and the converted format. General statistics are reported in Table 6.1 which shows the label distribution of explicit roles, DNIs, and INIs for both the original FrameNet and the converted PropBank versions, respectively. It should be noted that as a result of the conversion, some information is lost, however the overall proportions remain similar among the two formats (cf. the last colum which shows the ratio of explicit to implicit roles). The differences are partly also due to the fact that for instance for adjectives no mappings were defined, even though some instances link null complements in the base format. Furthermore, the mapping rules have only been created for core roles A0-A4, i.e. agent, patient, etc. We want to point out again, that our proposed method is flexible enough to also handle non-core roles. This means that, for instance, AM-TEMP (temporal modification) or AM-LOC (location) are also part of $\mathcal{R}$.[5] However, in order to properly evaluate our method on the iSRL data set of Ruppenhofer et al. (2010) we limit our analysis to these five unique roles. In order to assess the applicability of our proposed model, we report on three experiments in the following.

[5] We describe an application of the BKB involving non-core roles in Chapter 8.

Role	**Verbs**		**Nouns**	
	Overt	NIs	Overt	NIs
A0	40	**45**	24	23
A1	83	39	29	**33**
A2	3	11	10	6
A3	-	7	-	1
A4	-	24	-	-
totals:	126	126	63	63

Table 6.2: Label distributions of all roles in both data sets from Experiment 1; majority NI classes in bold.

6.3 Evaluation

6.3.1 Experiment 1

To evaluate the general usefulness of our memory-based approach to detect implicit roles, we set up a simplified framework for predicates with exactly *one overt argument and one NI* annotated in the SemEval data (for all verbs and all nouns and from both the train and test files to obtain a reasonably large sample; no differentiation of DNIs and INIs). This pattern accounts for 189 instances—roughly 9% of the data samples in the SemEval set. We divided the instances into two subsets based on the predicate's part of speech. The label distributions over overt and null instantiated roles for both verbal and nominal predicates are given in Table 6.2.

6.3.1.1 Task Description

Predict the role of the single missing NI (A0–A4) for each given predicate instance.

6.3.1.2 Predicting Null Instantiations

We trained one sense-disambiguated (*SD*) gold model for verbs (*PB*) and one for nouns (*NB*) according to Sect. 6.2.3 on the complete PropBank and the complete NomBank, respectively. This was compared with 30 separate *SD* and *SI* models on varying portions of the automatically annotated *WaCkypedia_EN* dump: These were trained on the first k sentences each, in order to make their prediction quality comparable, while k ranges from 50 sentences for the smallest model to $k = 10$ million for the largest model ($\approx \frac{1}{5}$ of the whole corpus). For NI role prediction, we return n_i, i.e. the maximally probable unrealized semantic role given the overt argument o_j plus the predicate:

$$n_i = \arg \max_{n \in \mathcal{R} \setminus R} P(n|o_j, predicate),$$

where $R = \{o_j\}$, the predicate's single explicit role and $\mathcal{R} = \{\texttt{A0..A4}\} \supset R$, the role inventory.

6.3.1.3 Results & Evaluation

The prediction accuracies for verbal and nominal predicates are illustrated in Figure 6.1. Although the number of instances in the data sets is small, some general trends are clearly visible. Our major findings are:

By increasing the number of training sentences the performance of the *SD* and the *SI*-based classification models steadily increases as well. The trend is the same for both verbs and for nouns, even though training in the nominal domain requires more data to obtain similarly good results. More precisely, models trained on only 50k sentences already have an adequate performance on test data for verbs (≈76% with the *SD* model). To reach a similar performance on nouns, we need to increase the training size roughly by a factor of 5.

Likewise, the performance of the *SD* models is better in general than the one of the *SI* models throughout all models analyzing verbal predicates, but only marginally better for nouns.

Both the *SD* and the *SI* models outperform the majority class baseline for both parts of speech.[6]

Also, with 800k sentences for nouns and only 50k sentences for verbs, both *SD* model types reach accuracies equal to or greater than the supervised *PB* and *NB* (gold) models which have been trained on the complete PropBank and NomBank corpus including sense distinctions, respectively.

The classification accuracies for the *SD* models reach their saturated maxima for verbs at around 91.27% (115/126) with 6 million training sentences and 85.71% (54/63) with 2.85 million sentences for nouns. For verbs, a χ^2 test confirms a significant ($p < .01$) improvement of our best model over the *PB* gold model. On the sparse evaluation data for nouns, the improvement over the *NB* gold model is, however, not significant.

Taken together, the improvements confirm that memory-based learning over mass data of automatically annotated (explicit) semantic roles can actually outperform gold models constructed from corpora with manual SRL annotations, even if the tools for automated mass annotation were trained on the very same corpora used to build the gold models (PropBank, NomBank). Also, the experiment demonstrated the feasibility of predicting implicit roles solely using information about the distribution of explicit roles. For the artificially simplified NI patterns in Experiment 1, already small portions of automatically annotated SRL data are sufficient to yield adequate results for both types (DNIs and INIs). Sense disambiguation of predicates generally increases the performance.[7]

[6] 35.71% with only 1k training sentences (verbs), 52.38% with 50k sentences (nouns).

[7] A simple error analysis of the misclassified noun instances revealed that classification on the test data suffers from sparsity issues: In the portions of the *WaCkypedia_EN* that we used for model building, three predicates were not attested (twice *murder.01* and once *murderer.01*). This has a considerable impact on test results.

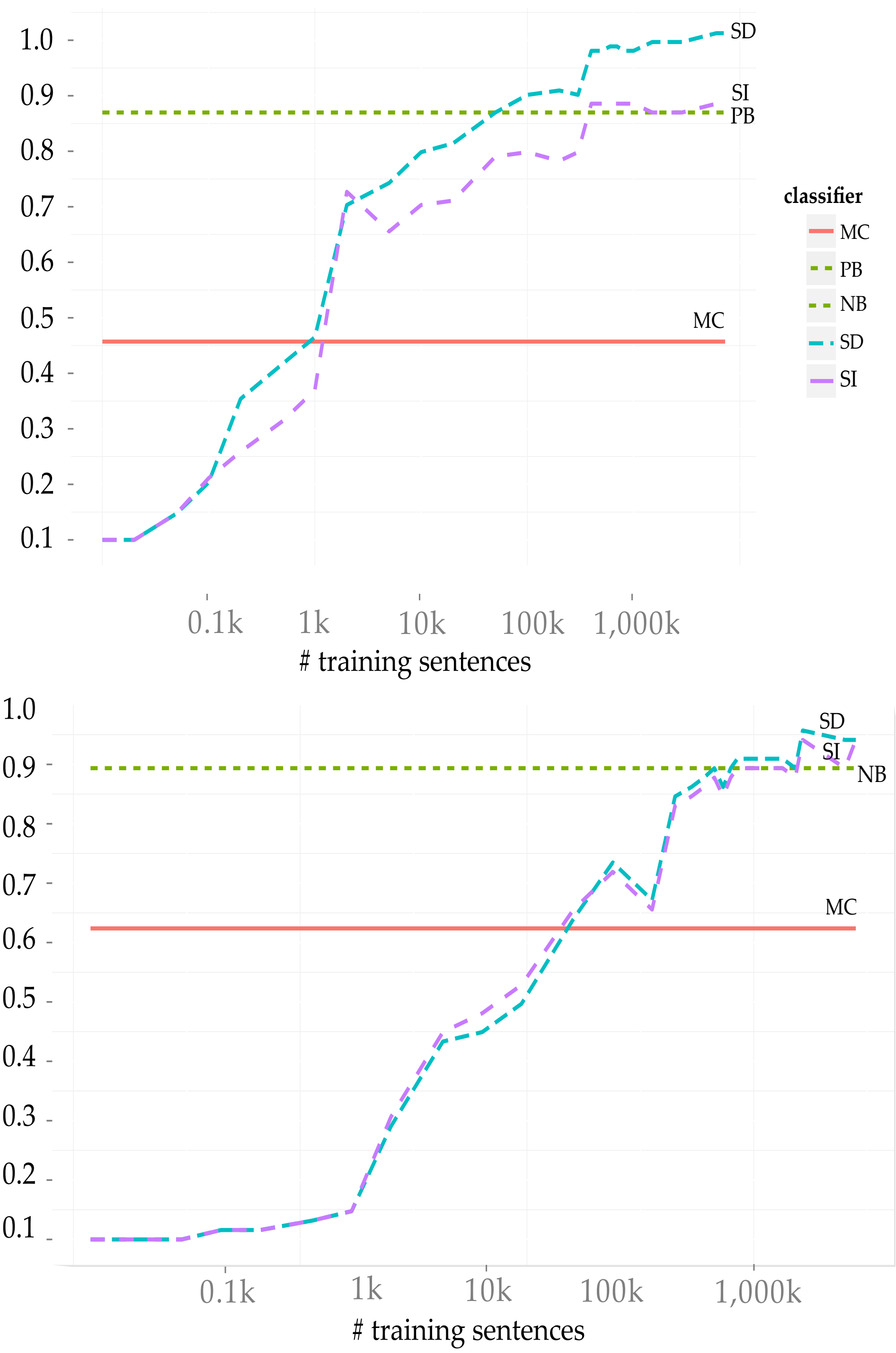


Figure 6.1: Prediction accuracies for verbal (top figure) and nominal predicates (bottom figure) from Experiment 1. Majority class (*MC*) baselines in red, PropBank (*PB*) and NomBank (*NB*) gold models in green. The log-scaled x-axis only refers to the *SD* and *SI* models and indicates first *k* sentences used for training.

NI Pattern	Freq	NI Pattern	Freq
-	706	A0 A2	7
A1	86	A1 A2	6
A0	51	A3	5
A2	35	A1 A4	3
A4	18	A0 A1 A2	1
A0 A1	11		

Table 6.3: The 929 NI role patterns from the test set sorted by their number of occurrence. Most of the predicates are saturated and do not seek an implicit argument. Only one predicate instance has three implicit roles.

6.3.2 Experiment 2

The setup from the previous experiment is by far too simplistic compared to a real linguistic scenario. Usually, a predicate can have an arbitrary number of overt arguments, and similarly the number of missing NIs varies. To tackle this problem, we take the original train and test split (744 vs. 929 unrestricted frame instances of the form: any combination of overt roles vs. any combination of NI roles per predicate). Again, we do not draw a distinction between DNIs and INIs, but treat them generally as NIs. Table 6.3 shows the distribution of the different NI role patterns in the test data.

6.3.2.1 Task Description

Given a predicate and its overtly expressed arguments (ranging from any combination of A0 to A4 or none), predict the correct set of null instantiations (which can also be empty or contain up to five different implicit elements).

6.3.2.2 Predicting Null Instantiations

We distinguish two main types of classifiers: *supervised classifiers* are directly obtained from NI annotations in the SemEval training data, *mildly supervised classifiers* instead use only information about (automatically obtained) explicitly realized semantic roles in a given corpus, *hybrid classifiers* combine both sources of information. We estimated all parameters optimizing F-measure on the train section of the SemEval data set. Their performance is evaluated on its test section. We aim to demonstrate that mildly supervised classifiers are capable of predicting implicit roles, and to study whether NI annotations can be used to improve their performance.

Baseline: Given the diversity of possible patterns, it is hard to decide how a suitable and competitive baseline should be defined: predicting the majority class means not to predict anything. So, instead, we predict implicit argument roles randomly, but in a way that emulates their frequency distribution in the SemEval data (cf. Tab. 6.3), i.e. predict no NIs with a probability of 76.0% (706/929), A1 with 38.6% (86/929), etc. The baseline scores are averaged over 100 runs of this random 'classifier', further referred to as A.

Supervised classifier: Supervised classifiers, as understood here, are classifiers that use the information obtained from manual NI annotations. We set up *two* predictors B_1 and B_2 tuned on the SemEval training set: B_1 is obtained by counting for each predicate its *most frequent NI role pattern*. For instance, for *seem.02*—once annotated with implicit `A1`, but twice without implicit arguments—B_1 would predict an empty set of NIs. B_2 is similar to B_1 but conditions NI role patterns not only on the predicate, but also on its explicit arguments.[8] For prediction, these classifiers consult the most frequent NI pattern observed for a predicate (B_2: plus its overt arguments). If a test predicate is unknown (i.e. not present in the training data), we predict the majority class (empty set) for NI.

Mildly supervised classifier: Mildly supervised classifiers do not take any NI annotation into account. Instead, they rely on explicitly realized semantic roles observed in a corpus, but use explicit NI annotations only to estimate prediction thresholds. We describe an extension of our prediction method from Exp. 1 and present eight parameter-based classification algorithms for our best-performing *SD* model from Exp. 1, trained on 6 million sentences.

We define prediction for classifier $\mathbf{C}_0$ as follows: Given a predicate *predicate*, the role inventory $\mathcal{R} = \{\texttt{A0..A4}\}$, its (possibly empty) set of overt roles $R \subseteq \mathcal{R}$ and a fixed, predicate-independent threshold t_0. We start by optimizing threshold t_0 on all predicate instances with *no* given overt argument. If there is *no* overt role and an unrealized role $n_i \in \mathcal{R}$ for which it is true that $P(n_i|\mathit{predicate}) > t_0$, then predict n_i as an implicit role. If there is an overt role $o_j \in R$ and an unrealized role $n_i \in \mathcal{R} \setminus R$ for which it is true that $P(n_i|o_j,\mathit{predicate}) > t_0$, then predict n_i as an implicit role. Note that C_0 requires this condition to hold for *one* o_j, not all explicit arguments of the predicate instance (logical disjunction).

We refine this classifier by introducing an additional parameter that accounts for the group of overtly realized frames with exactly *one* overt argument, i.e. $\mathbf{C}_1$ predicts n_i if $P(n_i|o_j,\mathit{predicate}) > t_1$; for all other configurations the procedure is the same as in C_0, i.e. the threshold t_0 is applied.

Classifiers $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$ extend C_1 accordingly and introduce additional thresholds t_2, t_3, t_4 for the respective number of overt arguments. For example, $\mathbf{C}_3$ predicts n_i if $P(n_i|o_{j_1}, o_{j_2}, o_{j_3},\mathit{predicate}) > t_3$, for configurations with fewer arguments it relies on C_2, etc. Our general strategy here is to see whether the increasing number of specialized parameters for increasingly marginal groups of frames is justified by the improvements we achieve in this way.

A final classifier $\mathbf{C}_{4_{n,v}}$ extends C_4 by distinguishing verbal and nominal predicates, yielding a total of ten parameters $t_{0_n}..t_{4_n}, t_{0_v}..t_{0_n}$.

Hybrid classifier: To explore to what extent explicit NI annotations improve the

[8] Specifically, we extract finer-grained patterns, e.g., *evening.01*[`A1`] → {}=2, {`A2`}=3, where a predicate is associated with its overt role(s) (left side of the arrow). The corresponding implicit role patterns and their number of occurrence is shown to the right.

Classifier	A	B_1	B_2
Precision	*0.149*	0.848	**0.853**
Recall	*0.075*	*0.155*	*0.206*
F_1 Score	*0.100*	*0.262*	*0.332*

Classifier	C_0	C_1	C_2	C_3	C_4	$C4_{n,v}$	$C4_{n,v,B1}$	$C4_{n,v,B2}$
Precision	*0.368*	*0.378*	0.398	0.400	0.400	0.423	0.561	0.582
Recall	**0.861**	0.851	0.837	0.837	0.837	0.782	0.615	0.814
F_1 Score	*0.516*	*0.523*	*0.540*	*0.541*	*0.541*	*0.549*	0.589	**0.679**

Table 6.4: Precision, recall and F_1 scores for all classifiers introduced in Experiment 2. Scores are compared row-wise to the best-performing classifier $C4_{n,v,B2}$ (cf. last column). A significant improvement over a cell entry with $p < .05$ is indicated in *italics*.

classification results, we combine the best-performing and most elaborate mildly supervised classifier $C4_{n,v}$ with the supervised classifiers B_1 and B_2: For predicates encountered in the training data, $\mathbf{C4}_{n,v,B_1}$ (resp., $\mathbf{C4}_{n,v,B_2}$) uses B_1 (resp., B_2) to predict the most frequent pattern observed for the predicate; for unknown predicates, apply the threshold-based procedure of $C4_{n,v}$.

6.3.2.3 Results & Evaluation

Table 6.4 contains the evaluation scores for the individual parameter-based classifiers. All classifiers demonstrate significant improvements over the random baseline. Also the mildly supervised classifiers outperform the supervised algorithms in terms of F_1 score and recall. However, detecting NIs by the supervised classifiers is very accurate in terms of high precision. Classifier B_2 outperforms B_1 as a result of directly incorporating additional information about the overt arguments.

Concerning our parameter-based classifiers, the main observations are: First, the overall performance (F_1 score) increases from C_0 to C_4 (yet not significantly). Secondly, with more parameters, recall decreases while precision increases. We can observe, however, that improvements from C_2 to C_4 are marginal, at best, due to the sparsity of predicates with two or more overt arguments. Results for C_3 and C_4 are identical, as no predicate with more than three overt arguments occurred in the test data. Encoding the distinction between verbal and nominal predicates into the classifier again slightly increases the performance.

A combination of the high-precision supervised classifiers and the best performing mildly supervised algorithm yields a significant boost in performance (Tab. 6.4, last two columns). The optimal parameter values for all classifiers $C4_{n,v}$ estimated on the train section of the SemEval data set are given in Table 6.5.

In Table 6.6, we report the performance of our best classifier $C4_{n,v,B2}$ with detailed label scores. Its overall NI recognition rate of 0.81 (recall) outperforms the

Noun thresholds	t_{C0_n}	t_{C1_n}	t_{C2_n}	t_{C3_n}	t_{C4_n}
Values	0.35	0.10	0.20	0.35	0.45
Verb thresholds	t_{C0_v}	t_{C1_v}	t_{C2_v}	t_{C3_v}	t_{C4_v}
Values	0.05	0.25	0.25	0.30	0.20

Table 6.5: Optimal parameter values for the thresholds in all $C_{4_{n,v}}$ classifiers estimated on the train section of the SemEval data set.

state-of-the-art in implicit role identification: cf. Table 6.7.[9]

Roles	A0	A1	A2	A3	A4
# Labels	70	107	49	5	21
Precision	0.675	0.578	0.432	0.400	0.791
Recall	0.800	0.897	0.653	0.400	0.905
F_1 Score	0.732	0.703	0.520	0.400	0.844

Table 6.6: Evaluation of $C_{4_{n,v,B2}}$ for all 252 implicit roles.

Summarizing our results, Exp. 2 has shown that combining supervised and mildly supervised strategies to NI detection achieves the best results on the SemEval test set. Concerning the mildly supervised, parameter-based classifiers, it has proven beneficial to incorporate a maximum of available information on overtly expressed arguments in order to determine implicit roles. Our best-performing classifier achieves an NI recognition rate beyond state-of-the-art.

Interestingly, memory-based learning offers the capability to detect both DNIs (resolvable from context), as well as INIs (not resolvable from context), simply by learning patterns from local explicit role realizations. Although the higher-parametrized classifiers are not significantly better than the ones with only a few parameters, we believe that this is primarily due to data sparseness in the training set as a general trend of a performance improvement is nonetheless noticeable. Similar problems related to data sparsity have been reported in Chen et al. (2010).

6.3.3 Experiment 3

Our final experiment focuses on the comparison with previous research where DNI and INI predictions are separately evaluated. In our setting, however, we regard this evaluation as artificial as DNI/INI classification could alternatively be decided depending on distance and availability of potential antecedents, a problem we would like to address in subsequent experiments.

[9] Note that only an indirect comparison of these scores is possible due to the aforementioned difference between data formats and also because none of the other systems report precision scores for their pattern-based NI detection systems.

System	NI recall	DNI/INI interpret. prec relative	absolute
Laparra and Rigau (2012)	0.66	-	-
Chen et al. (2010, SEMAFOR)	0.63	0.55	0.35
Silberer and Frank (2012)	0.58	0.70	**0.40**
Tonelli and Delmonte (2011)	0.54	**0.75**	**0.40**
Tonelli and Delmonte (2010, VENSES++)	0.08	0.64	0.05
Schenk et al. (2015):	**0.81**	0.57	0.36

Table 6.7: Recognition rate (recall) for all NIs, relative (based on correctly recognized) and absolute precision scores comparing the different state-of-the-art systems to our best-performing classifier $C4_{n,v,B2}$.

6.3.3.1 Task Description

For every predicate, predict the set of null instantiations as in Exp. 2. Then, classify every predicted NI as DNI or INI.

6.3.3.2 Predicting Null Instantiations

We take the best-performing classifier $C4_{n,v,B2}$ from Experiment 2. Following Tonelli and Delmonte (2011), we then employ a rule-based classifier $\mathbf{C}_{\text{DNI,INI}}$ to separate predicted NIs into DNIs or INIs by application of the following constraints in this order:

1. predict INI for predicates with part of speech VBN/VBG (e.g., in passive voice).
2. predict the majority class according to DNI/INI frequencies for the predicate in the SemEval training set.
3. predict DNI if DNI/INI frequencies are equal or the predicate is missing in the SemEval training data.

6.3.3.3 Results

Incorporating $C_{\text{DNI,INI}}$ into the best performing NI classifier from Experiment 2 outperforms current state-of-the-art systems in terms of NI recall (Tab. 6.7) but has drawbacks in DNI/INI classification.[10]

A closer look at the individual NI types (upper part of Table 6.8) reveals that, overall, the performance of our predictor is competitive regarding the accuracies

[10] Note that our scores are not directly comparable as none of the other systems report precision scores for their pattern-based NI detection modules and our evaluation is based on the PropBank version of the data set whose label distribution, contrasting DNIs and INIs, is different from the FrameNet format (DNI majority class: 66.3% vs. 50.8%).

System	Type	Precision	Recall	F_1 Score
Tonelli and Delmonte (2011)	DNI	0.39	0.43	0.41
	INI	**0.46**	0.38	**0.42**
Chen et al. (2010, SEMAFOR)	DNI	**0.57**	0.03	0.06
	INI	0.20	**0.61**	0.30
Schenk et al. (2015)	DNI	0.43	**0.44**	**0.43**
	INI	0.24	0.51	0.32
Laparra and Rigau (2012)	DNI	**0.50**	0.66	**0.57**
Schenk et al. (2015):	DNI	0.41	**0.86**	0.55

Table 6.8: INI vs. DNI classification compared to previous works (upper part). Silberer and Frank (2012) do not report individual NI type scores. Laparra and Rigau (2012) focus only on DNI detection. Our results on this subtask are shown in the last row.

by the systems reported by Tonelli and Delmonte (2011) and Chen et al. (2010, SEMAFOR). More specifically, there is no single best performing system. The system by Tonelli and Delmonte (2011) is generally powerful in predicting INIs, SEMAFOR has high recall and high precision for both, while we outperform the others on DNI analysis. That our system is particularly efficient in detecting DNIs as opposed to INIs (i.e. powerful on those null complements which are resolvable from the context) seems plausible, because the outlined memory-based strategy relies on *explicit* role patterns. We conclude that, when a specific null complement is resolvable from the context, its probability of appearing locally in the immediate syntactic context of the predicate is greater as opposed to null complements which are not resolvable, and thus, typically tend to occur less frequently in close local proximity of the predicate.

Clearly, the best results are obtained by Laparra and Rigau (2012). However, the authors only report accuracies for the identification of DNIs, as INIs are beyond their scope. The last row of Table 6.8 gives the scores of our tool when we substitute $C_{\text{DNI,INI}}$ by predicting the majority class (DNI). Outperforming all other systems, we are able to detect 86% of all DNIs in the test set with an F_1 score only marginally worse than L&R.

6.4 Summary

We have presented a novel, statistical method to infer evidence for implicit roles from their explicit realizations in large amounts of automatically annotated SRL data. We conclude that—especially when annotated training data is sparse—memory-based approaches to implicit role detection seem highly promising. With a much greater degree of flexibility, they offer an alternative solution to static rule- and template-based methods.

Despite its simplicity, we demonstrated the suitability of our approach: It is competitive with state-of-the-art systems in terms of the overall recognition rate, however, still suffers in precision of the respective null instantiated arguments. Thus, directions for future research should consider integrating additional contextual features, and would benefit from the *complete* role inventory of our models—including non-core roles. Regarding this extended setting, it could be interesting to experiment with other machine learning approaches to assess whether the accuracy of the detected NIs can be increased. In particular, regarding the estimation of the parameters of our proposed model it would be highly beneficial to learn the weights automatically by means of a neural network, for instance, in place of the present ad-hoc solution. Along similar lines, Do et al. (2017) have recently proposed the first recurrent neural network to sequentially learn explicit realizations of semantic frame arguments and applied their method to iSRL. Their approach is highly promising and—even though deep learning has had a great impact on many related areas in NLP—up to now this has not been the case for iSRL.

In this chapter, we have presented a novel method for iSRL role identification which we extend in the next chapter. It draws on the main ideas introduced here and presents a resource-lean approach to implicit role resolution.

Chapter 7

Role Resolution with Prototypical Fillers

7.1 Motivation

The previous chapter has outlined a novel resource-lean method for iSRL role identification. The method described in this chapter extends its main ideas and introduces a closely related approach to **implicit role resolution**, i.e. linking a locally uninstantiated role with an appropriate antecedent in the discourse. Crucially, the approach presented in this chapter is situated in the same knowledge-lean framework. In this context, we pointed out in the previous chapter that traditional state-of-the-art approaches to implicit SRL (iSRL) are supervised and need a groundwork of hand-annotated training data—which is costly, extremely sparse, limited to only a handful of predicates, and requires careful feature engineering (Gerber and Chai, 2012; Silberer and Frank, 2012; Li et al., 2015). A first attempt has been made to combine the scarce resources available by Feizabadi and Padó (2015), but given the great diversity of predicate-specific roles and enormous complexity of the task, the main issues regarding feature engineering remain, cf. Chen et al. (2010) and the description in Section 5.2.3. A promising exploratory effort recently made by Gorinski et al. (2013) aims to overcome the annotation bottleneck by using distributional methods to infer evidence for elements filling null instantiated roles. The authors do not rely on gold annotations but instead learn distributional properties of fillers induced from a large corpus.

In this chapter, we propose an extension of the distributional idea for unsupervised iSRL to reduce the need for annotated training data. Specifically, we propose to induce predicate and role-specific **prototypical fillers** from large amounts of SRL annotated texts in order to resolve null instantiations as (semantically and syntactically) similar elements found in the context. Parts of our approach have been successfully applied in traditional SRL (Hermann et al., 2014), but not yet to implicit roles. Our work differs from Gorinski et al. (2013) in that we extend discrete context vectors to SRL-guided embeddings and experiment with a variety of different configurations. We intend *not* to set a new benchmark beating the current state-of-the-art for *supervised* iSRL, but rather provide a simple and alternative strategy which does not rely on manually annotated gold data. Still, we demonstrate that our method is highly competitive with supervised methods on one out of two standard evaluation sets and that it can easily be extended to other predicates for which no implicit gold annotations are available.

7.2 Learning Prototype Representations

7.2.1 Prototypical Fillers

We use large amounts of *explicit* SRL annotations to compute predicate-specific *protofillers* (prototypical fillers) for each frame element (role) individually:

$$\vec{v}^{\,protofiller} = \frac{1}{N} \sum_{i=0}^{N} E(w_i) \tag{7.1}$$

where N is the total number of tokens filling a particular role and $E(\cdot)$ is an embedding function which maps a word w_i to its distributed representation, i.e. a precomputed vector of d dimensions. Note that only those words contribute to the protofiller of a frame element which occur in this role.

As an illustration, consider again the motivating example (11) taken from the WallStreet Journal texts of the Penn Treebank which is reproduced here.

(16) “The answer isn’t [price$_{\boldsymbol{pred}}$] reductions, [. . .]”, he said.[1]

In the Example (16), the nominal predicate *price* is not associated with any local arguments and it is unclear from the restricted context what the price reductions refer to. However, given the domain of newswire texts in Gigaword, we can straightforwardly compute a domain-specific prototypical filler for a specific role of the predicate by simply collecting all instances of *price* associated with the overtly realized constituents filling that role, e.g., realized by *the price for gold*, the *energy price*, the *price of a stock*, etc. Each constituent contributes to the latent syntactic and semantic properties which are captured in the generated protofiller. An illustration of this process for the commodity/goods role of *price* is depicted in Figure 7.1.

[1] PDTB Document ID `wsj_2396`.

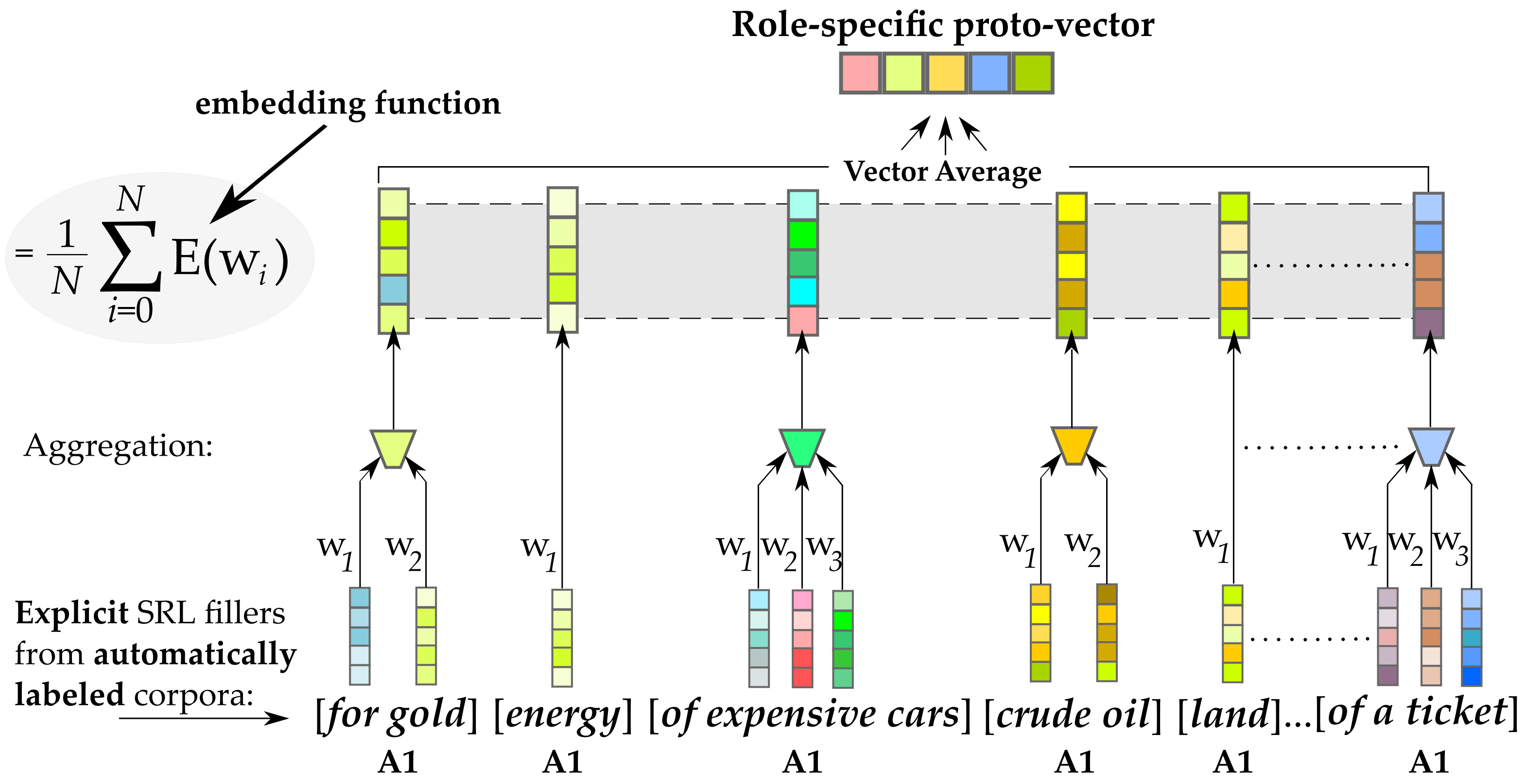

Role-specific proto-vector
embedding function
Vector Average
$= \frac{1}{N}\sum_{i=0}^{N} E(w_i)$
Aggregation:
w_1
w_2
w_1
w_1
w_2
w_3
w_1
w_2
w_1
w_1
w_2
w_3
Explicit SRL fillers from automatically labeled corpora:
[for gold] [energy] [of expensive cars] [crude oil] [land]...[of a ticket]
A1
A1
A1
A1
A1
A1

Figure 7.1: Illustration of the generation of a prototypical filler for the `A1` (commodity/goods) role for the nominal predicate *price*.

7.2.2 Identifying Null Instantiations

In this section, we illustrate how the learned protofillers can be applied to *resolve* implicit semantic roles. To this end, consider another example with implicit roles, this time in FrameNet style, from Ruppenhofer et al. (2010):

(17) [In the centre of this room$_{\text{GOAL}/\mathbf{NI}}$] there was an upright beam, [which$_{\text{THEME}}$] had been [placed$_{\boldsymbol{pred}}$] [at some period$_{\text{TIME}}$] as a support for the old worm-eaten baulk of timber which spanned the roof.

Here, the predicate *place* evokes the PLACING frame, with two frame elements (roles) overtly expressed (THEME and TIME) but with one null instantiated role (GOAL) beyond the embedded relative clause and thus beyond the scope of traditional SRL. Similar to the example in (16), our approach generalizes over labeled filler instances of the frame (PLACING) as found in corpus data, e.g., *planted on the top of the church*, *placed on the middle picture*, *hung over the river*, *laid on the table*, etc. We exploit their syntactic (in this case the prepositional) and semantic properties (i.e. inanimate, spacial NPs) in order to capture a composed meaning and thus to approximate the correct implicit role *in the centre of this room*. We measure similarity between a trained protofiller $\vec{v}^p$ and a candidate constituent $\vec{v}^c$ by cosine similarity

$$\cos(\theta) = \frac{\vec{v}^p \cdot \vec{v}^c}{\|\vec{v}^p\|\|\vec{v}^c\|}$$

and predict a candidate as null instantiation which maximizes the inner product with the protofiller. As candidate constituents for an implicit argument we initially consider all terminal and non-terminal nodes in a context window of the predicate, ruling out those categories which never occur as implicit arguments, which do contain the target predicate and/or which are already overt arguments. The result set comprises mainly nouns, verbs and PPs. Candidate constituents in our evaluation data are available from their respective (manual) syntax annotation, but could easily be extracted using automated phrase-structure parsers. The candidate *vectors* for arbitrary length n-grams are derived in the same way (by means of Equation 7.1).

7.2.3 Training Resources & Tools

In accordance with domain-specific evaluation data, we chose to learn protofillers on two distinct corpora: *The Corpus of Late Modern English Texts, CLMET* (Smet, 2005) (≈35M tokens, 18th–20th century novels) and a subset of the English Gigaword corpus (Graff and Cieri, 2003) (≈500M tokens of newswire texts). We label the first one with *SEMAFOR*[2] (Das et al., 2014), a FrameNet-style semantic parser.

[2] http://www.cs.cmu.edu/~ark/SEMAFOR/

We employ *mate-tools*[3] (Björkelund et al., 2009) to obtain a PropBank/NomBank analysis for each sentence in Gigaword.

Table 7.1 highlights general statistics on the number of predicates collected from both corpora. Two observations are worth noting: While on average the number of explicitly realized roles/frame elements per predicate/frame in both data sets is similar, we find more predicate instances in CLMET that in Gigaword. This is due to FrameNet and its fine-grained modeling of lexical units. Also note that the FrameNet lexicon specifies 9.7 frame elements per lexical frame[4] which – despite the fact that this number also comprises non-core arguments – is much larger than what can explicitly be labeled by the SRL systems.

	CLMET	Gigaword
# explicit roles	21.9M	264.0M
# predicate instances	9.5M	122.5M
# roles per predicate	2.3	2.2
# predicates per sentence	7.6	4.2

Table 7.1: Statistics on the number of explicit fillers used for training protofillers.

Regarding the distributional component, we experimented with a variety of distributed word representations: We chose out of the box vectors; Collobert et al. (2011a), dependency-based word embeddings (Levy and Goldberg, 2014) and the pre-trained Google News vectors from *word2vec*[5] Mikolov et al. (2013a). Using the same tool, we also trained custom embeddings (bag-of-words and skip-gram) with 50 dimensions on our two corpora.

7.3 Evaluation

In order to assess the usefulness of our approach, a quantitative evaluation has been conducted on two iSRL test sets which have become a de facto standard in this domain: a collection of fiction novels from the SemEval 2010 Shared Task with manual annotations of null instantiations (Ruppenhofer et al., 2010), and Gerber and Chai (2010)'s augmented NomBank data set. Table 7.2 shows some general statistics on the number of implicit roles and candidate phrases involved in our experiments. In order to have a comparison with the supervised approaches referred to in this study, we also provide the size of the training data.

7.3.1 SemEval Data

In Table 7.3, we report the classification scores for the (*NI-only*) null instantiation *linking* task on the SemEval data, given the parsed candidate phrases and the

[3] `https://code.google.com/p/mate-tools/`

[4] `https://framenet.icsi.berkeley.edu/fndrupal/current_status`, accessed March 2016.

[5] `https://code.google.com/p/word2vec/`

	SemEval	NomBank
# predicate instances		
in training set	1,370	816
in test set	1,703	437
# implicit arguments		
in training set	245	650
in test set	259	246
# of candidate phrases per predicate instance	27.6	52.2
proportion of single tokens	63.4%	47.9%
proportion of phrases	36.6%	52.1%
∅ length of candidate phrase (in tokens)	5.8	7.1

Table 7.2: Statistics on implicit arguments and candidate phrases from the test sections of the two evaluation sets.

gold information about the missing frame element.[6] We include the results of our best-performing configuration, obtained from protofillers trained on the late modern English texts and Collobert et al. (2011a) embeddings (C&W) with the search space for candidate NIs limited to the current and previous sentence. As a reference, we compare our results to the two best models (M_1 and $M_{1'}$) by Silberer and Frank (2012), the vector-based resolver (vec) by Gorinski et al. (2013) – which is most similar to ours – and, finally, their ensemble combination of four semantically informed resolvers by majority vote (4x).

		P	R	F_1
Silberer and Frank (2012)	M_1	30.8	25.1	**27.7**
Silberer and Frank (2012)	$M_{1'}$	**35.6**	20.1	25.7
Gorinski et al. (2013)	vec	21.0	18.0	19.0
Gorinski et al. (2013)	4x	26.0	24.0	25.0
Schenk and Chiarcos (2016)	C&W embeddings	27.2	**25.7**	26.4

Table 7.3: NI linking performance on the SemEval test data.

The figures in Table 7.3 suggest that our approach clearly outperforms the vector-based method by Gorinski et al. (2013) and is best in terms of overall recognition rate (recall) among all systems. One potential reason for that might be that, in contrast to the vec resolver, we do not compute mere context vectors but do rely on the valuable annotations obtained from explicit SRL structures. Also, we do not restrict our analysis to *head* words only, as we have seen that syntactic

[6] This avoids error propagation from NI *detection* and allows us to directly compare our results to previous approaches on the same task. Note that Laparra and Rigau (2012) only report their accuracies for the full pipeline.

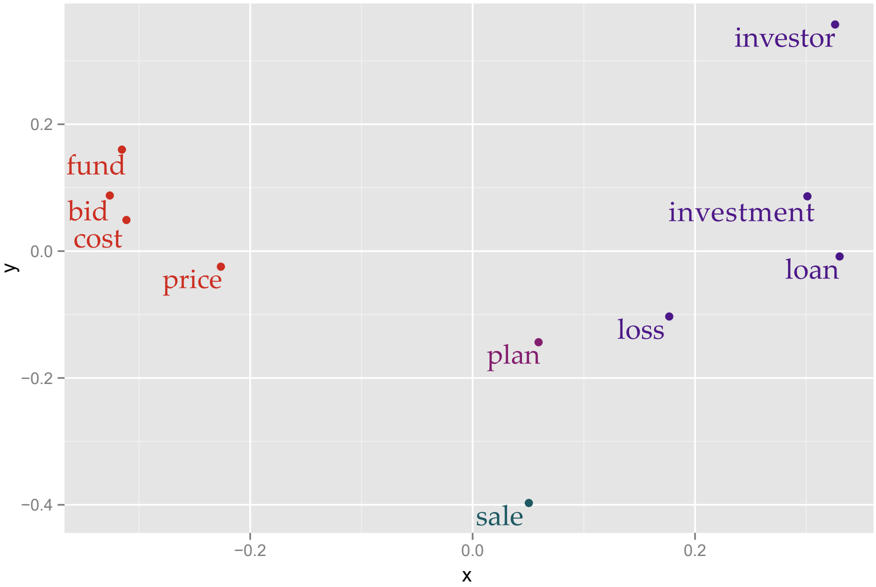


Figure 7.2: Clustered projection of the ten nominal predicates from Gerber and Chai (2010) in protofiller space.

information from function words is crucial for the resolution of null instantiated roles, too. Moreover, our distributional protofiller method is highly competitive with the state-of-the-art performance by Silberer and Frank (2012), yet does not yield better results in terms of F_1 score. Note however that, in contrast to their approach, ours is largely unsupervised and neither relies on gold coreference chains, nor do we need to train on implicit semantic roles in a supervised setting. An error analysis of our method reveals that it is particularly effective for NIs encountered in the *same sentence* as the target predicate (44.4% accuracy), which seems plausible given the contextual setup in which protofillers are derived.

7.3.2 NomBank Data

Compared to the SemEval data, Gerber and Chai (2010)'s augmented NomBank resource covers only ten nominal predicates, which allows us to nicely visualize the distributional profile based on their prototypical fillers. For each predicate, we simply concatenate all per role computed protofillers and apply multidimensional scaling to project the so obtained vectors onto two dimensions (cf. Figure 7.2).

We observe that the predicate grouping is now based on the prototypical fillers that they co-occur with: In the Wall Street Journal texts, *loss*, *loan* and *investment* are similar because their proto-agents (A0 fillers) who lose, lend and invest resp. are semantically shared (i.e. companies, banks). Similarly, *bid*, *cost* and *fund* are related in that the targets or commodities (A2) are all money-financed. Finally, the predicates *sale* and *plan* are to be expected as outliers as they are less homogeneous in their prototypical argument structure.

	B	Gerber & Chai			Laparra & Rigau			**Proto C&W**			**Proto W2Vcbow**		
predicates:	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
sale	36.2	47.2	**41.7**	**44.2**	41.2	39.4	40.3	**61.0**	29.6	39.8	60.8	26.8	37.2
price	15.4	36.0	32.6	34.2	**53.3**	**53.3**	**53.3**	14.7	25.8	18.7	21.8	36.6	27.3
investor	9.8	36.8	40.0	38.4	**43.0**	39.5	**41.2**	22.5	48.3	30.7	24.1	**57.2**	33.9
bid	32.3	23.8	19.2	21.3	**52.9**	**51.0**	**52.0**	30.4	31.5	30.9	40.0	41.5	40.7
plan	38.5	**78.6**	**55.0**	**64.7**	40.7	40.7	40.7	41.1	43.2	42.1	44.3	51.0	47.4
cost	34.8	**61.1**	**64.7**	**62.9**	56.1	50.2	53.0	32.5	19.1	24.0	49.9	29.3	36.9
loss	52.6	**83.3**	**83.3**	**83.3**	68.4	63.5	65.8	54.8	73.1	62.6	54.7	63.8	58.9
loan	18.2	**42.9**	33.3	37.5	25.0	20.0	22.2	33.9	**49.0**	**40.1**	33.2	44.2	37.9
investment	0.0	40.0	25.0	30.8	**47.6**	**35.7**	**40.8**	29.1	21.8	24.9	39.2	34.3	36.6
fund	0.0	14.3	16.7	15.4	66.7	33.3	44.4	**100.0**	**33.3**	**50.0**	75.0	25.0	37.5
Overall	26.5	44.5	40.4	42.3	**47.9**	**43.8**	**45.8**	30.2	35.2	32.5	33.5	39.2	36.1

Table 7.4: Classification scores of the implicit argument labeling on the NomBank test section. Baseline *B* from Gerber & Chai (2010): uses previous occurrence of same predicate. Gerber & Chai (2010): supervised logistic regression classifier trained on implicit fillers. Laparra & Rigau (2013): algorithm based on coherence relationship between predicates and fillers. Our best-performing protofillers are obtained by Collobert et al. (2011) embeddings (**Proto C&W**) and custom trained vectors (**Proto W2Vcbow**) using Gigaword SRL annotations.

We have empirically evaluated our protofiller method also on this data set: Table 7.4 reports the classification scores for implicit argument resolution compared to the state-of-the-art (Laparra and Rigau, 2013a). We restrict the search for implicit arguments to certain predicate-specific parts-of-speech, since some syntactic constituents (e.g., SBAR) never occur as implicit arguments. For choosing the final implicit arguments for each individual predicate instance, we follow the same deterministic strategy as described in Gerber and Chai (2010), which informally states that, if a certain role is not overtly expressed (within a chain of mentions of the same predicate in previous sentences), it is an implicit candidate. POS lists and cosine similarity thresholds which trigger an actual prediction have been optimized on the development set. The context window for candidate NIs is optimal for the current and previous two sentences in our setting, which explains why the number of candidate constituents is approximately twice as large for the NomBank predicates (cf. Table 7.2).

Our best-performing protofillers are again obtained by Collobert et al. (2011a) embeddings substituting explicit SRL annotations in the Gigaword corpus, and with custom-trained embeddings using the continuous bag-of-words model. Overall, our results significantly exceed the highly informed baseline but cannot beat the state-of-the art on this test set. For some predicates, the protofillers seem to generalize better (higher recall), and in particular for the low-frequency predicates (*fund*), precision can be increased. Also, we found that the dependency-based word embeddings perform slightly worse (not shown) compared to our optimal two configurations. This might be due to the fact that the inherent properties of dependency-based contexts mostly focus on relations between semantically valuable nouns, ignoring ("skipping") functional words and categories.[7] The same pertains to the pre-computed Google News vectors which come with a frequency cutoff excluding stop words, again a constraint which is harmful for the correct identification of implicit roles. Furthermore, skip-gram embeddings perform significantly worse than the embeddings derived by the continuous bag-of-words implementation (relative decrease in F_1 by more than 30%). Finally, we observed that inferring implicit roles for nominal predicates is much more challenging because our collected fillers exhibit a much greater variation. For example, the protoagents of *loan* can roughly be divided into two categories, institutions and countries. This in turn introduces noise and has a negative effect on the quality of the singleton protofillers which by vector average capture neither of the two groups perfectly. Promising alternatives could operate on (topic-like) protofiller **clusters** which we leave for future work.

7.4 Summary

In this chapter, we have described a lightweight approach for the resolution of implicit semantic roles which does not rely on manual gold annotations. For each

[7] This is also nicely illustrated in Levy and Goldberg (2014).

predicate-specific role, our method generalizes over explicit SRL-guided annotations incorporating pretrained word embeddings. This allows us to capture their idiosyncratic properties and use the so-inferred protofillers to find null instantiated roles by means of distributional similarity.

Our method has proven to be generally useful, in particular on the SemEval data, where it is competitive with supervised systems. Its greatest benefit stems from its simplicity and from the fact that it allows us to induce null-instantiated roles for arbitrary predicates. As it is applicable even if no iSRL training data is available, it represents a promising technique to address iSRL data scarcity issues.

In our experiment, we employed PropBank/NomBank-style (i)SRL annotations, and our general design clearly benefits from using small-scale inventories of semantic roles. It should be noted though, that our approach is not restricted to any particular SRL tagset, but can be equally applied to other role inventories with similar degrees of consistence and size. Beyond SRL annotations in a strict sense, this might even extend to syntactic dependency annotations that are occasionally taken as a substitute for semantic roles proper, and in this regard, our approach resembles Peñas and Hovy (2010). In particular, we see potential in combining our experiments with on-going efforts to cross-lingual projection, adaptation and harmonization of syntax annotations along the lines of Sukhareva and Chiarcos (2014, 2016) and related approaches based on frameworks such as the Universal Dependencies (Nivre, 2015, UD).[8] If successful, an adaptation using grammatical relations rather than semantic roles represents a promising possibility to create iSRL annotation and iSRL annotation tools for other languages, as Universal Dependencies are becoming increasingly available for major and low-resourced languages and can be projected to others.

Data: The collection of pretrained prototypical vectors described in this chapter is publicly available at `http://www.acoli.informatik.uni-frankfurt.de/resources.html`.

[8] `http://universaldependencies.github.io`

Summary

This part of the dissertation has been concerned with textual implicit information as typically evoked on the word or phrase level. As a theoretical foundation, the first section of Chapter 5 introduced predicate-argument structure and the closely related thematic relationship between predicates and their argument realization—modeled in terms of semantic roles. We have seen that when arguments are locally unexpressed they can sometimes be linked to an antecedent in the discourse and thus be recovered from the context. The previous literature on such *implicit semantic roles* has distinguished two types: anaphoric and generally resolvable vs. existential and non-resolvable null instantiations (cf. Section 5.1). The specification of semantic roles and their interaction with distinct predicates (for different word senses and parts-of-speech) has been encoded in hand-designed lexicons (for instance, in FrameNet) or by large-scale corpus annotation efforts (PropBank or NomBank).

These resources, described in Section 5.1, have established the basis for automated implicit semantic role labeling (Section 5.2). In general, iSRL is cast as a three-step pipeline approach involving i.) the detection of a missing implicit role for a given predicate instance, ii.) the classification of the role's interpretation, i.e. to decide whether it is resolvable or not, and, finally, iii.) the linking process to an appropriate candidate filler in the preceding or following (non-local) discourse context. Various approaches have been summarized in Section 5.2.2 distinguishing the semantically fine-grained FrameNet-style from the more general PropBank-style parsing.

Based on the prior work on iSRL, we have addressed specific issues related to data scarcity of manually annotated iSRL resources, failure of supervised systems to generalize to unseen events, and have motivated the need for improvements both in implicit argument detection and in the resolution process.

In particular, Chapter 6 has introduced a novel (memory-based) method for the detection of implicit roles in PropBank format. The novelty can be seen in that—instead of a static dictionary lookup—implicit roles are detected based on their likelihood of occurrence as estimated from explicit role annotations in large corpora. The advantage is that the proposed method is language-independent, applies to the detection of both resolvable and non-resolvable, as well as core and non-core arguments, and is more flexible than costly, hand-designed rules and dictionaries—given as the only requirement the availability of a standard SRL parser. Future work should readdress the manually introduced parameters of the model, their configuration and optimization, which in its current version seems to be a rather ad-hoc procedure, but we conjecture that the model (esp. the weights) could be straightforwardly tuned and automatized within a neural network setting.

Following the fruitful idea of a statistical generalization for implicit argument detection, an adaptation of the method has been applied to implicit role resolution, too, which was presented in Chapter 7. We have described a largely unsupervised approach based on the idea to generate predicate-specific prototypical role fillers induced from large amounts of explicit SRL annotations. The technique finds fillers for implicit roles by means of distributional similarity, and generalizes to the same extent over both FrameNet and PropBank labels. Moreover, it has demonstrated to be knowledge-poor, as no manual gold annotations are required. Most importantly, the proposed method is even competitive with supervised systems on a standard evaluation set. We have observed that our method captures important syntactic and semantic aspects within a singleton prototype, however, it remains an interesting research direction to pursue for future work, whether the induction of more than just one filler for the same semantic role will lead to performance improvements in the resolution process. We have already hinted at the fact that proto-agents (as a result of a dimensionality-reduced projection) for a particular predicate can be classified into two or more semantic groups, in which entities share similar (sub)properties. Prototypical clusters could be the result of a more fine-grained assessment of our vectors. Ultimately, they will be more sparse, but could represent a useful resource, e.g., when integrated into a machine learning framework, or by supplementing the standard out-of-the-box embeddings used in recurrent neural network iSRL models, e.g., similar to the work described in Do et al. (2017).

In the last part of this thesis, we thoroughly investigate a potential synergy effect between the two phenomena of implicit semantic roles and implicit discourse relations. We have already pointed out in the introduction of this thesis that theoretically motivated interfaces between (i)SRL and discourse parsing exist. For example, verbs of implicit causality (Garvey and Caramazza, 1974) affect our interpretation, for instance, when we encounter the word *frighten*. These cues lead to certain pragmatic inferences made by the hearer, who would expect specific information to be conveyed in the continuation of a text. Theoretical models assume that subsequent sentences will in fact provide answers to implicit questions that are being generated during comprehension. Interestingly, the connection between implicit causality verbs and the follow-up information provided by a subsequent sentence is oftentimes expressed in the form of an implicit discourse relation.

In the ensuing Chapter 8, we make direct reference to these observations and base our proposed methodology on the popular *Question Under Discussion (QUD)* model of discourse interpretation, cf. von Stutterheim and Klein (1989); Kuppevelt (1995); Benz and Jasinskaja (2017), inter alia. We demonstrate that apart from single local cues, such as implicit causality verbs, single predicate instances, negation markers, or negatively connoted words (Webber, 2013), *complete predicate-role con-*

stellations together with their implicit semantic arguments are strong indicators in licensing global discourse coherence. To this end, we introduce a correlation study to assess the mutual dependence of sentence-internal cues on the discourse. Our study builds on large-scale generalizations to quantify this relationship in capturing statistical regularities between them. Our methodology is novel in the sense that the interaction between both local and global structures is modeled in terms of implicit information, while recent similar approaches only considered overt roles, cf. Roth (2017). We demonstrate that there exists a measurable interplay between akin local and global information. We shed some further light on entity-based coherence and show that these relations are special insofar as they are influenced by cues which are distinct from those that account for the remainder of discourse senses in the Penn Discourse Treebank.

As part of the motivation for our work in this thesis we have also seen, that according to Centering Theory (Grosz et al., 1995) and studies on the interpretation of discourse phenomena (Rohde et al., 2007; Kehler and Rohde, 2017), *pronouns* are a driving force in the establishment of local coherence in entity relations. However, apart from few similar constructions in English or German, some major languages such as Chinese or Japanese heavily lack the presence of pronouns in many contexts. Linguistic studies refer to these cases as instances of *zero anaphora* (Fillmore, 1986; Tao, 1996). It is noteworthy that Givón (1983) and Tao (1996), respectively, place them on the extreme end of a topic continuity scale, arguing that zero anaphors typically relate to easily accessible (i.e. continuously mentioned) subjects or referents, whereas, for instance, full NP's go along with a rather discontinuous topic shift in discourse.[9] Based on these observations, we argue that these theoretically motivated transition shifts can be indirectly associated to discourse senses, for instance, in the Penn Discourse Treebank, where continuations would relate to instantiations of EXPANSION relations and shifts most likely to COMPARISON.

Crucially, we conjecture that, in order to account for an overall integration of events and their participants into the cohesive structure of discourse, it becomes evident that the resolution of implicit semantic roles is required, for instance, for core agents in the local context (zero anaphora), as well as the recovery of links to appropriate antecedents or postcedents in the non-local context. A number of studies and computational approaches—methodologically very similar to iSRL—have been suggested already for the resolution of zero anaphora in Chinese (Yeh and Chen, 2001; Chen and Ng, 2013; Iida et al., 2007; Chen and Ng, 2016, inter alia). Moreover, in practical iSRL applications it has been shown that this link between local implicit information and global discourse structure seems reasonable, for example, by making use of anaphoric mentions as part of coreference chains, cf. Silberer and Frank (2012).

In Chapter 9, we therefore outline a specifically-tailored experiment to test this mutual dependence in the form of SRL beyond the sentence boundary. The ex-

[9] Note that this view is analogous to the principles described in Centering Theory which postulate that CONTINUE transitions are to be considered the norm.

periment has the goal to set discourse coherence on the basis of local predicate role patterns and to test whether an existent discourse architecture can feasibly be ported to the joint domain of labeling semantic arguments in a cross-sentential setting. We demonstrate that coreferentiality can, in fact, be encoded in local semantic arguments, and that our proposed computational model treats coreferential roles in a very similar way to hand-annotated entity relations which stem from a discourse treebank.

Chapter 10 revisits entity-based coherence in narrative stories and presents an extension to our previous methods in terms of an algorithm to distinguish correct from incorrect story continuations. Our method models continuation transitions, the most anticipatory coherence transition in free texts according to Centering Theory. We cast the task of story understanding as a special case of discourse continuity, instead of entity continuity as typically implemented in anaphora or coreference resolution systems. Specifically, the detection of implicit discourse structure will be adapted to the challenging task at hand. We demonstrate that story coherence modeling is complex and requires the proper detection of latent implicit temporal and causal relationships. In the next chapter, we show that an incorporation of implicit discourse information can be a driving factor in successfully handling such complex semantic processing tasks.

To summarize, the ensuing two chapters are concerned with experiments to bridge the gap between the two types of implicit information which we broadly classified as holding between sentences and evoked within sentences. We introduce two experiments and again roughly distinguish two variants: one to measure the effect of implicit roles on implicit discourse in a bottom-up fashion (Chapter 8), and one—vice versa—to map discourse top-down onto the basis of semantic roles (Chapter 9). Finally, Chapter 10 presents an extension of our proposed methods and describes its applicability to model discourse-driven entity coherence in narrative stories.

Part IV

Bridging the Gap

Chapter 8

Extending Intra-Sentential Semantics to the Discourse

8.1 Motivation

As part of the theoretical motivation in this thesis, we claimed that discourse relations and implicit semantic roles are closely related insofar as an account of global discourse coherence is properly licensed by implicit roles and can only then be computationally realized when uninstantiated roles and their links to antecedents are successfully recovered. Practical iSRL implementations and approaches to zero anaphora have successfully built on this assumption (Silberer and Frank, 2012; Chen and Ng, 2013) but—to the best of our knowledge—no quantitative study has explicitly been designed to test whether this relationship exists, and—if it exists—how strong the interaction is, and how it could be measured. In order to confirm the hypothesis of a mutual interdependence between implicit roles and implicit discourse structure, we postulate that this interrelation should manifest itself in a correlation of implicit information of *similar or related types*. We elaborate on details in the following.

The experiment in this chapter is inspired by the recent work of Asr and Demberg (2012) and, in particular, by Asr and Demberg (2015) who aim at finding a principled explanation for the presence versus absence of discourse markers between discourse relations in free text. Their work is grounded on information theory and builds on the three pillars of the *Uniform Information Density* theory (Levy and Jaeger, 2006), the Gricean *Maxim of Quantity* (Grice, 1975), and the notion of *Surprisal* in text (Hale, 2003). A combination of different aspects of these theories relates to the hypothesis that if informative information is uniformly distributed and submitted from speaker to hearer, there is no surprisal to be expected on the hearer side, and thus the discourse relation is predictable. Therefore, a discourse marker can (or even should) be omitted, resulting in an implicit discourse relation.

However, the *Uniform Information Density* principle states that also *unexpected* relations can have no markers, provided that *other* strong cues are present in the local context. To this end, Asr and Demberg (2015) address this phenomenon by looking into one specific implicit discourse relation (namely EXPANSION:Alternative:chosen alternative, which is in its explicit form typically signaled by *instead*) and a single cue—a negation maker. The authors demonstrate in a statistical experiment based on the Penn Discourse Treebank (Prasad et al., 2008, PDTB) that when a relation is highly predictable due to the presence of the negation in the first argument the discourse marker can be dropped. Specifically, a negation and sentence polarity in implicit relations tend to be highly indicative of this particular sense. These findings are also in line with previous corpus experiments by Webber (2013).

Furthermore, Asr and Demberg (2015) note that a special type of cue, namely *implicit causality verbs* (Garvey and Caramazza, 1974), have a strong tendency towards implicit discourse relations of type CONTINGENCY:Cause:reason, and refer in this context to an original lab study by Rohde and Horton (2010), who conducted a detailed investigation on these cues. Asr and Demberg (2015) assert, however, that the effect of implicit causality verbs involved in this study cannot be verified and tested on a larger scale, for instance on the PDTB, partly because of the nature of the experiment that involved only artificially short constructions in which the specific verbs appeared—as opposed to real expository texts which normally exhibit a far more sophisticated syntactic structure.

On a related note, Kehler and Rohde (2017) have very recently investigated the significance of the *Question Under Discussion (QUD)* model of discourse interpretation, cf. von Stutterheim and Klein (1989); Kuppevelt (1995); Benz and Jasinskaja (2017), inter alia. The general concept of QUDs states that implicit questions are being generated, updated and answered by subsequent sentences—as the discourse unfolds. A participant in a conversation, say, a hearer, has certain expectations not only on what kind of utterances the speaker will produce next, but also on what types of coherence relations are likely to follow from the current point in time. These expectations together with the QUDs are dependent on contextual cues, e.g., influenced by implicit causality verbs. To this end, Kehler and Rohde (2017) presented a dialogue experiment, which demonstrated this effect, by showing that some predicate contexts tend to evoke causal discourse relations, whereas others are more likely to be followed by elaborations or explanations.

8.1.1 A Correlation Study

In the experiment outlined in this chapter, we build on the work of Asr and Demberg (2015) and Kehler and Rohde (2017) and propose a large-scale generalization of sentence-internal cues, for which we assume that they affect the expectation of certain types of implicit discourse relations. Instead of just looking at artificially constructed examples, a single negation, or a single type of verb, we consider all predicate instances together with their co-occurring semantic roles in large corpora as potential features to explain, in a more sophisticated model, the driving force behind implicit discourse relations. Specifically, we address the question of whether and how sentence-internal information could give rise to its superordinate extra-sentential discourse structure. To explain the interaction between local semantic roles on the one hand and non-local discourse structure on the other, *we consider both sources of information as implicit*. This means that in order to account for implicit discourse relations, we assume a certain contribution by implicit semantic roles which stems from within the embedded discourse arguments. Note that this strategy stands in a great contrast to all previous attempts to describing implicit discourse relations. The reason is that prior works have only dealt with explicit surface features. For example, overt semantic roles were considered useful evidence (Roth, 2017). Other approaches have pinpointed overt words as indicators or—in order to obtain better generalizations—have substituted those with word embeddings (Feng and Hirst, 2012; Mihaylov and Frank, 2016a). This chapter, however, describes the first experiment of its kind to interrelate two types of non-overt evidence, and to build a bridge between implicit semantic roles and implicit discourse relations. Ultimately, this means that the local scope and extent of "missing" information needs to be inferred from individual predicates occurring in natural contexts. Later on, this acquired information needs to be employed as the sole basis for modeling implicit (global) discourses. To this end, we propose a *correlation study* intended to quantitatively measure the effect of the interaction. Since no discourse treebanks with implicit semantic role annotations exist that could be directly exploited for our purposes, we consider iSRL predictions and propose a two-step approach, whose details are described as follows.

In the first step, we acquire general co-occurrence information on semantic roles for any given predicate. More precisely, we compile a *background knowledge base* of predicate instances together with the various realizations of explicit semantic role patterns from automatically annotated instances in large amounts of free texts. This procedure of collecting co-occurrence information is restricted to a within-sentence basis.

Then, by means of corpus statistics over the knowledge base, we estimate the relative importance of any (predicate-specific) role as part of a specific role constellation. The goal here is to develop a mechanism for pinpointing those semantic roles which—in case they are *not* overtly expressed in a given sentence—could nonetheless represent a crucial piece of additional (implicit) information and could thus be potentially relevant *outside of the local predicate context* towards the extra-sentential discourse structure. The described procedure works in a bottom-up fashion, i.e. by collecting statistics sentence-internally and applying them to the inter-sentential discourse.

This procedure is best illustrated with an example. For this purpose, consider the following two (unrelated) sentences in examples (18) and (19).[1]

(18) [Some economists and government officials$_{\texttt{A0}}$] here are[n't$_{\texttt{AM-NEG}}$] [applauding$_{pred}$].[2]

[1] Taken from the PropBank (Palmer et al., 2005) online framesets available at: `http://propbank.github.io/`

[2] `http://verbs.colorado.edu/propbank/framesets-english-aliases/applaud.html`

(19) [I$_{\texttt{A0}}$] do[n't$_{\texttt{AM-NEG}}$] [know$_{pred}$] and [I$_{\texttt{A0}}$] do[n't$_{\texttt{AM-NEG}}$] [care$_{pred}$].[3]

In these two sentences, three predicates and their accompanying semantic roles are overtly realized. Each occurrence of *applaud*, *know* and *care* comes with exactly one explicit agent role (`A0`) and one adjunct negation role (`AM-NEG`), resulting in three structurally *equivalent* role patterns. However, given the fact that we are dealing with completely different predicates (which typically appear in different meaning contexts), it might be worth investigating to what extent one would expect *additional* information to be encoded in either of the two sentences, provided that these sentences stand in a certain discourse relation with other sentences.[4] We assume that this type of information is *latent*, i.e. distinct for each predicate and only present in an underlying form, but *not* overtly expressed in the respective sentences. For the benefit of generalizability, we conjecture that this additional information should best be *measured in terms of implicit roles*, and specifically by their probability of occurrence with reference to the background knowledge base. Intuitively, *not applauding* strongly evokes an implicit question asking for a *reason* (i.e. it invokes a causality role, `AM-CAU`, explaining *why?*), whereas *not knowing* and *not caring* would rather not—at least not to the same extent. Note again that this view is analogous to the well-established QUD models of discourse interpretation in which clauses provide *answers* to implicit questions stated in the preceding context, cf. Kehler and Rohde (2017).

In sum, the main contributions of this correlation study are two-fold:

1. the establishment of a fine-grained quantitative estimate of the likelihood of missing, unexpressed, i.e. implicit information among different predicate contexts and their role patterns, and

2. an assessment of how this unexpressed information contributes to the global discourse context.

Ultimately, in the above example (18), one might find the reason for *not applauding* in the immediate discourse context, which is indeed the case when we inspect the following sentence in the Penn Discourse Treebank from which the examples are drawn. Here, the implicit sense that holds between the two arguments is of type CONTINGENCY:Cause:reason.

(20) `Arg1`: Some economists and government officials here aren't applauding.

`Arg2`: They fear that the boom may be too big for Japan's or anyone else's good.

[3] `http://verbs.colorado.edu/propbank/framesets-english-aliases/care.html`

[4] Asr and Demberg (2015) refer to this phenomenon as *relational surprisal* which is influenced by the different cues within the sentences.

Implicit discourse sense: CONTINGENCY:Cause:reason[5]
Inferred connective: *because*.

Finally, in order to demonstrate the effect of sentence-internal implicit information on the sentence-external discourse structure, we present a lightweight evaluation in which we aggregate predicate-wise implicit information for specific semantic roles in analogy with a set of hand-annotated discourse relation pairs from the Penn Discourse Treebank. We illustrate that certain semantic roles of "missing" within-sentence information do indeed positively correlate with the surrounding discourse senses of similar types in which the respective sentences are embedded. To the best of our knowledge, this is the first study which involves a quantitative assessment of implicit information by transition between local to non-local contexts.

In what follows, we first describe the setup of the background knowledge base (Section 8.2) and how we quantify implicit information from explicit role patterns. In Section 8.3 we evaluate our hypothesis on the PDTB senses and finally, in Section 8.3.4, discuss our findings in the light of related research. Section 8.4 concludes this chapter with a short summary.

8.2 Estimating Implicit Information from Large Corpora

8.2.1 Compilation of a Predicate-Role BKB

For this experiment, we follow the approach described in Chiarcos and Schenk (2015b) and set up a background knowledge base (BKB) of predicates together with their overtly realized argument and adjunct roles, as they co-occur naturally in large amounts of texts. The idea here is to first obtain generalizations over mass data which will then be applied to specific predicate contexts. For these purposes, we employ *mate-tools*[6] (Björkelund et al., 2009) and automatically label the complete English Gigaword corpus (Graff and Cieri, 2003), a large collection of English newswire texts, with semantic role annotations for nominal and verbal predicates distinguished by their word senses.[7] For an overview of the annotated corpus and predicate/role statistics, cf. Table 8.1. The source of the BKB consists of approximately two billion words, annotated with semantic roles. Note that, overall, verbal predicate instances are more frequent (55%) compared to nominal ones (45%). However, for the number of predicate types, we observe an opposite

[5] Document ID `wsj_1037` of PDTB training section.

[6] `https://code.google.com/archive/p/mate-tools/`

[7] Sentence boundary detection is performed with Dan Gillick's *splitta* implementation (`https://pypi.python.org/pypi/splitta/0.1.0`). All sentences longer than 60 tokens were removed. We use *mate*'s built-in tokenization mechaninsm along with the complete NLP pipeline for all annotations and remove all "noisy" predicates consisting mainly of digits and temporal expressions such as *90's*, *25-year old*, etc.

trend. The number of extracted semantic roles amounts to almost 720 million. Note, however, that roles are not double-counted, i.e. two separate mentions of an A0 role for the same predicate instance are counted only once towards the BKB.

	Overall		
# documents	4.1M		
Newswire sources:	AFE, APW NYT, XIE		
# sentences	94.9M		
# tokens	**1.97B** (orig. 1.76B)[a]		
avg sentence length (in tokens)	13.1		
	Overall	Nominal	Verbal
# predicate instances	357.4M	160.7M (45.0%)	196.7M (55.0%)
# predicates per sentence	3.8	1.7	2.1
# predicate types (filtered)	567k	317k (55.9%)	250k (44.1%)
# explicit roles	718.8M	240.1M (33.4%)	478.7M (66.6%)
# explicit roles per predicate	2.01	1.49	2.43
# distinct role patterns	31.8k		
cumulative frequency of all role patterns	351.4M		

Table 8.1: Global corpus statistics and counts of automatically labeled predicate instances, explicit roles, and role patterns in Gigaword

[a] This number counts only whitespace separated tokens as reported in the official Gigaword statistics: `https://catalog.ldc.upenn.edu/ldc2003t05`.

Detailed information on the individual semantic roles and their proportions in Gigaword are given in Table 8.2. For better interpretation we include role descriptions from both the original shared task paper by Carreras and Màrquez (2005) and the excellent introduction by Jurafsky and Martin (2017). With both nominal and verbal predicates, (proto-)patient roles are most frequent, followed by arguments A0 and A2, which is plausible as these are most well-defined across predicates. Interestingly, for the verbal predicates the distribution among modifier roles exhibits a greater variability compared to the nominal counterparts. Temporal adjuncts are predominant for verbs, while modifiers of manner are most frequent for nouns. Semantic roles which are realized in other parts of the sentence (R-X) are listed for reasons of completeness but are statistically underrepresented in the automatically annotated data.

Finally, Tables 8.3 and 8.4 give an overview of the most frequent role patterns, separately for the number of instances and types, respectively. In total, there are 31,801 distinct role patterns. The most frequent role pattern in terms of instances counted is `A0 A1`, a prototypical subject–object tuple. The most frequent role pattern type is `A1`, which appears most frequently among all predicate types extracted, i.e. roughly for every second predicate type (52.1%). We count the absence of any roles as a separate pattern, which appears approximately 9 million times in the Gigaword data.

8.2.2 Quantifying Implicit Information through the BKB

Given the BKB of role pattern counts extracted form the text corpus, we seek a way to model the predicate-specific effect that individual roles have among the various constellations of role patterns. Returning to the example sentence (18) from the previous section, we want to measure the contribution of an unexpressed **implicit** `AM-CAU` on a verbal predicate *pred*, *given* the joint occurrences of two explicit roles `A0` and `AM-NEG`. This can be modeled straightforwardly in terms of conditional probabilities which we approximate from the BKB by relative frequency estimation:

$$\mathrm{P}(\texttt{AM-CAU} \mid \textit{pred}, \texttt{A0}, \texttt{AM-NEG}) \approx \frac{\#(\textit{pred}, \texttt{A0}, \texttt{AM-CAU}, \texttt{AM-NEG})}{\#(\textit{pred}, \texttt{A0}, \texttt{AM-NEG}) + \#(\textit{pred}, \texttt{A0}, \texttt{AM-CAU}, \texttt{AM-NEG})}$$

Note that for reasons of simplicity and a more intuitive understanding of the effect that individual roles have towards a specific constellation, this probability estimate does not take frequencies of other role (sub-)constellations into account.[8]

Ultimately, we attempt to measure and compare the importance of individual implicit roles as part of same role constellations across different predicate instances—in our specific case between *applaud*, *know* and *care*. Intuitively, two generalizations should hold: An implicit semantic role (here: `AM-CAU`) is of particular relevance to the non-local discourse context of the sentence in which the role is embedded, provided that

1. $\mathrm{P}(\texttt{AM-CAU} \mid \textit{applaud}, \texttt{A0}, \texttt{AM-NEG}) > \mathrm{P}(\texttt{AM-CAU} \mid \textit{applaud}, \texttt{A0})$
2. $\mathrm{P}(\texttt{AM-CAU} \mid \textit{applaud}, \texttt{A0}, \texttt{AM-NEG}) > \mathrm{P}(\texttt{AM-CAU} \mid \textit{know}, \texttt{A0}, \texttt{AM-NEG})$

The first condition states that conditioning on the *full* context is necessary rather than conditioning only on a partial role (sub)set, while the second restriction states that the implicit contribution is *predicate-specific* and does not show the same strong effect among all predicate types. Given the motivation from the previous section, in our specific examples (18) and (19), the probability of a causal role

[8] For instance, the pattern [*pred*, `A0`, `AM-NEG`, `AM-LOC`] shares a common subset with [*pred*, `A0`, `AM-NEG`, `AM-CAU`] but is not considered in the computation.

Role	Description	Nominal	Verbal
A0	(prototypical) agent	26.0%	25.9%
A1	(prototypical) patient	**44.1%**	**35.8%**
A2	benefactive, instrument, end state	14.4%	7.5%
A3	start point, benefactive, instrument	2.0%	0.8%
A4	end point	0.1%	0.7%
A5	miscellaneous	0.0%	0.0%
AM-ADV	general purpose	0.2%	3.2%
AM-CAU	cause/*why?*	0.0%	0.3%
AM-DIR	direction/*where to/from?*	0.0%	0.6%
AM-DIS	discourse marker	0.0%	1.5%
AM-EXT	extent	0.3%	0.0%
AM-LOC	location/where?	3.3%	3.2%
AM-MNR	manner/*how?*	5.4%	2.5%
AM-MOD	modal verb	0.0%	3.4%
AM-NEG	negation marker	0.6%	1.3%
AM-PNC	purpose/*why?*	0.0%	0.9%
AM-PRD	predication	0.0%	0.0%
AM-REC	reciprocal	0.0%	0.0%
AM-TMP	temporal/*when?*	3.5%	7.3%
R-A0		0.0%	1.9%
R-A1		0.0%	1.0%
R-A2		0.0%	0.0%
R-A3		0.0%	0.0%
R-A4		0.0%	0.0%
R-A5		0.0%	0.0%
R-AM-ADV		0.0%	0.0%
R-AM-CAU		0.0%	0.0%
R-AM-DIR		0.0%	0.0%
R-AM-DIS		0.0%	0.0%
R-AM-EXT		0.0%	0.0%
R-AM-LOC		0.0%	0.2%
R-AM-MNR		0.0%	0.1%
R-AM-MOD		0.0%	0.0%
R-AM-NEG		0.0%	0.0%
R-AM-PNC		0.0%	0.0%
R-AM-PRD		0.0%	0.0%
R-AM-REC		0.0%	0.0%
R-AM-TMP		0.0%	0.5%
# instances:		240.1M	478.7M
		100.0%	100.0%

Table 8.2: Statistics on the proportions of extracted core arguments (AX), modifier roles (AM-X), and roles which are realized in other parts of the sentence (R-X)

Rank	Role Pattern	**Instance** Frequency	Proportion
1	A0 A1	71.4M	20.3%
2	A1	68.9M	19.6%
3	A0	19.0M	5.4%
4	A1 A2	18.9M	5.4%
5	A0 A2	10.6M	3.0%
6	A0 A1 AM-TMP	9.7M	2.8%
7	A2	9.4M	2.7%
8	no roles	9.0M	2.6%
9	A0 A1 A2	8.7M	2.5%
10	A1 AM-TMP	6.3M	1.8%
11	A1 AM-MNR	6.0M	1.7%
12	A1 AM-LOC	4.3M	1.2%
13	A0 A1 AM-LOC	4.0M	1.1%
14	A0 A1 AM-MNR	3.8M	1.1%
15	A0 A1 AM-ADV	3.7M	1.1%
16	A0 A1 R-A0	3.2M	0.9%
17	A0 A1 AM-MOD	3.0M	0.8%
18	AM-MNR	2.6M	0.7%
19	A0 AM-MNR	2.5M	0.7%
20	A0 A1 AM-DIS	2.2M	0.6%
21	A0 AM-LOC	2.1M	0.6%
22	A1 A3	2.0M	0.6%
23	A0 AM-TMP	2.0M	0.6%
24	A1 A2 AM-TMP	2.0M	0.6%
...	...	...	...
699	A2 AM-MOD AM-TMP	9,666	2.7e−5%
700	A1 AM-DIS AM-LOC AM-MOD	9,579	2.7e−5%
701	A1 A2 AM-ADV AM-PNC	9,572	2.7e−5%
...	...	...	...
20,299	A2 AM-ADV AM-DIS AM-LOC AM-MOD AM-NEG	2	5.7e−9%
...	...	...	...
	Cumulative frequency of pattern instances:	351.4M	100.0%
31,801	# distinct patterns		

Table 8.3: Statistics on different role pattern **instances** in Gigaword

Rank	Role Pattern	**Type** Frequency	Proportion among Predicate Types
1	`A1`	295,667	52.1%
2	`A0 A1`	211,190	37.2%
3	`A0`	140,177	24.7%
4	`A1 AM-MNR`	78,325	13.8%
5	`no roles`	74,019	13.0%
6	`A1 AM-TMP`	53,086	9.3%
7	`A1 AM-LOC`	48,087	8.5%
8	`AM-LOC`	47,453	8.4%
9	`A1 A2`	44,297	7.8%
10	`A0 A1 AM-TMP`	39,429	7.0%
11	`A0 AM-MNR`	37,974	6.7%
12	`A0 A1 AM-MNR`	37,119	6.5%
13	`AM-MNR`	35,238	6.2%
14	`A0 A1 A2`	32,429	5.7%
15	`A0 AM-TMP`	31,072	5.5%
16	`AM-TMP`	31,026	5.5%
17	`A0 A1 AM-LOC`	30,155	5.3%
18	`A0 AM-LOC`	28,831	5.1%
19	`A2`	27,535	4.9%
20	`A0 A2`	26,574	4.7%
21	`A1 AM-ADV`	22,919	4.0%
22	`A0 A1 AM-ADV`	21,601	3.8%
23	`A1 AM-NEG`	19,029	3.4%
24	`A0 A1 R-A0`	18,312	3.2%
...	...	...	...
699	`A0 AM-ADV AM-CAU`	728	0.1%
700	`A0 A1 AM-ADV AM-LOC AM-MNR AM-TMP`	725	0.1%
701	`A0 A1 AM-ADV AM-DIS AM-LOC AM-MOD`	723	0.1%
...	...	...	...
20,299	`A0 AM-ADV AM-DIR AM-EXT AM-TMP`	5	3.5e−6%
...	...	...	...
	Total number of **predicate types**:	567,225	
31,801	# distinct patterns		

Table 8.4: Statistics on different role pattern **types** in Gigaword

appearing in this particular context should be greater for *applaud* than for *know*. In what follows, we illustrate the mechanism of these two constraints on the basis of the collected instances obtained from the BKB.

8.2.2.1 Using the Full Role Context

In order to demonstrate the usefulness of the first principle, i.e. conditioning on the full role context, instead of only a partial role subset, we compute a significance score over all predicate types in the BKB. The significance score is simply the ratio of a full (`A0` *and* `AM-NEG`) versus a partial condition on roles (only `A0`) when computing the implicit role probability of an `AM-CAU`, i.e. $\frac{\mathrm{P}(\texttt{AM-CAU}|\mathit{pred},\texttt{A0},\texttt{AM-NEG})}{\mathrm{P}(\texttt{AM-CAU}|\mathit{pred},\texttt{A0})}$. Table 8.5 is a rank-sorted list according to the significance score of each predicate. Here, larger significance scores indicate a greater discrepancy between the two ways of computing the implicit role probability, and therefore indicate a greater contribution of the full role context including the additional `AM-NEG` role. Intuitively, many predicates that co-occur only with standard agent roles are not accompanied by a causality role (low probabilities of second column in Table 8.5), however—once their context is negated—there is a sudden increase in probability towards an implicit causality (cf. second column and the following discussion Section 8.3). Interestingly, verbal predicates such as *enter*, *attend*, *participate*, *apply*, in their negated contexts strongly elicit an implicit causality role. For the predicates at the bottom of Table 8.5, whose significance scores are below 1.0, adding additional information in terms of the negation in fact *decreases* the probability of an implicit `AM-CAU`. Note that this is, for instance, also the case for the predicate *know*. Finally, it is worth mentioning that there is a clear influence of sentiment polarity on the predicates with high probabilities in the partial context (see, for instance, the negative words *suffer*, *worry*, *resign*). Negating these words in turn results in a positive context (which is considered the norm) and commonly do *not* evoke the response in asking for a reason.

8.2.2.2 Comparing Different Predicates

Returning to the second condition of measuring the effect of implicit information for *equivalent* role constellations among *different* predicate types, consider Tables 8.6 and 8.7. Obviously, the probability for an implicit `AM-CAU` occuring in this local context is much greater for *applaud* ($\sim$ 22.6%) given the explicitly realized role constellation (`A0, AM-NEG`) than for *know* ($\sim$ 1.5%) or *care* ($\sim$ 3.1%), respectively. This result is a purely quantitative estimate and it is analogous to the motivation from the example in the introduction, which gives us further evidence for an interdependence between local implicit semantic role information and the non-local discourse context, whose quantitative assessment we will further approach in the next section.

predicate	**partial** role context P(AM-CAU | *pred*, A0)	**full** role context P(AM-CAU | *pred*, A0, AM-NEG)	Score
enter.01.V	0.0038	**0.2476**	64.92
compete.01.V	0.0022	0.1336	60.61
investigate.01.V	0.0068	**0.3**	43.63
practice.01.V	0.0094	0.1854	19.69
attend.01.V	0.0101	0.1332	13.09
speak.01.V	0.0080	0.0918	11.37
participate.01.V	0.0060	0.0611	10.04
apply.01.V	0.0901	**0.7346**	8.15
give.01.V	0.0184	0.1274	6.89
talk.01.V	0.0032	0.0222	6.85
testify.01.V	0.0138	0.0774	5.61
hope.01.V	0.0073	0.0408	5.57
flee.01.V	0.0258	0.1392	5.37
react.01.V	0.0176	0.0949	5.37
start.01.V	0.0183	0.0973	5.31
sign.01.V	0.0931	**0.4878**	5.23
vote.01.V	0.0107	0.0529	4.90
...	...	...	...
...	...	...	...
worry.01.V	0.0522	0.0570	1.09
cry.02.V	0.0205	0.0222	1.08
have.03.V	0.0066	0.0071	1.07
object.01.V	0.0921	0.0890	0.96
do.02.V	0.0073	0.0063	0.85
hit.01.V	**0.1666**	0.1377	0.82
know.01.V	0.0199	0.0152	0.76
suffer.01.V	**0.1735**	0.1145	0.66
worry.02.V	**0.1972**	0.1276	0.64
act.02.V	0.0525	0.0293	0.55
quit.01.V	0.0898	0.0461	0.51
resign.01.V	0.0929	0.0284	0.30

Table 8.5: Predicates and their role probabilities for an implicit causality role AM-CAU given *two* distinct contexts—a partial (second column) and a full context with a negation (third column). Predicates are sorted by significance scores, i.e. by the strengths of discrepancy of the two probabilities, measuring the effect of the full context. Frequencies for role patterns are obtained from the BKB; threshold for role pattern frequencies $>$ 10 occurrences. Predicate suffixes indicate word-sense and part-of-speech information.

applaud.V.01 # predicate instances: 17,794 # distinct patterns: 964	
A0 A1	5,717
A0	1,263
A0 A1 AM-ADV	974
A0 A1 AM-TMP	943
A0 AM-TMP	783
...	
A0 AM-NEG	24
A0 AM-CAU AM-NEG	7
P(AM-CAU|*applaud*, A0, AM-NEG)	≈**22.6%**

Table 8.6: Role pattern statistics for the verbal predicate *applaud*.

know.V.01 # predicate instances: 1,369,197 # distinct patterns: 2,795		***care.V.01*** # predicate instances: 81,013 # distinct patterns: 1,010	
A0 A1	408,101	A0 A1	16,959
A1 A2	126,748	A0 A1 AM-NEG	9,544
A0 A1 AM-NEG	115,675	A1	6,456
A1	69,316	A0 A1 R-A0	3,248
A0 A1 AM-TMP	56,253	A0 AM-ADV AM-NEG	3,020
...		...	
A0 AM-NEG	13,392	A0 AM-NEG	2,866
A0 AM-CAU AM-NEG	208	A0 AM-CAU AM-NEG	91
P(AM-CAU|*know*, A0, AM-NEG)	≈**1.5%**	P(AM-CAU|*care*, A0, AM-NEG)	≈**3.1%**

Table 8.7: Role pattern statistics for the verbal predicates *know* and *care*.

8.3 Assessing the Effect of Local Implicit Roles to the Discourse

The intention of the leading example in this chapter was to demonstrate the effect of a single implicit role on a single implicit discourse relation. Although the example seems intuitive, its plausibility on a general level remains to be tested. Also, it needs to be tested whether the effect generalizes to other predicates and other discourse senses, as well. To this end, we propose to evaluate our method on the hand-annotated implicit discourse relations of the Penn Discourse Treebank (Prasad et al., 2008) by measuring the strength of implicit role contribution of a larger collection of predicate instances towards the specific discourse relations. In the following, we elaborate on the experimental setup (Section 8.3.1), describe how sense-wise implicit information is computed (Section 8.3.2), and evaluate our findings focusing on the important relations from the literature of causality and time, but also give new insights into less frequently encountered modifier roles of type purpose and predication. We wrap up the section with a final discussion.

8.3.1 Experimental Setup & Preprocessing

The (training) data set of the PDTB consists of approximately 17,000 argument pairs with manually labeled implicit discourse relations for 12 distinct senses; cf. Figure 8.1. For this experiment, gold argument spans and their respective tokens are taken as input to the same processing pipeline which has been applied to construct the background knowledge base of predicate co-occurrences. To be precise, each argument is considered a sentence[9] and is equipped with SRL annotations using *mate-tools*[10] (Björkelund et al., 2009).

8.3.2 Sense-Wise Computation of Implicit Information

We infer the contribution of non-local implicit semantic role information towards a particular implicit discourse sense as follows: for all N predicate instances appearing in either of the two arguments of an implicit discourse relation of type DISCREL, we compute an average probability by aggregation over all conditional probabilities for an implicit role `A-IMPL`, given a predicate instance $pred_i$ and its (automatically annotated) explicit role realization (`explicit roles` of $pred_i$).

$$\mathrm{P}(\texttt{A-IMPL}_{DISCREL}) = \frac{1}{N}\sum_{i=0}^{N} \mathrm{P}(\texttt{A-IMPL} \mid pred_i, \texttt{explicit roles}_{pred_i}) \tag{8.1}$$

Note that for the computation only those predicate instances are considered for which the BKB has information. This applies to both the predicate with the explicit roles (as automatically inferred from the SRL tool), as well as the predicate with the explicit roles *plus* the additional implicit role for which the implicit role probability needs to be estimated (by means of the previous equation). Predicate instances which are—for reasons of data sparsity—not fully covered by the BKB in this way are not considered. When a predicate has no explicit roles, the BKB must contain the pattern of the predicate and the single implicit role. When the implicit role is already explicitly realized in a discourse argument, the predicate is discarded from the analysis.

Note that for the purpose of this correlation study, Equation 8.1 can be specifically tailored to corresponding role-discourse relation pairs, for instance for `AM-CAU` and the (ideally) related **causal** discourse relations, CONTINGENCY:Cause:reason and CONTINGENCY:Cause:result as follows:

[9] Note that PDTB arguments can be a sequence of tokens, phrases, complete sentences (mostly missing sentence-final punctuation), or even multiple sentences. In order to keep the setup simple, we treat each discourse argument as a standalone unit and feed it one instance at a time into the SRL pipeline. Sentence-final periods are appended to argument spans which do not contain one.

[10] `https://code.google.com/archive/p/mate-tools/`

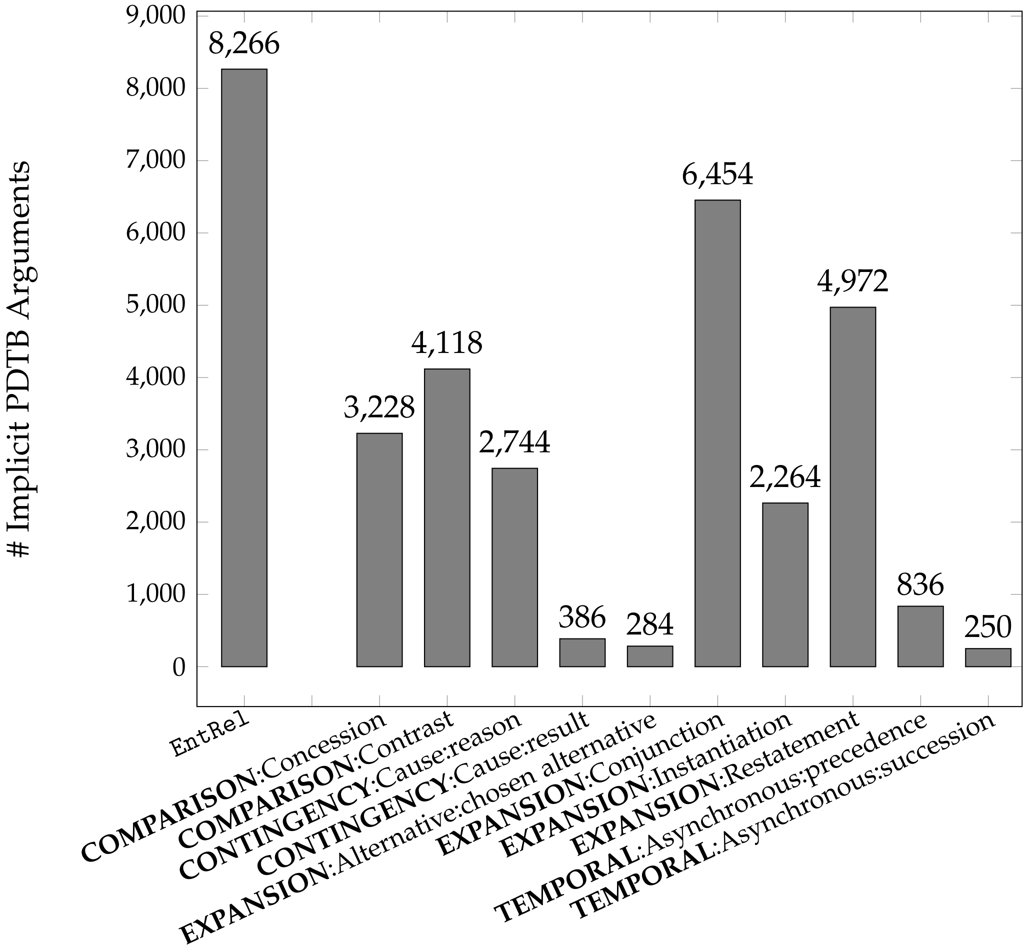


Figure 8.1: Distribution of implicit PDTB discourse sense relations involved in the correlation study.

$$\mathrm{P}(\texttt{AM-CAU}_{\text{CONTINGENCY:}\textsc{Cause}\text{:}*}) = \frac{1}{N}\sum_{i=0}^{N}\mathrm{P}(\texttt{AM-CAU} \mid \mathit{pred}_i, \texttt{explicit roles}_{\mathit{pred}_i})$$

The same applies to the temporal discourse relations, whose implicit semantic contribution should stem from temporal roles:

$$\mathrm{P}(\texttt{AM-TMP}_{\text{TEMPORAL:}*}) = \frac{1}{N}\sum_{i=0}^{N}\mathrm{P}(\texttt{AM-TMP} \mid \mathit{pred}_i, \texttt{explicit roles}_{\mathit{pred}_i})$$

Ideally we would assume that implicit causal semantic roles contribute most significantly towards implicit causal discourse relations, while implicit temporal discourse senses can be explained through the effect of local implicit temporal roles. We evaluate these two types and also elaborate on further constellations in the following section.

8.3.3 Evaluation

8.3.3.1 Implicit Relations of Causality & Time

The horizontal bars of Figure 8.2 indicate for every implicit discourse sense the average probability of implicit semantic role information of a certain role type (the top chart for `AM-CAU`, below for `AM-TMP`). The probabilities are averages obtained from all predicates in the respective argument spans for which the BKB holds information. Overall, there is a trend visible, in that implicit `AM-CAU` roles contribute stronger towards implicit causality relations (CONTINGENCY:Cause:reason and CONTINGENCY:Cause:result are within the three highest ranked relations with >0.01% contribution). The effect is even stronger for the temporal roles (>0.12%), with TEMPORAL:Synchrony and TEMPORAL:Asynchronous:precedence being those relations which on average appear to gather most of the implicit role probabilities from the temporal adjunct role `AM-TMP`.

Interestingly, implicit causality information has its strongest effect within relations of type EXPANSION:Alternative:chosen alternative (cf. the accompanying discussion section 8.3.4). Also TEMPORAL:Asynchronous:succession seems to behave differently in this context with respect to the two other temporal senses. Notice also that `EntRel`, entity relations, are ranked among the senses with lowest implicit probabilities in both charts.

Table 8.8 shows for each of the three top-ranked discourse relations those predicates and explicit role patterns which contribute most strongly towards a certain discourse relation. Note that again for the implicit causal relations, explicit predicate patterns contain predominantly negations and adverbial roles—more often than with the other senses. An illustration is given by the verbal predicate *communicate* and its explicit role pattern `A2 AM-MOD AM-NEG`. It is part of the second argument of an implicit relation of sense CONTINGENCY:Cause:reason.

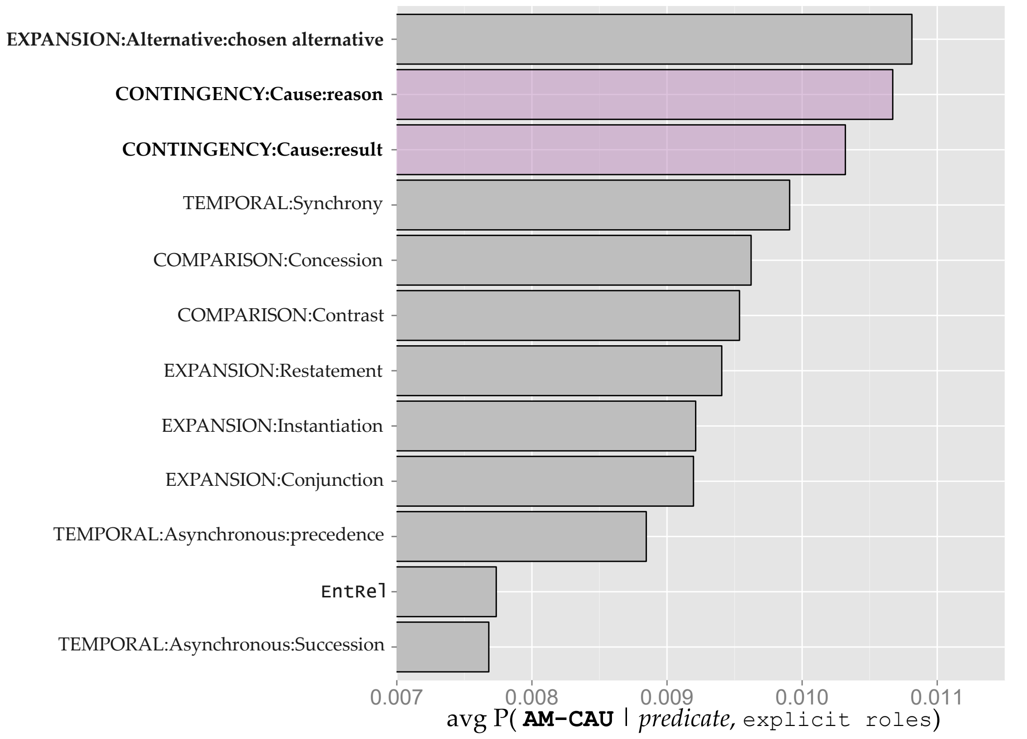
(21) `Arg1`: [...] Japanese offices tend to use computers less efficiently than American offices do

`Arg2`: [...]In Japan, many desktop terminals are limited to one function and [ca$_{\texttt{AM-MOD}}$][n't$_{\texttt{AM-NEG}}$] [communicate$_{pred}$] [with other machines$_{\texttt{A2}}$]

Implicit discourse sense: CONTINGENCY:Cause:reason[11]
Inferred connective: *as*.
P(`AM-CAU` | *communicate.01.V*, `A2`, `AM-MOD`, `AM-NEG`) $\approx$ 0.5.

Here, the fact of not being able to communicate is an elaboration on the reason for the statement in the first argument. In approximately 50% of all cases in the

[11] Document ID `wsj_0445` of PDTB training section. Note that in the second argument, an `A0` role is not properly detected by the automated SRL pipeline. Also note that there is a second predicate *limit.V.01* which also contributes towards the average implicit information of that sentence.



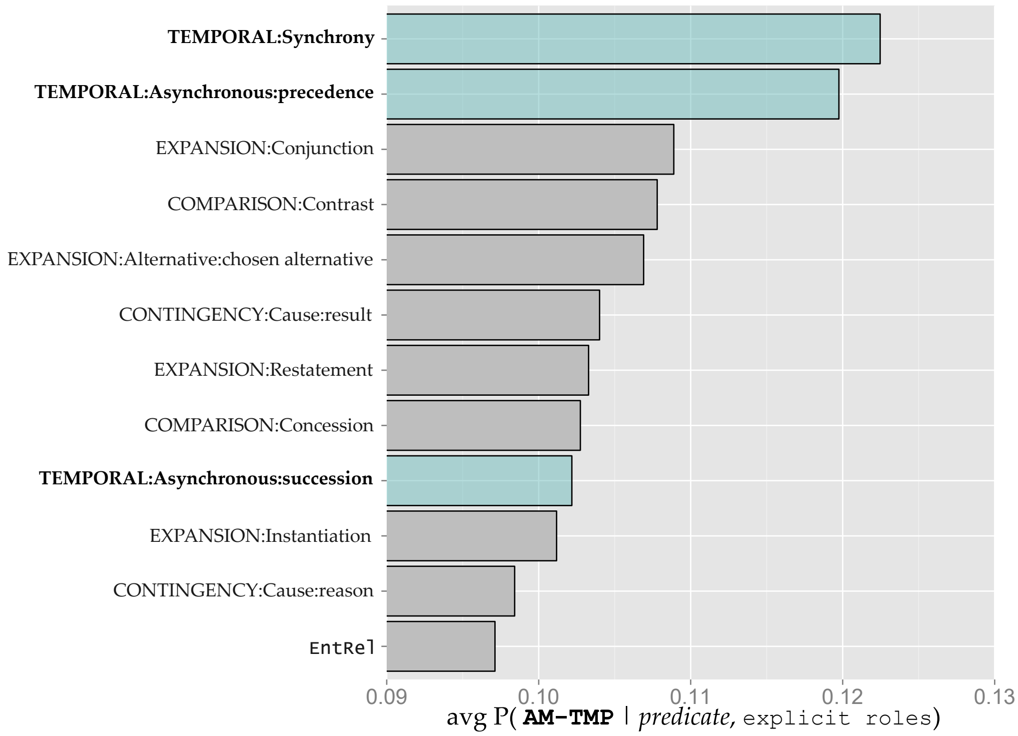


Figure 8.2: Contribution of implicit causal (AM-CAU/top) and temporal (AM-TMP/bottom) information on all implicit sense relations in the PDTB.

BKB (and thus in Gigawod), *communicate* together with this explicit role pattern is accompanied by an AM-CAU (realized by *because* or *as*, etc).

Another example to illustrate the effect of predicate-specific role patterns on the discourse structure—this time on a temporal relation—is given by the following argument pair.

(22) Arg1: the commission voted, as expected, to formally object to the accord between Air France, the state-owned airline, and state-controlled domestic carrier Air Inter

Arg2: [the two companies$_{\texttt{A1}}$] [will$_{\texttt{AM-MOD}}$] be [notified$_{pred}$] [so they can begin negotiations with Brussels on how to modify the pact$_{\texttt{AM-PNC}}$]

Implicit discourse sense: TEMPORAL:Asynchronous:precedence[12]
Inferred connective: *next*.
P(AM-TMP | *notify.01.V*, A1, AM-MOD, AM-PNC) $\approx$ 0.8.

In this relation pair, the first argument temporally precedes the second argument. The predicate *notify.V.01* is part of an event description in the future (relating to the negotiations which will happen). This temporal ordering information is partly expressed by the overt role pattern, including AM-MOD (modal *will*) and the purpose role AM-PNC. Also note that in this specific pattern the agent role (A0) of *notify* is *not* overtly expressed but its filler can be recovered in the previous sentence (cf. *the commission*). Thus, it is this characteristic constellation of predicate-role co-occurrences which is likely to trigger the non-local temporal discourse sense; or put simply, by notifying someone first a decision has to be made (*voting*) and this is the direct connection to the immediately preceding discourse unit.

8.3.3.2 Implicit Roles of Type Purpose & Predication

The upper chart in Figure 8.3 shows the effect of an implicit adjunct role of type purpose AM-PNC. Note that for purpose roles, the contribution towards (most notably) causality relations is again greater compared to the majority of other discourse senses. Examples for explicit realizations of purpose roles deliver additional information in terms of explanations and are realized, for instance, by *in exchange for such goodies, to ease the country's financial crisis, so that his avaricious brother can succeed, small coins given as change*.

Finally, the lower part of Figure 8.3 illustrates the impact of the adjunct predication role AM-PRD on various discourse senses. Examples of overt predications include, e.g., *joined in full strength, the staff standing ready, Toshiba Corp. busted open that sector, all hell broke loose with the finances, ate the meat raw*[13] etc. These examples are rather infrequent, as they represent stylistic markers in the corpus

[12] Document ID wsj_0743.

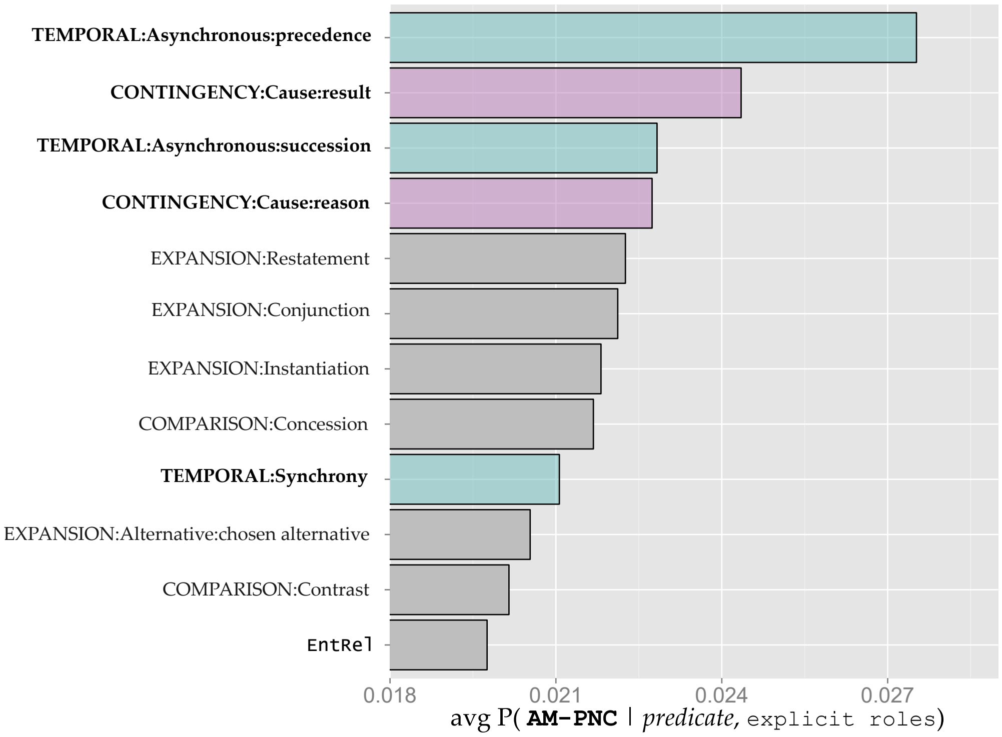
[13] This example is taken from Jurafsky and Martin (2017).

DISCREL	Pattern	P(AM-CAU | Pattern)
EXPANSION: Alternative: chosen alt	*close.01.V* [A1, AM-TMP]	0.191
	sign.01.V [A0, A1, AM-ADV, AM-MOD, AM-NEG]	0.103
	suspend.01.V [A1, A2, AM-TMP]	0.087
CONTINGENCY: Cause: reason	***communicate.01.V*** [A2, AM-MOD, AM-NEG]	0.500
	damp.01.V [A0, A1, AM-ADV, AM-DIS]	0.500
	leap.02.V [A0, AM-ADV, AM-DIR]	0.384
CONTINGENCY: Cause: result	*bury.01.V* [A1, A2, AM-MOD, AM-NEG, AM-TMP]	0.334
	die.01.V [A1, AM-ADV]	0.327
	restore.01.V [A1, AM-ADV, AM-MOD, AM-NEG]	0.312
		P(AM-TMP | Pattern)
TEMPORAL: Synchrony	*tumble.01.V* [A1, A2, AM-MOD]	0.666
	announce.01.V [A0, A1, AM-ADV]	0.600
	acquire.01.V [A0, A1, R-A1]	0.593
TEMPORAL: Asynchronous: precedence	***notify.01.V*** [A1, AM-MOD, AM-PNC]	0.800
	arrest.01.V [A1, R-A1]	0.735
	come.01.V [A1]	0.729
EXPANSION: Conjunction	*announce.01.V* [A1, R-A1]	0.935
	begin.01.V [no roles]	0.901
	sensation.01.N [no roles]	0.859

Table 8.8: Most influential predicate instances and their explicit role patterns in terms of implicit contribution towards discourse senses. For predicates in boldface explanations are given in the text.

data, however their implicit effect is predominant for the discourse senses Expansion:Instantiation, and—for the first time—also observable for entity-based coherence relations (EntRel). One possible explanation for the joint occurrence of both senses might be that instantiations of second arguments provide *descriptions in further detail*[14]. Entity relations behave very similar in this respect. That these two discourse senses are highly correlated has also been recently demonstrated by McKinlay (2013) who showed that *entity instantiations* (a form of entity relations with subset member mentions) appeared significantly more often within arguments of EXPANSION:Instantiation relations than other discourse relations. To summarize, both relation types are driven by implicit predication roles, i.e. the presence of a specific predicate combined with the absence of a (secondary) predication role increases the probability towards implicit instantiation and entity relations. We noticed that it is difficult, however, (in fact, almost impossible) to pinpoint this effect to single, individual examples in the PDTB. Nonetheless,

[14] For more information, the interested reader is referred to the specification in the PDTB annotation manual: https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf



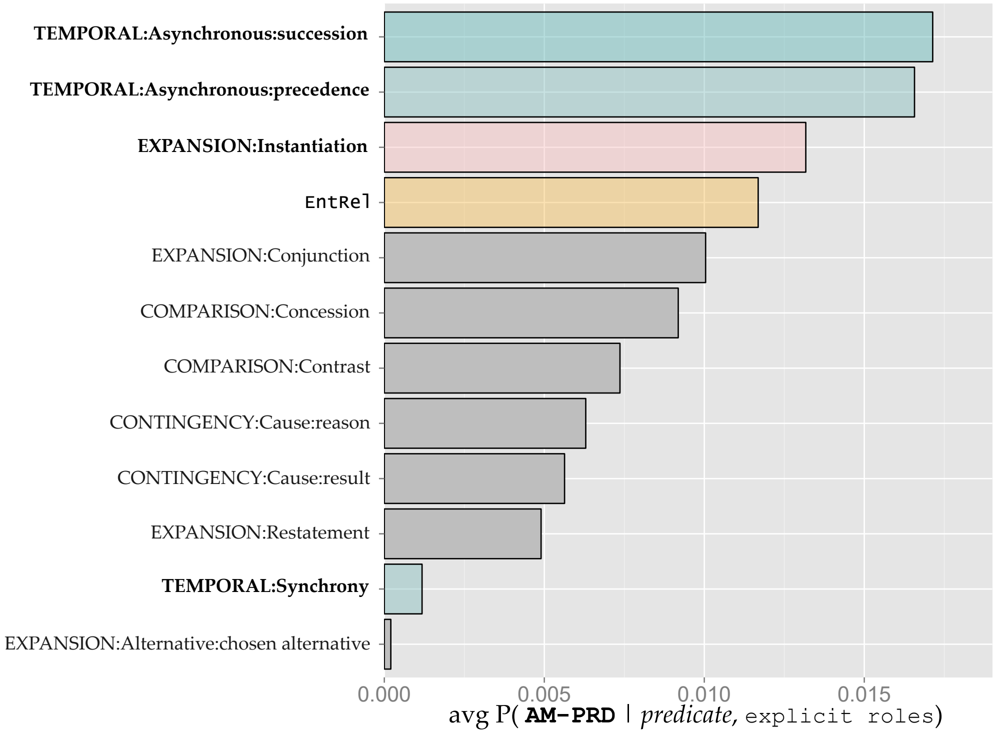


Figure 8.3: Contribution of implicit purpose (AM-PNC/top) and predication (AM-PRD/bottom) information on all implicit sense relations in the PDTB.

we believe that the influence of implicit predication information towards entity relations and instantiations can be best explained by an expansion of the example phrase *ate/eats the meat raw* from Jurafsky and Martin (2017). Instead of treating the secondary predication as a singleton pattern, one could extend it to form *two* adjacent discourse arguments whose implicit discourse senses share properties of either EXPANSION:Instantiation, `EntRel` (or even both), as in the following example:

(23) `Arg1`: John eats the meat.
`Arg2`: He eats it raw.

Implicit discourse sense: EXPANSION:Instantiation/`EntRel`
Inferred connective: *specifically*.

Note that this way the additional predication (*raw*) of the (originally) single-sentence predicate is moved into the second discourse argument while the general meaning is preserved. We believe that such constructions, once expanded over two argument representations, as in the PDTB, are typical for relations of instantiations or entity-based coherence when the predicate is likely to carry implicit predication information.

8.3.4 Discussion

In this correlation study and the accompanying experiments, we have observed that within-sentence predicative role constellations of overt and non-overt arguments evoke a quantitatively measurable amount of latent information which tends to positively correlate with extra-sentential discourse relations of similar implicit properties. These constellations—measured in terms of implicit roles—are predictive features which strengthen or weaken the expectation towards a certain discourse relation and which can be regarded as a large-scale generalization of the negation cue by Asr and Demberg (2015). One important difference, however, to the work in Asr and Demberg (2015) is that our approach is verifiable also on a greater number of implicit discourse senses. One reason for this is that we make use a larger quantity of patterns obtained from generalizations over mass data, instead of a single theoretically-motivated feature. Another reason might be that we deliberately consider *both* arguments for the extraction of cues instead of just the first one. Consider the following example from Asr and Demberg (2015), for which the authors claim that the first argument evokes an implicit causality/reason relation towards its continuation.

(24) John admires Mary. Shes plays the piano very well.

Even though this is clearly the case, according to our intuitions, however, it seems necessary to consider also the contribution of the *second* argument as, in general, the global sense of an argument pair arises from the *joint* occurrence of both. As an illustration of this idea, consider the following made-up example of type COMPARISON:Contrast with accentuation on the personal pronoun.

(25) John admires Mary. *She* still dislikes him.

Interestingly, the senses of examples (21) and (22) from Section 8.3.3 are also primarily featured by their second, instead of first arguments. Note further that this particular view on the mutual combination of an argument pair has been the major inspiration in the traditional, application-oriented approaches to discourse sense classification, in which *word-pair features* (i.e. the cross product of the combination of all tokens) of both arguments have been the core source of information (Blair-Goldensohn et al., 2007; Lin et al., 2009; Biran and McKeown, 2013). The motivation for these types of features, however, is purely machine learning-based. Unfortunately, it turned out that they are dramatically sparse and among the best predictors are linguistically rather uninterpretable function words with no real semantic content, as found by Pitler et al. (2009).[15] Our approach including semantic roles can thus be seen as a promising generic middle ground between these low-level features and the mainly corpus-oriented studies of related work.

On a related note, Webber (2013) presented such a corpus study to find evidence and more predictive, linguistically-motivated features for a certain type of coherence relation, namely EXPANSION:Alternative:chosen alternative. In its explicit form, the sense relation is signaled by *instead*, implying that one of the two alternatives in either of the two arguments is semantically taken. She illustrates that the most relevant features identified are negation and polarity words, downward-entailing constructions of the sort described in Danescu-Niculescu-Mizil et al. (2009), and event modals. By demonstration, these features are even more predominant for implicit relations than for the explicit counterparts and in line with our findings as, for instance, the most predictive patterns (with the greatest probability of am implicit causal role) contain either a negation or a modal verb role (or both) in over 65% of all cases. It should again be emphasized that we find a strong correlation between causality and the "chosen alternative" sense (top chart of Figure 8.2) for which very similar predictors are at work.

Interestingly, a striking opposite, i.e. an almost non-existent effect on causal and "chosen alternative" relations is evident from the contribution of implicit *predicative* roles (bottom chart of Figure 8.3). We believe that a possible explanation for this trend is that relations of type *instead* are semantically contrastive relations; cf. Stede (2011). Predication, however, signifies further attribution and is more likely to be associated with instantiation relations or entity-based coherence by further elaboration on the first argument.

Finally, regarding entity-based coherence relations, it should be noted that they are peculiar and behave very differently from the majority of discourse senses (e.g., they represent outliers in three of the four distributions in Figures 8.2 and 8.3). A related discussion has been raised by Knott et al. (2001) who focused on the semantically equivalent ELABORATION—by far the most frequent discourse relation—, claiming that under the RST framework there are general modeling issues regarding this specific type of relation. Among other things, the authors

[15] Cf. Section 2.2.1 and the ensuing discussion in Section 2.2.3.1 for further details.

point out structural problems with these relations, e.g., by the continuous constituency principle which is most often violated, because entity-based coherence typically expands across long-distance dependencies instead of between immediately adjacent arguments. This in fact implies that discourse relations of type `EntRel` are underrepresented in the PDTB, because here the gold annotation does not allow for a combination of far-away discourse units. Moreover, it should be noted that, in the PDTB, approximately 20% of all entity relations cannot be explained by coreferential properties, i.e. they do not share at least one coreferenced entity within the respective discourse arguments.[16] We thus argue that—on the basis of the experiments outlined in this chapter—detecting and mining entity-based coherence relations requires efforts which are grounded in a deeper semantic level, in particular, their proper analysis can benefit from the recovery of implicit semantic roles and the associated effect on the extra-sentential discourse structure.

Open Issues & Further Directions: A final remark should be made about the general setup of the correlation study conducted here and how it could be extended. Although we found the desired correlation between implicit role and implicit relation pairs of the same type, for instance for causality and time, it still remains an open question how other combinations should be interpreted and to what extent an effect can be interpretable at all. For instance, it remains unclear how implicit manner information should be related to asynchronous temporal relations, or how implicit location roles are related to, for instance, expansion senses. A thorough inspection of individual argument pairs from the PDTB may help, but the effects under investigation are rather subtle nuances which in the majority of cases are not trivially detectable from single examples. Here, only fine tendencies are evident to a certain extent (even when generalized over mass data), which this correlation study fortunately allowed us to quantify, yet whose interpretation is still far from being clear-cut. Another objection related to this can be seen in data sparsity of the knowledge base. Even though almost two billion words have been processed for semantic roles, rare predicates are still an issue for the evaluation on the PDTB senses and the contribution of these predicates is definitely questionable.[17] Finally, it remains an interesting direction to pursue further whether other types of arithmetic compositions of implicit information in argument spans—instead of plain average—would yield stronger correlations and better effects for interpretation. Even though it has been suggested that average computations seem to capture generalizations over (syntactic and) semantic effects pretty well, in particular with respect to word embeddings, cf. Socher and Manning and in particular Manning (2015b,c), additional experiments might build on finding alternatives to these standard approaches.

[16] This information can be obtained by mapping gold discourse and gold coreference annotations from PDTB and OntoNotes (Hovy et al., 2006), respectively, as these resources annotate subsets of the same texts.

[17] For instance, *charge.07.V* with roles `[A0, A1, AM-ADV, AM-TMP]` is observed only two times in Gigaword.

8.4 Summary

This chapter has introduced an experiment in the form of a correlation study—the first of its kind—to assess the impact of sentence-internal cues on their immediate implicit discourse context (Section 8.1.1). The study is unique and novel in the sense that the specific type of information in the local context has been measured in terms of *implicit semantic roles*, i.e. by approximation of a latent factor which is assumed to be the driving force for non-local implicit discourses.

In order to quantify the effect of implicit semantic roles, the setup of the correlation study has been two-fold (cf. Section 8.2): First, a background knowledge base of predicate-role constellations has been acquired in an unsupervised manner. Specifically, frequencies for automatically annotated predicate-role constellations, both nominal and verbal, have been obtained from large text corpora, here from the newswire domain. Second, conditional probabilities have been pre-computed for all constellations in the knowledge base in order to assess the relative importance of a specific (implicit) semantic role, given a specific predicate context of overtly expressed roles.

For the purpose of an evaluation, the so-obtained role probabilities have been aggregated over both argument spans of an implicit discourse relation within the Penn Discourse Treebank (cf. Section 8.3). The evaluation has shown that the hand-annotated implicit discourse relations of the Penn Discourse Treebank correlate with implicit roles of differently strong association effects. Interestingly, the strongest associations were yielded between implicit causality relations and implicit causality roles. The same pattern holds true for temporal relations and roles, respectively, which strongly suggests the presence of a *synergy effect* in terms of a hidden interaction between implicit relations and implicit roles. Other effects have been evident as well, for example for purpose or predication roles on causality and entity relations/instantiations, respectively, but their interpretation has been less straightforward.

While prior corpus-based studies have dealt with only fine-grained, individual hand-collected examples in order to illustrate an effect of local to non-local structures, application-oriented approaches have instead made use of uninterpretable low-level features only (specifically in a machine learning context). A big advantage of using (implicit) semantic roles for the assessment of implicit discourse phenomena is that they represent a very generic way—both interpretable and measurable by effect—to model local cues and can be considered to reside somewhere in between the two extremes. Finally, the outlined correlation study and the corresponding methodology is easily adaptable to other domains or even languages. The only requirement for the setup of a corresponding BKB is an adequate SRL system in the requested language.

In the style of an additional bridge experiment, the next chapter describes an novel framework for the harmonic treatment of both local semantic roles and global discourse relations.

Chapter 9

Modeling Discourse Coherence on the Basis of Semantic Roles

9.1 Motivation

The correlation study in the previous chapter has shown that there exists a bottom-up effect of sentence-internal information (semantic roles) on the extra-sentential context (discourse structure), and that this effect can be quantitatively measured. This chapter aims at demonstrating that the way of assessing this specific interaction can also be approached top-down from the *reverse* direction. In order to bridge the gap between discourse structure on the one hand, and semantic roles on the other, we propose a joint framework for a task which we term *cross-argument semantic role labeling*. Practically, this task can be regarded as a special case of SRL which goes *beyond* the sentence level. In this setting, we intend to model local predicate role realizations which are structurally linked to previous utterances and which are likely to capture coreferential information as engendered by the global discourse context.

In the ensuing experiment, our focus is on a particular type of discourse relation, namely entity-based coherence. Various theories and accounts to explain entity relations have already been subject to a whole range of prior research studies, cf. Knott et al. (2001); Grosz et al. (1995); Rohde et al. (2007); Kehler and Rohde (2017), inter alia. In the previous chapter we were able to confirm that these relations are special insofar as they behave in large measure differently from the remainder of discourse senses in an analysis based on the Penn Discourse Treebank. We argue that, in analogy to previous (mostly theoretical) works, entity-based coherence relations are particularly worth investigating because they exhibit structurally unique properties related to anaphoricity and coreference, for which we propose a computational approach in order to map their global properties onto the local semantic role level. It should be noted, however, that our experiment involves non-coreferential relations as well in order to test whether our proposed model can in fact learn structural differences between the two.

Our applied methodology described hereafter is inspired by the famous work on *centering* (Grosz and Sidner, 1986; Grosz et al., 1995), whose foundations and main principles were briefly sketched in Section 2.1.1. This particular theory is, among other things, concerned with the anaphoric relationship between entities across sentence boundaries and the choice and realization of referring expressions in a coherent discourse. These key aspects are crucial for a holistic treatment of both inter-sentential and intra-sentential information. We give an account of this interconnection by means of a novel label scheme for transition relations and demonstrate that it offers a greater degree of flexibility in a computational approach to entity relations as opposed to a single backward-looking center. The work presented hereafter is not so much concerned with implicit information, which has been the main focus of all previous experiments. Instead, we demonstrate that our presented framework (it is structurally *identical* to the one outlined in Chapter 4 on implicit discourse parsing), which has originally fulfilled the purpose of modeling extra-sentential discourse structure, is in fact highly adequate and generally flexible enough to be applied to the local syntactic context of arbitrary predicates.

9.1.1 Entity Relations & Centering

As outlined in the introductory chapter, Centering Theory explains perceived discourse coherence on the basis of four distinct transition types, CONTINUE, RETAIN, SMOOTH-SHIFT, and ROUGH-SHIFT, cf. Grosz et al. (1995); Krifka (2006). These relations are not discourse relations in the strict sense; yet, they can be regarded as a coreferentially motivated class of relations and are thus particularly suited to explain various phenomena arising from entity-based coherence and elaboration relations. Four simplified examples are shown hereafter. We assume that in all cases *John* is in the backward-looking center C_b of the first argument.[1]

(26) `Arg1`: John likes Mike.
`Arg2`: He is always willing to lend a hand. (he=John)
C_b = *John*, C_f = {*John*}
Transition relation: CONTINUE

(27) `Arg1`: John likes Mike.
`Arg2`: However, Mike hates John.
C_b = *John*, C_f = {*Mike*}
Transition relation: RETAIN

(28) `Arg1`: John likes Mike.
`Arg2`: Mike is a great man.
C_b = *Mike*, C_f = {*Mike*}
Transition relation: SMOOTH-SHIFT

[1] We use PDTB notation for discourse units. For illustration purposes, we omit the discourse-initial sentences, and show only the preferred element in the forward-looking centers C_f.

(29) `Arg1`: John likes Mike.
`Arg2`: Peter has recently met Mike on one of these NLP conferences.
C_b = *Mike*, C_f = $\{$*Peter*$\}$
Transition relation: ROUGH-SHIFT

Examples (26), (27), (28), and (29) obviously represent four distinct discourses and result in differences in perceived discourse coherence. In the second argument of (26), *John* is in the backward looking center and still the most salient entity in the forward-looking center of the new utterance (CONTINUE), whereas in (27), *Mike* (in place of *John*) is now the highly ranked element in the forward-looking center of the second discourse argument, resulting in a RETAIN transition. In the SMOOTH-SHIFT of (28), the backward-looking center *changes* (to become *Mike*), and it is equal to the forward-looking center of the new utterance. However, finally, in (29), the backward-looking center is still *Mike* as in the previous example, but note that this time the forward-looking center changes as well (to become *Peter* / ROUGH-SHIFT).

It should be noted that, Centering Theory establishes a link between such pairs of utterances solely based on a single referent per utterance, namely the backward-looking center. In this chapter, we suggest an extension of the salience ranking between two discourse segments and encode their connection into semantic roles—allowing for a diversification in assessing the coherence among distinct realizations of entities, and any two discourse segments in general. In our proposed method, we refer to an argument pair and consider a unique relation between two predicates—one from each utterance. These predicates typically differ in their role realizations, i.e. distinct argument types and number of explicit roles. A role pattern is said to be *coreferential* if at least one role filler in the second utterance is coreferent with an entity realized by any other semantic role of the first predicate. To this end, we introduce a specific *cross-argument label* for any pair of utterances. This label is predicate-dependent. It is inspired by Centering Theory and serves the purpose of the backward-looking center, thus establishing a direct backward link to the predicate of the previous discourse argument. When coreferentiality exists, this implies that any of the four transition types hold, and the cross-argument label should express this information. Otherwise, we refer to the relation as non-coreferential, i.e. `NoRel` holds (or possibly a ROUGH-SHIFT when another entity is introduced).[2]

In the following section, we describe the adaptation of an existing discourse architecture to the task of cross-argument SRL (Section 9.2). In particular, we elaborate on data acquisition for cross-argument SRL patterns and the connection between classical transition types from centering and our labels (Section 9.2.1). We define the experimental setup (Section 9.3) and distinguish individual tests for the

[2] Note that, in this bridge experiment our focus is not on PDTB-style discourse parsing, i.e. we disregard any other PDTB sense relations.

classification of coreferential and non-coreferential labels, respectively. Finally, we report on evaluation results, and conclude in Section 9.4.

9.2 A Discourse Model for Cross-Argument Semantic Role Labeling

9.2.1 Data Acquisition

Data Preparation: In order to acquire a reasonable amount of cross-argument SRL patterns as training data for our experiments, which capture both coreferential and non-coreferential discourse properties, we focus on three main data sources: i.) PropBank (Palmer et al., 2005) for gold-annotated verbal predicate-argument structure, ii.) OntoNotes (Hovy et al., 2006) for gold-annotated coreference information, and iii.) the Penn Treebank (Marcus et al., 1994). Note that both PropBank and OntoNotes cover the same Wall Street Journal base texts of the Penn Treebank, which makes these resources especially suitable for our purposes.

For the joint combination (mapping) of semantic roles with coreference patterns integrated into the primary Penn Treebank tokenization layer, we employ *conll-merge*[3] (Chiarcos and Schenk, 2018), a toolkit for the harmonization of concurrent linguistic annotations based on the same underlying texts into one shared output format. Since not every document in PropBank is annotated in OntoNotes—and vice versa—we obtain a final set of 597 documents with approximately 410k merged tokens and 39,500 predicates.

Training Data Generation: We treat each sentence as a discourse argument in the shallow style of the Penn Discourse Treebank (Prasad et al., 2008) and derive training instances as follows: For the current sentence i (denoted as the second argument, i.e. `Arg2`), we generate an argument pair with its previous sentence $i-1$ (which serves as `Arg1`). We repeat the procedure for the pre-previous one $i-2$ and generate another pair between i and $i-3$. This way, we ensure to keep long distance dependencies up to three sentences prior to the second argument.

For each argument pair (`Arg1`–`Arg2`) we generate a *cross-argument label*. The label is a 5-tuple and it represents a *predicate's argument realization in the second discourse argument*. The exact form of the SRL pattern is generated as follows: Semantic roles are *consecutively* indexed from 0 to 4 for arguments `A0` to `A4` in `Arg1`, and from 5 to 9 for arguments `A0` to `A4` in `Arg2`. We first consider a single predicate instance in the *second* (discourse) argument together with its (SRL) argument realization. For example, if a predicate in the second argument has an overt agent (`A0`) and a patient (`A1`) role, its label would be `56XXX`, because the first two indices (5 and 6) are filled by overt roles and all remaining (unfilled) arguments are represented by dummy placeholders `X`. Then, for the predicate and its associated role pattern in `Arg2`, *we check if a semantic role is coreferential with a constituent in the*

[3] `https://github.com/acoli-repo/conll`

first argument which itself fills a semantic role in Arg1. If so, we replace the corresponding label index by the semantic role index in Arg1 to denote coreferentiality. If a semantic role in Arg2 (more precisely, a constituent filling a semantic role) is not coreferential with any phrase in Arg1 the label index is not modified.[4]

As an illustration, consider Example (30).[5] In the second (discourse) argument, the predicate *learn* along with its (SRL) arguments evokes the 5-tuple 56XXX, however, since the A0 of *learn* is coreferential with the A0 of *commit* in Arg1, the final label is 06XXX.

(30) Arg1: But Sony ultimately took a lesson from the American management books and fired Mr. Katzenstein, after [he$_{\mathtt{\mathbf{A0}}}$] [committed$_{pred}$] [the social crime of making an appointment to see the venerable Akio Morita, founder of Sony$_{\mathtt{\mathbf{A1}}}$].

Arg2: [Mr. Katzenstein$_{\mathtt{\mathbf{A0}}_{\text{coref}(\mathtt{\mathbf{Arg1}}-\mathtt{\mathbf{A0}})}}$] certainly would have [learned$_{pred}$] [something$_{\mathtt{\mathbf{A1}}}$], and it's even possible Mr. Morita would have too.

Label: 06XXX (Continue)

The example in (30) is based on the two specific predicate senses *commit.v.02* in Arg1 and *learn.v.01* on Arg2. Note that, for any given discourse argument pair, the procedure is repeated for *all* predicate combinations in Arg2 and Arg1, thus deriving multiple labels for a given Arg1–Arg2 training instance.

Note that arbitrary and more complex patterns are possible. For example, an A0 in the second argument could be coreferential with an A2 in the first argument (label: 2XXXX, no other roles present), or an A1 in Arg2 could be coreferential with A0 in Arg1 (label: 507XX, A0 and A2 independently overt in Arg2) as in the following example (31).[6]

(31) Arg1: [...] [Mr. Keating$_{\mathtt{\mathbf{A0}}}$] had [gathered$_{pred}$] [the money$_{\mathtt{\mathbf{A1}}}$] [for him$_{\mathtt{\mathbf{A3}}}$] about two weeks before the meeting with regulators.

Arg2: [...] shortly after [the government$_{\mathtt{\mathbf{A0}}}$] formally [accused$_{pred}$] [Mr. Keating$_{\mathtt{\mathbf{A1}}_{\text{coref}(\mathtt{\mathbf{Arg1}}-\mathtt{\mathbf{A0}})}}$] [of defrauding Lincoln$_{\mathtt{\mathbf{A2}}}$].

Label: 507XX (Retain)

Finally, consider the example (32)[7], in which the local SRL structure in Arg2 is not affected by coreferentiality with Arg1, which is the case for the majority of instances:

[4] Only exact matches (no substring matches) are regarded as coreferential.

[5] Document ID wsj_0037

[6] Document ID wsj_2446

[7] Document ID wsj_2388

(32) Arg1: [Honeywell's contract$_{\texttt{A1}}$] [totaled$_{pred}$] [\$69.7 million$_{\texttt{A2}}$], and IBM's \$68.8 million.

Arg2: [Grumman Corp.$_{\texttt{A0}}$] [received$_{pred}$] [an \$18.1 million Navy contract to upgrade aircraft electronics$_{\texttt{A1}}$].

Label: 56XXX (Rough-shift)

Mapping Cross-Argument Labels to Centering Transitions: Obviously, there exists a correspondence between our fine-grained labels and the original centering transition relations. This correspondence is non-trivial in some of the cases, but a few interesting observations can be made. For instance, Example (30) with cross-argument label 06XXX can be straightforwardly mapped to a Continue. Example (31) introduces another entity (*the government*) in the form of a retention (Retain), however keeps *Mr. Keating* still in focus. Finally, in the second discourse argument of Example (32) another entity is introduced and there is obviously no coreferential relation to the first argument, resulting in a Rough-shift. Note, that in the absence of an entity in Arg2 typically NoRel would hold.

Table 9.1 gives an overview of the label distribution for all extracted patterns (occurring at least 100 times) and roughly equivalent transitions from Centering Theory. Role subscripts in the third column of the table indicate coreferential roles. It becomes evident that the most frequent patterns (ranks 1–7) are non-coreferential. These relations account for almost 90% of all patterns. Also note that this distribution (obtained from gold annotations) is very similar to the number of automatically inferred patterns in one of our previous experiments (cf. Chapter 8, Table 8.3). The cross-argument label can be mapped to NoRel when either the agent role in Arg2 is non-coreferential within Arg1, or when no agent role is present in the second discourse argument. Note that when it is present, as for instance in Example (32) with label 56XXX, this can also indicate a Rough-shift. It is noteworthy that the distribution shown in Table 9.1 is highly skewed and the first coreferential label (06XXX/Continue) appears at rank 8. This is in line with the claims made by Centering Theory that sequences of continuation are to be preferred over other coherence transitions, cf. Grosz et al. (1995, Rule 2, p. 215). Yet, the overall number of continuations account for only 2.5% of all instances.

Finally, it should be noted that the mapping can also be performed the other way round, i.e. from centering transitions to SRL role labels. Interestingly, it turns out that the classical examples used to motivate centering are very rare phenomena when we consider the exact semantic role patterns. For example, some of the few retentions from Grosz et al. (1995) are not even part of the overall distribution in Table 9.1. We reproduce an example hereafter from Grosz et al. (1995, p. 217, ex. 20) and give it an appropriate cross-argument SRL label.

(33) Arg1: [He$_{\texttt{A0}}$] [called up$_{pred}$] [Mike$_{\texttt{A1}}$] [...] (He=John)

Arg2: [Mike$_{\mathtt{A0}_{\text{coref}(\mathtt{Arg1}-\mathtt{A1})}}$] has [annoyed$_{\textit{pred}}$] [him$_{\mathtt{A1}_{\text{coref}(\mathtt{Arg1}-\mathtt{A0})}}$] [...]

Label: 10XXX (Retain)

In Example (33), the SRL role pattern contains nested double-coreferentiality (10XXX/Retain), however, in our derived data set, there are only two of these instances out of approximately 260k relations. One of these argument pairs is shown hereafter in Example (34)[8].

(34) Arg1: Mr. Waggoner has been involved in a dispute with the board since August , when [he$_{\mathtt{A0}}$] [ousted$_{\textit{pred}}$] [all the directors$_{\mathtt{A1}}$].

Arg2: Later they said [they$_{\mathtt{A0}_{\text{coref}(\mathtt{Arg1}-\mathtt{A1})}}$] [fired$_{\textit{pred}}$] [him$_{\mathtt{A1}_{\text{coref}(\mathtt{Arg1}-\mathtt{A0})}}$], and two directors attempted to place the company under bankruptcy-law protection.

Label: 10XXX (Retain)

9.2.2 Network Architecture

The goal of all ensuing experiments (described in detail in the next section) is to construct a classifier which is capable of generalizing over the context information in both discourse arguments, and to correctly produce the appropriate cross-argument label for any given discourse relation by detection of potential coreferentiality among semantic roles. The general idea here is to re-use the powerful discourse framework from Chapter 4 and to test whether it can be applied to the task of "backward-looking" semantic role labeling, thus mapping global discourse parsing onto the local sentence basis of the second discourse argument.

In this specific resource-lean setting, we will only make use of the plain surface-level information, i.e. the *tokens* in the two discourse arguments. Thus, technically, a suitable training instance representation comprises the tokens in Arg1 and Arg2 together with the *two associated predicates* (one from each discourse argument), as well the label as a 5-tuple of the cross-argument SRL realization in the second discourse argument—indicating potential coreferentiality among shared semantic roles. As an illustration, consider Example (35).[9] The coreferential SRL-pattern has the label 06XXX (Continue) and the predicate pair is *expect.v.01-think.v.01.*

(35) Arg1: And IBM said [it$_{\mathtt{A0}}$] [expects$_{\textit{pred}}$] [the costs to continue climbing$_{\mathtt{A1}}$].

Arg2: IBM said [it$_{\mathtt{A0}_{\text{coref}(\mathtt{Arg1}-\mathtt{A0})}}$] [thought$_{\textit{pred}}$] [more companies would become interested as the project progresses$_{\mathtt{A1}}$].

Label: 06XXX (Continue)

Label	Transition Rel.	Description	Frequency	Proportion
56XXX	NoRel/R-shift	A0 A1	100,551	38.45%
X6XXX	NoRel	A1	50,285	19.23%
X67XX	NoRel	A1 A2	42,551	16.27%
567XX	NoRel/R-shift	A0 A1 A2	14,547	5.56%
5XXXX	NoRel/R-shift	A0	11,701	4.47%
XXXXX	NoRel	no roles	8,358	3.20%
XX7XX	NoRel	A2	6,753	2.58%
06XXX	Continue	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A1	6,449	2.47%
5X7XX	NoRel/R-shift	A0 A2	3,628	1.39%
X6X8X	NoRel	A1 A3	1,796	0.69%
16XXX	Smooth-shift	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A1	1,576	0.60%
X6XX9	NoRel	A1 A4	1,394	0.53%
56X8X	NoRel/R-shift	A0 A1 A3	1,326	0.51%
067XX	Continue	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A1 A2	845	0.32%
X07XX	Retain	A1$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A2	782	0.30%
X67X9	NoRel/R-shift	A1 A2 A4	731	0.28%
X17XX	R-Shift/Retain	A1$_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A2	673	0.26%
XXX8X	NoRel	A3	498	0.19%
X6X89	NoRel	A1 A3 A4	475	0.18%
0XXXX	Continue	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A2	456	0.17%
5XX8X	NoRel/R-shift	A0 A3	395	0.15%
X678X	NoRel	A1 A2 A3	386	0.15%
X1XXX	Retain/R-shift	A1$_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$	372	0.14%
X0XXX	Retain	A1$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$	347	0.13%
26XXX	Smooth-shift	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A2})}$ A1	272	0.10%
167XX	Smooth-shift	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A1 A2	237	0.09%
56XX9	NoRel/R-shift	A0 A1 A4	234	0.09%
XXXX9	NoRel	A4	229	0.09%
5678X	NoRel/R-shift	A0 A1 A2 A3	224	0.09%
51XXX	R-shift/Retain	A0 A1$_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$	221	0.08%
50XXX	Retain/R-shift	A0 A1$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$	207	0.08%
0X7XX	Continue	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A2	206	0.08%
507XX	Retain/R-shift	A0 A1$_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A2	197	0.08%
5XXX9	NoRel/R-shift	A0 A4	194	0.07%
1XXXX	Smooth-shift	A0$_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$	179	0.07%
X6789	NoRel	A1 A2 A3 A4	176	0.07%
56X89	NoRel/R-shift	A0 A1 A3 A4	149	0.06%
# instances:			234,746	100.0%

Table 9.1: Cross-argument labels and corresponding centering transitions

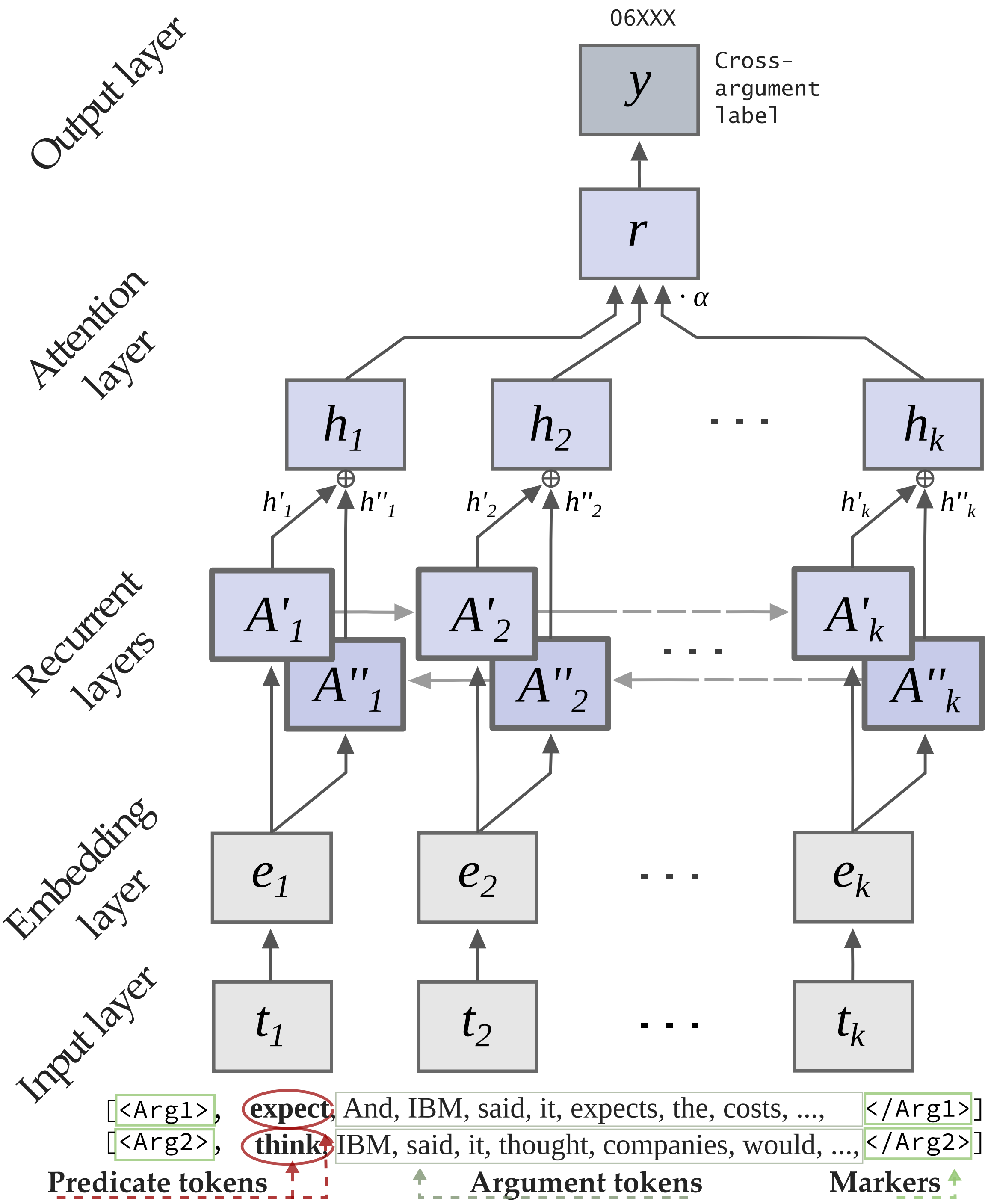


Figure 9.1: Modeling local coherence and cross-argument SRL structures with an attention-based bidirectional LSTM discourse network.

Figure 9.1 illustrates the overall network architecture for modeling cross-argument SRL in between two discourse segments. The graphics includes the training instance from Example (35) at the bottom of the visualization. Very similar to the method outlined in Chapter 4 on implicit discourse parsing, the tokens in each training instance are pre- and post-fixed with special argument boundary markers (<ARG1>, </ARG1>, etc.) to denote the beginning and end position of a relation, respectively. Directly following the start marker of an argument boundary, the *lemma of the predicate* is appended. These (normalized) predicates are duplicated in each argument to disambiguate discourse relations (highlighted in purple at the bottom of Figure 9.1).[10] Without any structural modification, overall, the architecture in Figure 9.1 is equivalent to the one presented in Chapter 4 on implicit discourse parsing. Technically, it is a recurrent neural network, in which input tokens (together with the pre-fixed predicates, highlighted in purple and green, resp.) are first encoded into an embedding layer by substituting raw words by distributed word representations in the first place. Three modular components are stacked on top: a layer of LSTM networks in a bidirectional manner, in order to better capture long-distance dependencies between extended parts of the input sequence by inspection of both left and right-hand-side contexts at each time step, an attention layer, and a final softmax output layer for cross-argument label classification (denoted by y in the visualization). More details can be referred to in the original description in Section 4.2. Also, note that the modeling procedure does not differ from the methodology outlined before: tokens are again analyzed *sequentially*, which has main advantages over a bag of embeddings representation. In what follows, we introduce three experiments and demonstrate the suitability of our method for the task of cross-argument SRL.

9.3 Evaluation

In this section, we outline a series of three experiments. In the first one (Section 9.3.1), we test whether our proposed architecture can model SRL patterns in general. Here, we restrict the analysis to *standard, i.e. non-coreferential SRL patterns*, the vast majority of cases as seen from Table 9.1. The second experiment (Section 9.3.2) focuses on *coreferential patterns* only. This task is harder because the label distribution is more skewed. A final experiment (Section 9.3.3) is concerned with *contrasting standard SRL with coreferential argument realizations*. Here, we try to test whether our method can detect the latent properties of coreferentiality in between discourse arguments in order to distinguish between the two types of relations. To this end, we provide visualizations of attention activity to selectively pinpoint those tokens which contribute to the decision on a coreferential SRL label.

In the first two experiments, we stick to the well-established data splits of

[8] Document ID wsj_1215

[9] Document ID wsj_1004

[10] Note that the two predicates need to be a distinct part of the representation, because the *same* Arg1-Arg2-pair can have *different* labels—depending on the chosen predicate combination.

the Penn Treebank and accompanying shared tasks by keeping the natural label distribution within the data sets, i.e. for training, we use WSJ sections 02-21, for development, WSJ sections 22 and 24, and for the test set, WSJ Section 23. In the third experiment, in order to assess the appropriateness of our classifier to distinguish standard SRL from coreferential SRL patterns, we chose to generate same sized sets of training instances. This setting seems most reasonable because the label distribution is highly skewed (≈5% coreferential labels vs. 95% non-coreferential labels). To be more precise, we selected random instances of the most frequent standard SRL pattern (`56XXX/NoRel` or ROUGH-SHIFT) and the most frequent coreferential SRL pattern (`06XXX/`CONTINUE), respectively, and trained a binary classifier. Note that we consider all experiments as a general proof of concept by re-using an existing discourse architecture and applying it to the task of local SRL. We refrain from extensively tuning hyperparameters for model optimization. Instead, we consider the majority class proportion in each data set as a solid and strong baseline for our experiments, in order to test whether our method can be generally useful and further extended.

A final remark is concerned with technical details of the training procedure: as embeddings, we employ the pre-trained Google News vectors for English with dimensionality $d = 300$ from *word2vec*[11] (Mikolov et al., 2013a). The parameters of the model are not modified with respect to the previous setting (Chapter 4), except for the maximum sequence length, which is shortened to 80 tokens for English sentences, and the dropout rate (for both recurrent and attention layer) set to 0.8.

9.3.1 Experiment 1—Standard SRL

Table 9.2 shows the results for the task of modeling *standard SRL patterns* of the seven most frequent realizations based on their natural distribution in the WSJ training, development and test sections.[12] The model performance is optimal after 9 epochs (development set performance 82.22%), and the test set majority class baseline for the label `56XXX` (non-coreferential agent and patient in `Arg2`) can be beaten by 16.5% absolute improvement in accuracy. The results demonstrate that our proposed technique is generally capable of modeling standard SRL patterns (`NoRel` or ROUGH-SHIFT), by obtaining a high degree of generalization, especially on the three most frequent labels (which account for more than 80% of all relations in this subset).

[11] `https://code.google.com/p/word2vec/`

[12] In all experiments, # indicates the number of instances in the respective data set, % the proportion, and majority class baselines are highlighted in bold in the first row. Model accuracies for development and test sets are shown in the last row.

	Pattern	Training Set		Development Set		Test Set		**Performance**		
label:		#	%	#	%	#	%	P	R	F_1
56XXX	A0 A1	84,585	43.41	7,170	37.78	8,796	**42.01**	60.1	87.7	71.3
X6XXX	A1	40,651	20.86	5,018	26.44	4,616	22.05	47.9	34.7	40.3
X67XX	A1 A2	35,200	18.07	3,397	17.9	3,954	18.89	64.8	68.2	66.5
567XX	A0 A1 A2	12,270	6.3	1,041	5.49	1,236	5.9	48.9	7.6	13.1
5XXXX	A0	9,739	5.0	1,007	5.31	955	4.56	57.2	7.0	12.5
XXXXX	no roles	6,848	3.51	815	4.29	695	3.32	43.2	6.9	11.9
XX7XX	A2	5,539	2.84	529	2.79	685	3.27	16.1	4.2	6.7
Model (accuracy):					**82.22**		**58.55**			

Table 9.2: Label distribution and test set performance for **standard SRL** (Exp. 1)

9.3.2 Experiment 2—Coreferential SRL

The top part of Table 9.3 shows the label distribution and classification scores (after 26 epochs) for *coreferential SRL* on the 10 most frequent labels as they occur in their natural distribution in the data sets. The results of our model only marginally outperform the majority class baseline of 57.44% by an absolute improvement of 1.83% on the WSJ test set patterns. This is mainly due to the highly skewed label distribution and data sparsity issues related to the less frequent labels. A second experiment has thus been concerned with a slight modification of the data set by removing the most frequent class label `06XXX`—resulting in only 9-way classification more evenly distributed data instances. As can be observed from the lower part in Table 9.3, the model is able to obtain a higher degree of generalization (≈6% performance improvement over the majority class baseline with default parameters), yet the overall challenge of modeling coreferential SRL patterns remains as our model can only successfully generalize over four out of nine labels (cf. performance scores in the lower part of Table 9.3). Still, we believe that the results of this experiment are promising as it turns out that our proposed architecture is capable of modeling local discourse coherence and that distinguishing between continuations, retentions, and shifts seems generally feasible.

9.3.3 Experiment 3—Standard vs. Coreferential SRL

The final experiment concerns the contrast between standard SRL patterns with coreferential patterns and to investigate whether a recurrent discourse architecture is capable of detecting latent properties within these two distinct types of argument pairs. We have specifically focused our attention on the two most frequent labels from each class, `56XXX` (agent and patient role present without coreferentiality, i.e. `NoRel` or ROUGH-SHIFT) and `06XXX` (agent and patient role present, but agent coreferential with `A0` in the first discourse argument, CONTINUE). Due to the highly unbalanced label distribution in the data sets, we have created random samples of two same sized sets. The performance scores of our model (after 9 epochs) based on 5,000 training instances, 500 development and 500 test set instances are shown in Table 9.4.

Generally, with our proposed technique, it is possible to outperform the majority class baseline (50%) by a large margin of 23.9% absolute improvement in accuracy. Even without (hyper-)parameter optimization, the existent discourse architecture can in essence be applied out-of-the-box to the task of modeling coreferential SRL patterns by distinguishing them from non-coreferential SRL realizations. The model seems capable of generalizing well across both label instances (F_1-scores between 70-77%). In what follows, we investigate the model's learned attention activities and elaborate in closer detail on the latent properties which drive the classification decisions.

Investigating & Visualizing Attention Activity: An informal, visual inspection on the distribution of attention activity scores in both types of relations conveys the impression that coreferential discourse relations exhibit a greater attention activity in *the second discourse argument* as opposed to non-coreferential ones; average Arg2 scores ≈ 0.400 for coreferential (06XXX) vs. ≈ 0.375 for non-coreferential (56XXX) relations based on all test instances of Exp. 9.3.3. What we can observe is that scores on the boundary between Arg1 and Arg2 for coreferential discourse arguments follow a rather smooth transition which mostly results in a slight *increase* towards the end of the discourse relation. However, for non-coreferential discourse arguments, there seems to be a decrease in attention activity along with the first tokens of the second discourse argument. We assume that the two types of relations exhibit distinct latent properties. Since this assumption needs to be tested formally, we addressed this phenomenon by two statistical tests i.) comparing average activities between coreferential arguments and non-coreferential arguments on all tokens, as well as ii.) between Arg2 means only. We report on the results in the following:

i.) **Comparison of overall average attention between coreferential and non-coreferential relations:**
According to a Welsh Two Sample t-test, there exists a statistically significant difference between the means of attention weights in both types of relations: t = -2.6178, df = 761.29, p-value = 0.009025, with means: 0.404 for label 06XXX, 0.412 for label 56XXX.

ii.) **Comparison of average attention in second discourse arguments between coreferential and non-coreferential relations:**
According to a Welsh Two Sample t-test, there exists a statistically significant difference between the Arg2 means of attention weights in both types of relations: t = 4.8302, df = 694.21, p-value = 1.68e-06, with means: 0.400 for label 06XXX, 0.375 for label 56XXX.

	Pattern	Training Set		Dev Set		Test Set		**Performance**		
label:		#	%	#	%	#	%	*P*	*R*	F_1
06XXX/CONTINUE	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A1	5,328	53.50	430	50.77	691	**57.44**	60.0	98.5	74.5
16XXX/S-SHIFT	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A1	1,282	12.87	130	15.35	164	13.63	42.4	17.0	24.3
067XX/CONTINUE	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$A1 A2	704	7.07	44	5.19	97	8.06	–	0.0	–
X07XX/RETAIN	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A2	650	6.53	62	7.32	70	5.82	50.0	1.4	2.7
X17XX/R-SHIFT/RETAIN	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A2	576	5.78	58	6.85	39	3.24	–	0.0	–
0XXXX/CONTINUE	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$	387	3.89	32	3.78	37	3.08	–	0.0	–
X1XXX/RETAIN/R-SHIFT	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$	310	3.11	30	3.54	32	2.66	–	0.0	–
X0XXX/RETAIN	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$	286	2.87	28	3.31	33	2.74	–	0.0	–
26XXX/S-SHIFT	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A2})}$ A1	234	2.35	16	1.89	22	1.83	–	0.0	–
167XX/S-SHIFT	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A1 A2	202	2.03	17	2.01	18	1.5	–	0.0	–
Model (accuracy):					**67.18**		**59.27**			

	Pattern	Training Set		Dev Set		Test Set		**Performance**		
label:		#	%	#	%	#	%	*P*	*R*	F_1
16XXX/S-SHIFT	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A1	1,282	27.68	130	31.18	164	**32.03**	35.8	87.1	50.7
067XX/CONTINUE	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$A1 A2	704	15.20	44	10.55	97	18.95	44.6	29.8	35.8
X07XX/RETAIN	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A2	650	14.04	62	14.87	70	13.67	46.8	31.4	37.6
X17XX/R-SHIFT/RETAIN	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A2	576	12.44	58	13.91	39	7.62	100.0	2.5	5.0
0XXXX/CONTINUE	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$	387	8.36	32	7.67	37	7.23	–	0.0	–
X1XXX/RETAIN/R-SHIFT	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$	310	6.69	30	7.19	32	6.25	–	0.0	–
X0XXX/RETAIN	$\texttt{A1}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$	286	6.18	28	6.71	33	6.45	–	0.0	–
26XXX/S-SHIFT	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A2})}$ A1	234	5.05	16	3.84	22	4.30	–	0.0	–
167XX/S-SHIFT	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A1})}$ A1 A2	202	4.36	17	4.08	18	3.52	–	0.0	–
Model (accuracy):					**47.72**		**38.09**			

Table 9.3: Label distribution and performance for **coreferential SRL** (Exp. 2) with natural label distribution (top), without most frequent class (bottom)

	Pattern	Training Set		Development Set		Test Set		**Performance**		
label:		#	%	#	%	#	%	*P*	*R*	F_1
56XXX/NoRel/R-SHIFT	A0 A1	5,000	50.00	500	50.00	500	**50.00**	83.1	60.0	69.6
06XXX/CONTINUE	$\texttt{A0}_{\text{coref}(\texttt{Arg1}-\texttt{A0})}$ A1	5,000	50.00	500	50.00	500	**50.00**	68.7	87.8	77.0
Model (accuracy):					**94.80**		**73.90**			

Table 9.4: Labels and performance for the two most frequent classes in **standard vs. coreferential SRL** (Exp. 3)

Turning to the attention weight visualizations, Figures 9.2 and 9.3[13] show three randomly selected coreferential sentence pairs along with their activity scores and the associated tokens.[14] Interestingly, as can be observed in the upper graphics of Figure 9.2 for instance, on almost every predicate (*drink*, *whistle*, *talk*, *read*) its attention weight is minimally greater than the attention weight of the immediately preceding token (highlighted in light and intense yellow colors). This suggests that the model places special focus on these words, which seems plausible, given that verbal word categories are a driving force within the semantic role labeling setting in which our learning framework is grounded. A very similar pattern is evident in the bottom graphics of Figure 9.2. Also, note that in Figure 9.3, we see a related trend; this time, however, do pronouns account for the step-up in the scores (esp. in Arg2), which seems intuitive given the fact that coreference is mostly indicated by referring expressions between discourse arguments.

As pointed out previously, the attention activities in coreferential argument pairs exhibit a more interesting overall "flow" as opposed to the non-coreferential ones, which typically show a monotone decrease in the strength of activity towards the end of a discourse relation. Both the coreferential sentence pairs of Figures 9.2 and 9.3 exemplify and confirm this hypothesized trend in the second argument. Note that we can straightforwardly relate this effect by analogy with the observations made in Chapter 4 on contrastive patterns between *entity relations* and *conjunction relations*. Here, we found that entity relations distinguish themselves from (ordinary) conjunction relations in a very similar behavior: attention activities in second arguments increase for entity relations, yet they remain rather stable in conjunction relations. We have attributed this effect to the model's learned capability to detect the appearance of additional semantic information (in Arg2) *related* to the *same* coreferential entity (in Arg1); cf. Chapter 4, Figure 4.3. Thus, we conclude that entity relations share structural properties with and behave very similar to coreferential ones, which seems plausible because both of these relations *are*, in fact, coreferential.

9.4 Summary

This chapter aimed at setting cross-sentential discourse coherence within the local context of semantic roles. To this end, we have presented a joint modeling framework for SRL beyond the sentence-level, which we termed cross-argument semantic role labeling. In this specific setting, we established a direct connection between two events (one predicate from each discourse unit and their realized roles) and associated them with a specific cross-argument label, for which the different types of discourse coherence determine the exact form of the label. Our main motivation for the concept of a cross-argument label is inspired from Cen-

[13] From document IDs wsj_0037, wsj_1366, and wsj_1273, respectively.

[14] The graphics have been produced with a low dropout rate of 0.1 on the attention layer. We observed that higher dropout rates lead to slightly better classification accuracies but less pronounced activity patterns.

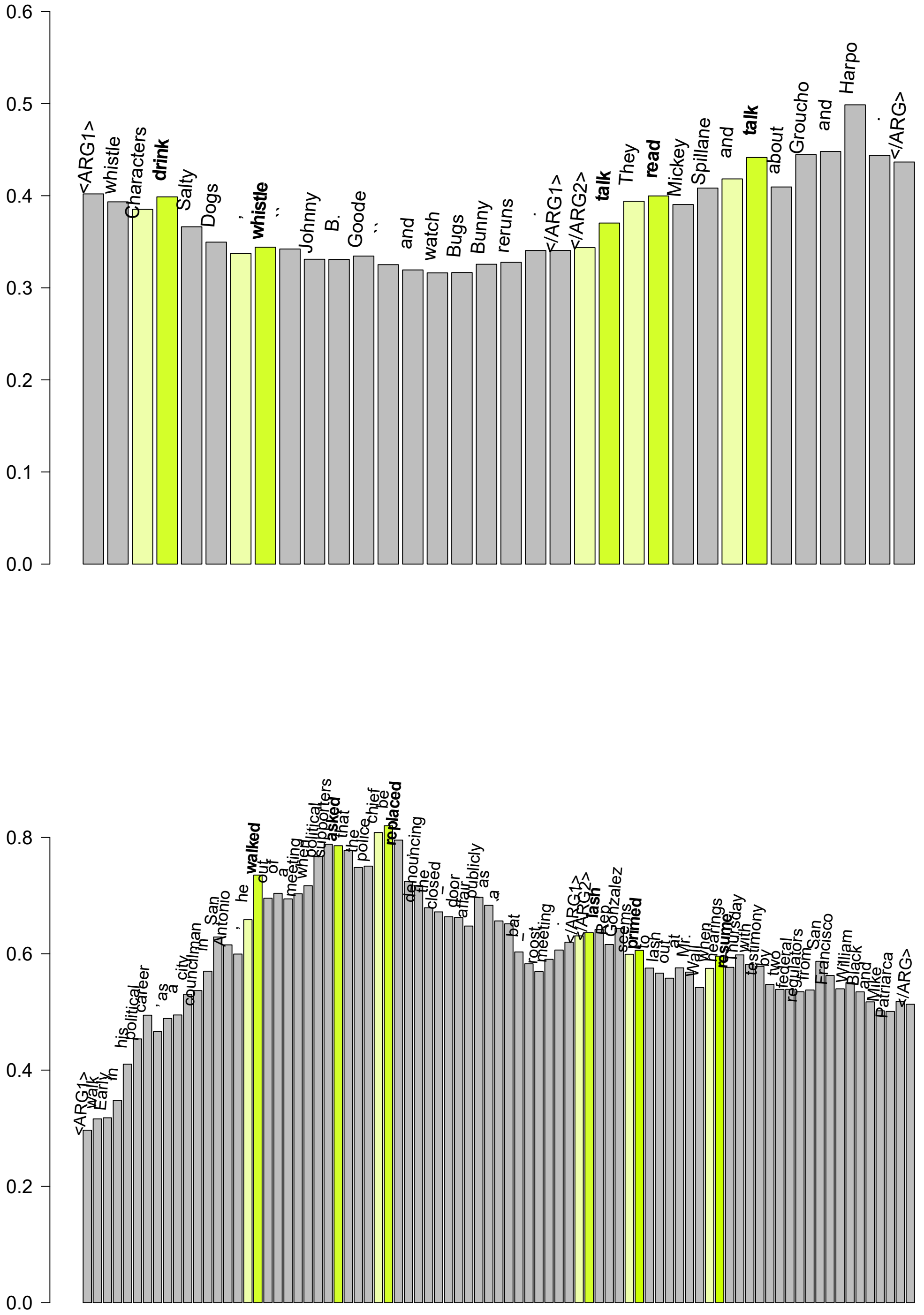


Figure 9.2: Attention activities of two coreferential sentence pairs (06XXX/Continue) with predicate scores highlighted in yellow.

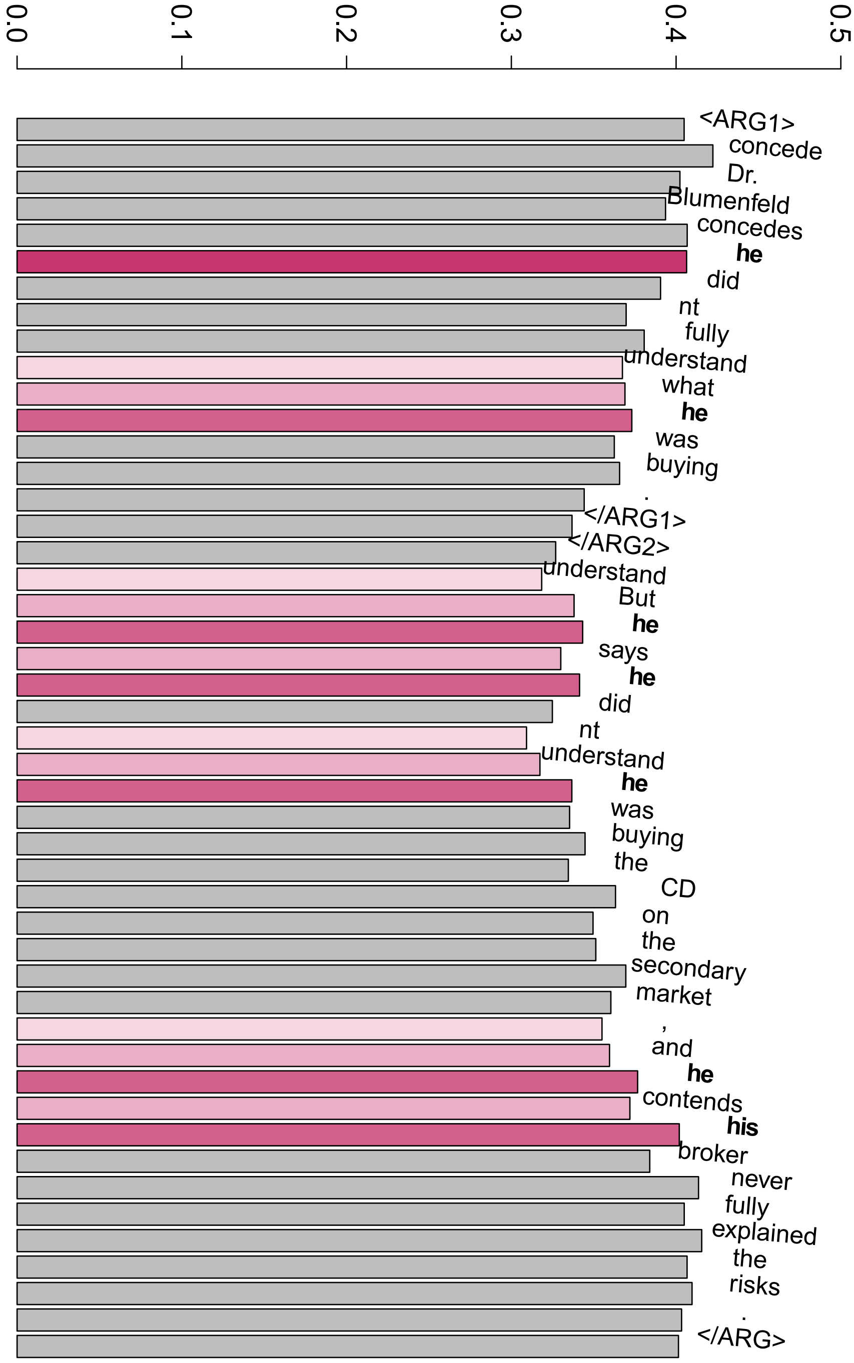


Figure 9.3: Attention activity for a coreferential sentence pair (06XXX/Continue) and pronoun scores highlighted in purple.

tering Theory and serves the purpose of the backward-looking center by back-reference from the current to any previous utterance under investigation of entity relations. An obvious advantage of having event-driven labels instead of a single backward-looking center can be seen in a more fine-grained and at the same time more flexible assessment of local discourse coherence, because our relation scheme is predicate-specific, it involves the interrelation of explicitly realized roles and their meaning, and does not rely only on a single salient entity. In a closely related series of experiments in this chapter, our focus was placed on a computational approach to entity relations as these relations typically exhibit anaphoric relationships between subsequent utterances, and their peculiarities have been extensively studied in the theoretical literature before.

Technically, we have realized cross-argument SRL by reusing an existing discourse architecture, which has already been shown to successfully recover implicit discourse senses in between adjacent sentence pairs. Without any structural modification, we have applied it to the task at hand—classifying both coreferential and non-coreferential labels; suitable training instances were extracted from gold annotations in large corpora. Note, that we considered a semantic role pattern as coreferential if it shares at least one argument slot in the local context with another entity found in the preceding discourse. Non-coreferential relations were either typically due to rough shifts (because another entity was introduced) or, in fact, due to the absence of any relation (when no entity was present).

A quantitative evaluation has shown that our proposed architecture is generally suitable and powerful enough to model, distinguish and highlight peculiarities of both types of relations (coreferential vs. non-coreferential). In the case of coreferential pairs, it turned out that our system can in fact distinguish between the classical transition types of continuations, retentions, and shifts, even though the model performance still leaves some room for improvement in terms of hyperparameter optimization. An closer investigation of the attention weights learned by the model lead to the following two major findings: First, it can evidently pinpoint the latent features which drive the classification of coreferential discourse arguments, i.e. our model shows intuitive preference for predicates and pronouns. Second, our acquired coreferential sentence pairs behave almost identical to previous observations made with regards to attention activity on entity relations from the Chinese Discourse Treebank. We conclude that our proposed model can indeed learn the structural peculiarities of entity-based coherence relations.

The next chapter revisits the different transition types in entity-based coherence relations. We introduce an algorithm based on implicit discourse structure and show that it can be successfully applied to the practical downstream task of narrative understanding.

Software: The code for the recurrent neural network model outlined in this chapter is publicly available from the following URL: `http://www.acoli.informatik.uni-frankfurt.de/resources.html`.

Chapter 10

Extension: Modeling Story Coherence

10.1 Motivation

The previous chapter has presented a computational framework for modeling the different transition types in entity-based coherence relations. In this extension, we revisit continuations, retentions, and shifts, and describe their application to the challenging downstream task of *narrative understanding*. In fact, semantic applications related to Natural Language Understanding (Allen, 1995) have seen a recent surge of interest within the NLP community, and *story understanding* can be regarded as one of the high-level disciplines in that field from which, for example, automated reasoners or question answering systems could greatly benefit. In order to enable deep language comprehension it is essential to build a system which can cope with the complexities and challenges of modeling procedural commonsense knowledge in natural language descriptions. Closely related to Machine Reading (Hovy, 2006) and script learning (Schank and Abelson, 1977; Mooney and DeJong, 1985), story understanding is built on top of a cascade of core NLP applications, including—among others—event extraction (UzZaman and Allen, 2010), (implicit) semantic role labeling (Gerber and Chai, 2012) and—most notably—discourse processing. Regarding the latter, there has been emerging research drawing upon related aspects of, for instance, causal or temporal relation recognition (Mirza and Tonelli, 2016), or inter- and extra-sentential relation classification for implicit discourse relations and entity relations. Concerning the latter, we have thoroughly outlined two approaches in Chapters 3 and 4, respectively.

In the related field of narrative understanding recent progress has been made, and a variety of successful approaches have been introduced, ranging from narrative chains (Chambers and Jurafsky, 2008) to script learning techniques (Regneri et al., 2010), or event schemas (Nguyen et al., 2015). What all these approaches have in common is that they ultimately seek to find a way to prototypically model the causal and correlational relationships between events, and also to obtain a structured (ideally more compact and abstract) representation of the underlying commonsense knowledge which is encoded in the respective story. Concerning this matter, as we pointed out in Chapter 2 (Section 2.2.3.1) the downside of these approaches is that they are feature-rich (potentially hand-crafted) and therefore costly and domain-specific to a large extent.

In the style of the previously outlined resource-lean attempts to processing language data, this chapter specifically addresses this matter and proposes a lightweight framework for modeling procedural knowledge in commonsense stories whose only source of information are again distributed word representations. By extension of the methodology introduced in Chapter 3, the problem of modeling text coherence is cast as a special case of discourse processing for implicit relations in which the proposed model jointly learns to distinguish correct from incorrect story endings. The approach is inspired by promising related attempts using event embeddings and neural methods for script learning (Modi and Titov, 2014; Pichotta and Mooney, 2016). The system that we present is an end-to-end implementation of the ideas sketched in Mostafazadeh et al. (2016b) of the *joint paragraph and sentence level* model (cf. Section 10.2 for details). The approach is evaluated in the *Story Cloze Test* (Mostafazadeh et al., 2017, 2016a, cf. Section 10.1.1), a task for predicting story continuations. Despite its simplicity, the system outlined in this chapter demonstrates superior performance on the designated data over previous approaches to script learning and—due to its language and genre-independence—it also represents a solid basis for further optimization towards other textual domains.

10.1.1 The Story Cloze Test—Task Description & Data

In the Story Cloze Test and its recent accompanying shared task (Mostafazadeh et al., 2017), an automated system is presented with a natural language description consisting of four sentences (the *core story*) along with two alternative single-sentence endings, i.e. a correct and a wrong one. The system is then supposed to select the correct ending based on a semantic analysis of the individual story components.

Recently, the shared task organizers have provided participants with a large corpus of approximately 98k five-sentence everyday life stories (Mostafazadeh et al., 2016a, *ROCStories*[1]) which have been obtained by crowdsourcing and which were released for training their narrative story understanding models. Also a validation and a test set have been made available (each containing 1,872 instances). The former serves for parameter optimization, whereas final performance is evaluated on the test set. The instances in all three sets are mutually exclusive. Note that in addition to the *ROCStories*, both validation and test sets include an additional *wrong* 5th-sentence story ending (either in first or second position) plus hand-annotated decisions about which story ending is the right one. As an illustration, consider the examples in 10.1, 10.2, and 10.3, each consisting of a core

[1] http://cs.rochester.edu/nlp/rocstories/

story and two alternative continuations (called quizzes).[2]

In the narrative structure of these three example stories, we can clearly observe the analogy to the classical transition relations from Centering Theory (Grosz et al., 1995) in the form of SHIFT, CONTINUE, and RETAIN, which find their expression in anaphoric references between the individual story parts. A manual inspection of the data reveals that the vast majority of examples are continuation transitions between entities—the default coherence transition according to centering, i.e. roughly speaking the most salient entity is carried over from one sentence to the next. Crucially, the global semantics of these stories can be attributed to one particular factor. For instance, when we inspect the *ROCStory* in Table 10.1, we can can determine a latent *discourse structure* by virtue of a temporal/causal relationship that governs the connection between the individual events in each sentence. Based on the positive outcome of the story we can clearly say that the right ending is the second quiz. It is noteworthy that for all stories in the data set, the task of choosing the correct ending is human solvable with perfect agreement according to Mostafazadeh et al. (2016a). Similar observations related to discourse coherence apply as well to the other examples from the data set. We elaborate on details in the next section.

Four-Sentence Core Story	Quiz 1	Quiz 2
I asked Sarah out on a date. She said yes. I was so excited for our date together. We went to dinner and then a movie.	I had a terrible time. (***wrong*** ending)	I got to kiss Sarah goodnight. (***correct*** ending)

Table 10.1: A *ROCStory* consisting of a core story and two alternative continuations (Mostly CONTINUATION transitions, one ROUGH-SHIFT).

10.2 Design Principles & Network Architecture

In this section, we describe the theoretical motivation, our proposed model architecture for finding the right story ending, and how the training procedure is implemented.

10.2.1 Theoretical Motivation

As pointed out before, most of the sentence transitions in the data set (especially between the core story and their associated endings) are structurally very similar

[2] These examples were randomly selected from the test and validation sets with IDs: `fc416bf8-23b2-41af-b4bf-ffc544321166`, `c6e09baa-51dc-454c-a77c-8ebe7c2c63a7`, `83c7ec4c-f2e3-4474-8897-5cab30a27042`.

Four-Sentence Core Story	Quiz 1	Quiz 2
Felipe had been interested in Germany since he was a young child. He learned about the language and culture by reading library books. Finally in college he was able to visit Berlin. Felipe loved the German art, food, and beer.	Felipe visited Germany again later in his life. (***correct*** ending)	Felipe never wanted to go to Germany again. (***wrong*** ending)

Table 10.2: A *ROCStory* consisting of a core story and two alternative continuations (predominantly Continue transitions).

Four-Sentence Core Story	Quiz 1	Quiz 2
Simon had a kitten called Tiny. Tiny was mischievous and often pushed things off of furniture. Once Tiny pushed a glass of the table. Simon's parents walked in and saw the smashed glass on the floor.	Simon's parents were mad at Tiny. (***correct*** ending)	They loved Tiny. (***wrong*** ending)

Table 10.3: A *ROCStory* consisting of a core story and two alternative continuations (all instantiations of Continue, Retention, Shift).

to continuations. We argue that continuous entity-based coherence in the form of sequences of narrative events can be modeled in two ways: either by means of entity continuity, or on the basis of discourse continuity. The former approach is rather superficial and employs coreference or anaphora resolution. It aims at connecting different (explicit) mentions of the same entity, and tries to derive a suitable meaning representation by mainly capturing the most salient entity at each time step in the story. We claim, however, that this approach is not sensitive enough to model the true semantic relationship that holds between the actions expressed in the individual interconnected story sentences. This particular relationship is in fact present only in an underlying form and (besides entities, anaphors, and pronouns) it also relates to specific predicates and other important co-occurring words in the context. Since anaphors can sometimes be unexpressed, the latter approach (the one which implements discourse continuity) is more flexible and operates on discourse relations as a means to capture the cohesive links between story components, and it is most notably concerned with the implicit senses that hold between the story units. For the realization of this experiment, we argue for the use of an (implicit) discourse coherence model, because their relationship is typically not signaled by connectives in the data (such as *but* or *because*).

On the one hand, we have already successfully demonstrated that properties of local coherence and esp. entity relations can be modeled by means of a neural network architecture. (cf. Chapters 3, 4, 9). On the other hand, we have proven the practicability of a lighweight feedforward system to recognize implicit discourse structure in between adjacent sentences. (Chapter 3). In this experiment we combine both aspects by adopting our feedforward approach and adapting it to the task at hand, thereby building on the works from (shallow) discourse parsing, most notably on the recent success of neural network-based frameworks in that field, cf. Xue et al. (2016); Wang and Lan (2016). Specifically for *implicit* discourse relations, i.e. for those sentence pairs which, for instance, can signal a temporal, contrast or contingency relation, but which suffer from the absence of an explicit discourse marker (such as *but* or *because*), it has been shown that the interaction of properly tuned distributed representations over adjacent text spans can be particularly powerful in the relation classification task. Very similarly, we argue that the Story Cloze test can be cast as a special case of implicit discourse relation recognition by attempting to model an underlying, latent connection between a core story and its correct vs. incorrect continuation.

As an illustration, consider the final sentence of the core story in Example 10.1 and the two adjacent quizzes which can be treated as argument pairs (`Arg1` and `Arg2`) in the classical view of the Penn Discourse Treebank (Prasad et al., 2008). Crucially, we can distinguish different types of implicit discourse senses that hold between them.

(36) `Arg1`: We went to dinner and then a movie.
`Arg2`: I had a terrible time.

Implicit discourse sense: TEMPORAL:**Synchronous**
Inferred connective: e.g., *while*.

(37) `Arg1`: We went to dinner and then a movie.
`Arg2`: I got to kiss Sarah goodnight.

Implicit discourse sense: TEMPORAL:**Asynchronous**:precedence
Inferred connective: e.g., *then*.

Here, in the first example, the label "Synchronous" indicates that the two situations in both arguments overlap temporally (which could be signaled explicitly by *while*, for instance), whereas in the second example Asynchronous:precedence implies a temporal order of both events. The distinction between different implicit discourse senses are subtle nuances and are highly challenging to detect automatically; however, they are typical of the *ROCStories*, as almost *no* explicit discourse markers are present between the individual story sentences. Finally, note that the motivation for this approach is also related to the classical view of recognizing *textual entailment* which would treat correct and wrong endings as the entailed and contradicted hypotheses, respectively Giampiccolo et al. (2007); Mostafazadeh et al. (2016a).

10.2.2 Training Instances

For the Story Cloze Test, a training instance is modeled as a triplet consisting of the four-sentence core story (C), a first quiz sentence (Q1) *and* a second quiz sentence (Q2) from which either Q1 or Q2 is the correct continuation of the story. Note that the original *ROCStories* contain only valid five-sentence sequences but the evaluation data requires a system to select from a pool of two alternatives. Therefore, for each single story in *ROCStories*, we randomly sample one negative (wrong) continuation Q_{wrong} from all last sentences, and generate two training instances with the following patterns: [C, Q1, Q2_{wrong}]:Label_1,[C, Q1_{wrong}, Q2]:Label_2, where the label indicates the position of the correct quiz. The motivation is to jointly learn core stories together with their true ending while at the same time discriminating them from semantically irrelevant continuations.

For each component in the triplet, we have experimented with a variety of different calculations in order to capture their idiosyncratic syntactic and semantic properties. We found the vector average over their respective words $\vec{v}^{avg} = \frac{1}{N}\sum_{i=1}^{N} E(t_i)$ to perform reasonably well, where N is the total number of tokens filling either of C, Q1 or Q2, respectively, resulting in three individual vector representations. Here, we define $E(\cdot)$ as an embedding function which maps a token t_i to its distributed representation, i.e. a precomputed vector of d dimensions. As distributed word representations, we chose out of the box vectors; GloVe vectors (Pennington et al., 2014), dependency-based word embeddings (Levy and Goldberg, 2014) and the pre-trained Google News vectors with $d = 300$ from *word2vec*[3] (Mikolov et al., 2013a). Using the same tool, we also trained custom embeddings (bag-of-words and skip-gram) with 300 dimensions on the *ROCStories* corpus. Punctuation symbols were in all settings.

10.2.3 Network Architecture

The feature construction process and the neural network architecture are depicted in Figure 10.1. The bottom part illustrates how tokens are mapped through three stacked embedding matrices for C, Q1 and Q2, each of dimensionality $\mathbb{R}^{d \times n}$. A second step applies the average aggregation and concatenates the so-obtained vectors $\vec{c}^{avg}$, $\vec{q_1}^{avg}$, $\vec{q_2}^{avg}$ (each $\vec{v}^{avg} \in \mathbb{R}^d$) into an overall composed story representation of dimensionality $\mathbb{R}^{3*d}$ which in turn serves as input to a feedforward neural network. The network is set up with one hidden layer and one sigmoid output layer for binary label classification for the position of the correct ending, i.e. first or second.

[3] `https://code.google.com/p/word2vec/`

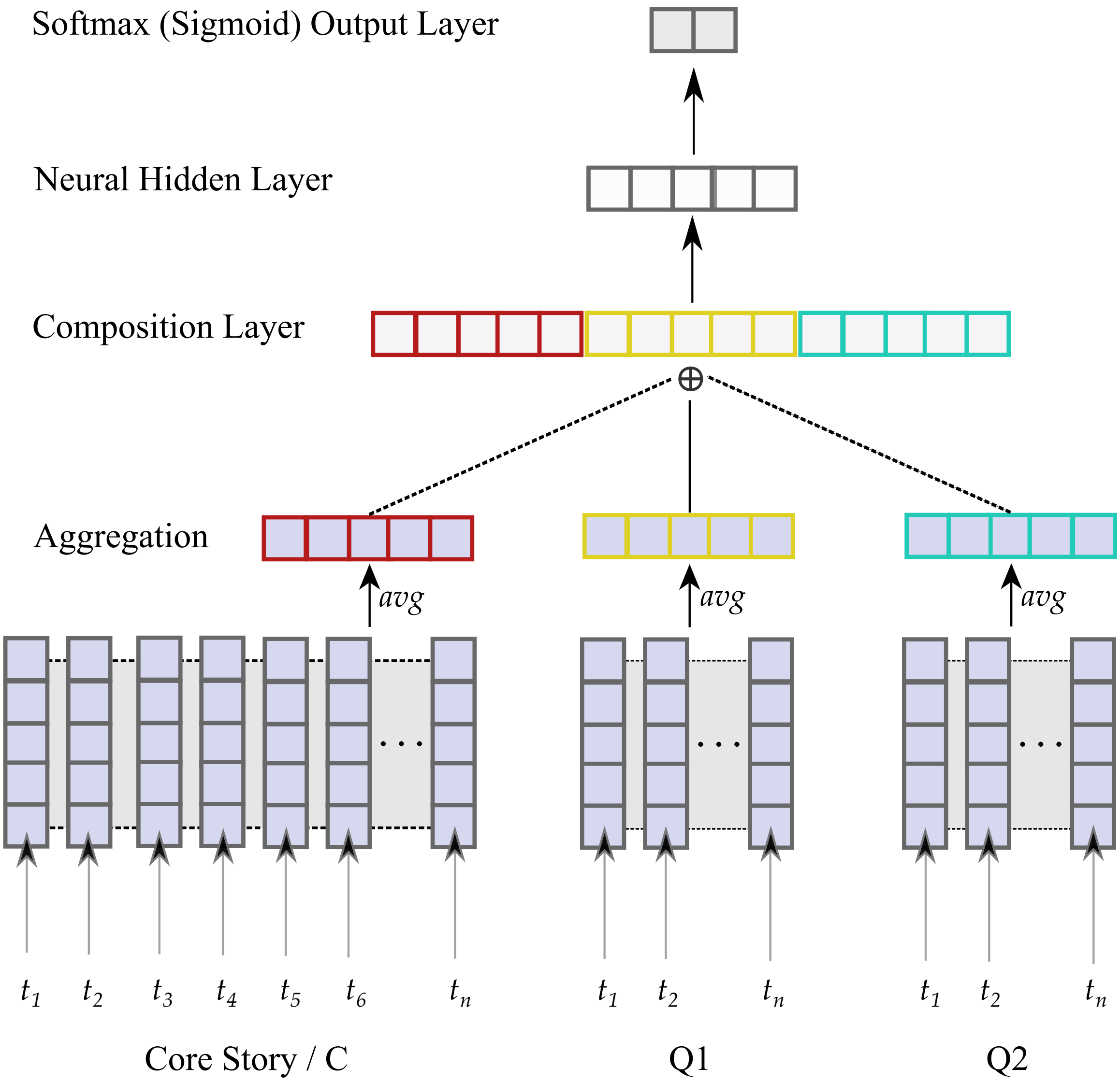


Figure 10.1: The proposed architecture for the Story Cloze Test. Depicted is a training instance consisting of three distributed word representation matrices for core story (C), quiz 1 (Q1) and quiz 2 (Q2), each component of varying length *n*. Note that either Q1 or Q2 is a wrong story ending. Matrices are first individually aggregated by average computation. Resulting vectors are then concatenated to form a composition unit which serves as input to the network with one hidden layer and binary output classification.

10.2.4 Implementational Details

The network is trained only on the *ROCStories* (and the negative training items), totaling approx. 200k training instances, over 30 iterations and 35 epochs with pretraining and a mini batch size of 120. All (hyper-)parameters are chosen and optimized on the validation set. We conduct data normalization, Xavier weight initialization (Glorot and Bengio, 2010) on the input layer, and employ rectified linear unit activation functions to both the composition layer and hidden layer with 220-250 nodes, and finally apply a sigmoid output layer for label classification. The learning rate is set to 0.04, l2 regularization = 0.0002 for penalizing network weights using the cross entropy error loss function. The network is trained using stochastic gradient descent and backpropagation.[4]

10.3 Evaluation

The model presented in this chapter is intrinsically evaluated on both the validation and the test set provided by the shared task organizers. As a reference, three re-implemented baselines are provided as well, which are borrowed from Mostafazadeh et al. (2016a) at the time when the data set was released, namely the best-performing algorithms inspired by Huang et al. (2013) (Deep Structured Semantic Model/DSSM) and Chambers and Jurafsky (2008) (Narrative-Chains). The fact that a system is supposed to select either of two possible continuations for any given *ROCStory* allows this binary choice to be straightforwardly evaluated on an accuracy level. Table 10.4 shows that correct endings appear almost equally often in either first or second position in the annotated data sets. The majority class is only significantly beaten by the DSSM model. The approach outlined in this chapter, denoted by *Neural-ROCStories*, however, can further improve upon the best system by an absolute increase in accuracy of 4.7%.

Only the best configuration is shown and has been achieved with the 300-dimensional pre-trained Google News embeddings (GloVe vectors and self-trained embeddings performed worse by $\approx$ 2% in accuracy). Interestingly, the performance of the model on the test set is slightly better that on the validation set but also very similar which suggests that it is able to generalize well to unseen data and is not prone to overfitting training or validation data. A manual inspection of a subset of the misclassified items reveals that the neural recognizer is struggling to properly handle story continuations which change the underlying sentiment of the core story either towards negative or positive, e.g. *fail test, study hard → pass test*.

It should be noted that in the official shared task initiated by Mostafazadeh et al. (2017) (whose detailed results can be referred to in the paper) the implementation of the system outlined in this chapter is ranked at position 6/9, which suggests that there is still potential for improvement. Nonetheless, this also shows that a lightweight feedforward system can be competitive with more sophisticated

[4] The model is implemented with the toolkit *deeplearning4j*: `https://deeplearning4j.org/`.

System		Performance	
		Validation	Test
DSSM	Huang et al. (2013)	0.604	0.585
Narrative-Chains	Chambers and Jurafsky (2008)	0.510	0.494
Majority Class	–	0.514	0.513
Neural-ROCStories	**Schenk and Chiarcos (2017)**	**0.629**	**0.632**

Table 10.4: Performances (in % accuracy) on the validation and test sets of *The Story Cloze Test* data. A reference comparison to two re-implemented baseline systems by Mostafazadeh et al. (2016a) is given.

tools, which is promising: In fact, most systems were based on full-fledged *recurrent* architectures, similar to the one described in the previous chapter, and have incorporated additional external resources. These architectures are more complex and harder to train and optimize and stand in contrast to the minimalist setting of a simple feedforward network which is capable of modeling story coherence, whose demonstration has been the main focus of this chapter.

Finally, we want to point out that such a simple model can be easily extended in various ways, without having to refer to the more complicated LSTM architectures. For instance, instead of having three individual (separate) components to model core story, quiz one and quiz two, respectively, an alternative strategy could follow the architecture of two sequential (feedforward) models.[5] These two sequential components could capture a tuple of core story plus additional ending directly in one joint representation. Alternative and very similar experiments on a variety of convolutional architectures have recently been proposed in the work by Feng et al. (2015). Here, the authors have applied convolutional neural network-based systems to the challenging task of question answering, in which for any given question a correct answer must be selected out of a set of candidates. Overall, this task is highly related to the Story Cloze Test as task setups are almost identical. Further explorations using a host of alternative network architectures for the Story Cloze Test data might be worth investigating.

10.4 Summary

In the style of a resource-lean method for detecting textual relationships between events, this chapter has introduced a highly generic neural recognizer for modeling text coherence, applied to a designated data set—the *ROCStories* for finding appropriate story continuations. The approach is inspired by successful models for implicit discourse relation classification and is a functional adaptation of the

[5] For instance, as outlined in `https://faroit.github.io/keras-docs/1.0.0/getting-started/sequential-model-guide/`.

methodology outlined in Chapter 3 to story coherence modeling: In particular, the recognition of latent, implicit discourse information between adjacent sentences was taken advantage of, and we practically extended the method to distinguish between correct and incorrect story continuations. Technically, the network architecture involved in this experiment is a lightweight feedforward system whose input only relies on the carefully tuned interaction of distributed word representations between story components. An evaluation of the model shows that the minimalist approach yields adequate results and can outperform prior state-of-the-art algorithms for script learning.

Future work should address different weighting schemes for embeddings, an incorporation of linguistic knowledge into the currently rather rigid representations, for example, including sentiment polarities, or experiment with other convolutional modules to represent the joint package of story sentences. It has been demonstrated already that more elaborate recurrent architectures can yield an additional boost in performance on the *ROCStories* data set. Similarly, Chapter 4 already presented an extension of the methodology applied here in which we outlined a more sophisticated model with an attention mechanism for relation recognition. We believe that attention and related external semantic memories can be a convenient and promising extension of the work presented here. Especially, regarding interpretability of the decisions made by such a system, the weightings provide pointed insight and can be a key component towards handling such complex semantic tasks—for the purpose of a deeper story understanding.

Software: An implementation of the methodology outlined in this chapter is publicly available from the following URL: `http://www.acoli.informatik.uni-frankfurt.de/resources.html`.

Summary

This part of the dissertation has aimed at crossing the bridge between the previous two large parts on implicit discourse parsing (Part II) and implicit semantic role labeling (Part III). Although the dimension of the former has for the most part been treated as a global, i.e. cross-sentential phenomenon, and the latter as a rather local one, a clear line separating the two could not be drawn so easily, however. Obviously, both tasks share local and global aspects at the same time: for example, implicit roles are evoked locally within the immediate syntactic context of the predicate, and can possibly (but must not necessarily) be resolved globally in the discourse context. Also, implicit causality verbs evoke the expectation towards a certain discourse relation on the word or phrase level, having an immediate effect on the specific type of sentences which follow. We have considered this specific interaction as a *bottom-up* effect—from the word level to the discourse level. Yet, we have also seen that this interaction can be approached from the reverse direction in a *top-down* fashion; for instance, in a speaker vs. hearer context, whenever a discourse relation is not uniquely predictable, an explicit (word-level) marker has to be inserted in order to disambiguate the type of relation. Another example for a top-down interaction can be seen in coreferential sentences pairs, such as entity relations, whose discourse coherence and sense relation can best be explained by the choice of referring expressions and anaphoricity on the word level within the arguments. In order to assess the mutual influence of both top-down and bottom-up effects, we have outlined specially tailored experiments in two Chapters.

Chapter 8 targeted at extending within sentence semantics to the discourse. To this end, we have introduced a correlation study whose setup has been two-fold. In a first step, a background knowledge base was acquired which contains the realizations of explicit predicate-role constellations as they naturally occur in large corpora. Based on the frequencies in the knowledge base, we have computed the probability of occurrence of a specific implicit role in a given predicate context. We found that with the aggregation of local implicit SRL information in argument pairs there seems to exist a synergy with global discourse coherence in the form of a strong mutual association between implicit roles, on the one hand, and semantically similar discourse senses, on the other. For example, we have observed that implicit causality (modifier) roles contribute mostly to implicit causality relations, as annotated in the Penn Discourse Treebank. This correlation study is—to the best of our knowledge—the first of its kind to assess the bottom-up contribution of local iSRL information on the senses of superordinate implicit discourse structure. As pointed out already, some of the collected predicate patterns simply occur too infrequent in the automatically annotated corpora for reliable predictions regarding implicit roles and the preceding or following discourse structures in which the predicates are embedded. Since the overall effects in the correlation with discourse senses (which we measured in aggregated probabilities over both arguments) were evident, yet rather weak, we conjecture that specific weightings could be useful to strengthen the effect of the interaction. For example, we argue that not all predicates are equally important in the assessment of global discourse coherence: auxiliary predicates could be less likely to contribute substantially to global discourse coherence as opposed to main verbs, for instance. In this connection, we propose a weighting scheme similar to the one introduced in Chapter 3 on implicit discourse arguments. Furthermore, we suggest that our correlation study could be easily reproduced in a fully automatized setting, and thus be extended to other languages, for example, to Chinese, for which a whole range of SRL and shallow discourse parsers already exist (Sun, 2010; Wang and Lan, 2016). Future research might directly address the issue of zero anaphora and investigate in closer detail which predicates are more likely to have unexpressed core agents and whether this phenomenon positively correlates with entity-based coherence relations, as one would expect given the theoretical literature on zero anaphora in Chinese.

Chapter 9 has taken up the reverse, top-down direction and focused on mapping discourse relations (esp. entity-based coherence) onto a local predicate level and their semantic arguments. For this purpose, we have introduced a novel SRL framework for the classification of predicate-argument structures beyond the sentence level. The particular the relation labels were motivated by Centering Theory, and we have demonstrated that they offer a more flexible (event-driven) account of anaphoric relationships as opposed to a single salient entity in the discourse. Technically, for this approach, we were able to recycle an existing discourse architecture—most notably without any structural modifications—by successful adaptation to the task at hand. Three experiments have shown that our model is flexible enough to reliably predict local SRL patterns (either coreferential, non-coreferential or both simultaneously), and that the learned attention weights can help in pinpointing those latent features on entity relations (in particular, predicates and pronouns) which drive the classification decisions of the model. It should be noted that this bridge experiment should be considered a proof of concept, as the classification performance of the model (which we simply ported to the new domain of coreferential SRL) was ultimately not optimal and could benefit from further parameter tuning. Also, further experimentation concerning long-distance dependencies is needed in order to account for the true nature of entity-based coherence which typically span across more than just two discourse arguments. Analogous to the first bridge experiment, we argue that the experimental setup can be straightforwardly ported to other genres or languages, as well. Automated coreference resolution systems exist for English or Chinese (Lee et al., 2013; Clark and Manning, 2016), and allow to directly encode coreferential information into the local argument realization (potentially also for non-core roles) and to establish implicit links to entity mentions from antecedents or postcedents in the discourse.

Chapter 10 has outlined a closely related experiment. It has been concerned with continuations as the most preferable coherence transition according to Centering Theory. Continuations are special insofar as they relate to aspects of both discourse continuity (discourse relations) and entity continuity (anaphoric relationships). We have focused on the former and explored ways to detect the unexpressed relationship that holds between subsequent sentences in narratives, in order to employ it as a key factor in modeling the structural properties imposed by entity-based coherence. In particular, we made use of implicit discourse relations in between adjacent sentences and applied our modeling concept to the task of narrative understanding. This task requires the automated identification of appropriate versus inappropriate story continuations. On the one hand, we have demonstrated that latent implicit discourse structure provides direct means to model text coherence; future work needs to investigate in closer detail the exact sense relations that hold. On the other hand, we have also proven that our previously introduced feedforward architecture (on implicit discourse relations) can be straightforwardly ported to the task at hand. In general, this experiment has shown that mining implicit information in free texts can be a driving force in successfully handling elaborate types of semantic downstream tasks. In this context, we want to point out that the Penn Discourse Treebank does not associate any finer-grained senses (e.g., type or subtype level) to entity relations. Broadly speaking, entity relations have always been treated as a special, self-contained class, distinct from the remainder of implicit sense relations in the PDTB.[6] Yet, as a result of this experiment, we argue that, even when two descriptions are about the same entity, this should *not* restrain us from considering additional discourse senses that could potentially hold between them, e.g., contrastive or temporal relationships. We elaborate on details in the final chapter of this thesis.

In the final part of this thesis, we conclude our work by further elaboration on particular aspects which we have not addressed so far, for instance, how our proposed methods could ideally contribute to practical applications. We discuss potential improvements of our methods, revisit entity-based coherence, and give an outlook on future research directions on the topics of implicit semantic role labeling and discourse processing.

[6] Cf. the descriptions of the shared tasks by Xue et al. (2015, 2016).

Part V

Conclusion

Chapter 11

A Review of this Thesis

This thesis has addressed phenomena surrounding implicit information in text. We have introduced a range of practical techniques for the recovery, analysis, and interrelation of textually unexpressed content as holding between sentences (Part II on implicit discourse parsing) and evoked within sentences (Part III on implicit semantic role labeling). The bridge experiments in Part IV aimed at a holistic treatment of the two types of information. In this final chapter, we revisit and re-emphasize the main contributions of this thesis. We shed light on potential improvements and discuss future research directions of our proposed techniques, also with a focus on practical applications.

11.1 Possible Improvements & Future Directions

11.1.1 Implicit Discourse Parsing

Design of the Network Architecture: A number of improvements concern the two implicit discourse parsers which we outlined in Chapters 3 and 4, respectively. In particular, the first system—the lightweight parser designed in the style of a feedforward neural network—distinguishes itself in terms of an elaborate argument composition function, featuring two aggregations: First, for each argument, we computed the sum over the average and pointwise product in the word embeddings matrix. Then, the resulting vectors for each argument were simply concatenated. Although we experimented with a variety of different configurations for the purpose of an efficient and semantically meaningful representation of the discourse units, we are aware of the fact that the present network architecture is not the only possible one. Similar performances in the shared task of Xue et al. (2016) have been achieved with slightly different composition operations; cf. Qin et al. (2016) for convolution and max pooling on the concatenation of word embeddings and part of speech vectors, or Mihaylov and Frank (2016b) on experiments with cross-argument convolution layers. We conclude that further experimentation on that front is necessary to derive more suitable representations, and thus, to achieve even better classification scores. We refer the interested reader to

Hu et al. (2015) for a further description on general concepts of sentence matching (e.g., using an *iterated* sequence of convolution and pooling operations) or to the work by Feng et al. (2015), who explored and compared a whole range of distinct deep learning architectures to represent question–answer pairs.

Parser Scope on Discourse Arguments & Integration of External Resources: Unfortunately, from an architectural point of view our system is simply a *bag-of-embeddings* implementation. This means that the order of words appearing in the two discourse arguments (and, crucially, even the order of the discourse arguments themselves), is irrelevant to the classifier during training and prediction. However, the fact that the network is insensitive to word order seems to be highly inconsistent with the way humans process and interpret discourse, namely, in a sequential manner. For that reason, we have proposed an extension of the bag-of-embeddings approach, capable of a sequential treatment of discourse units, and introduced a recurrent neural network in Chapter 4, which differs from its predecessor in terms of a more elaborate (yet, at at the same time, more complicated) network structure. Accuracies on implicit relations in the Chinese Discourse Treebank could be increased with this method.

However, we believe that there is still room for improvement. One direction worthwhile pursuing can be traced back to the original annotation of discourse units in the shallow PDTB/CDTB, where an (implicit) discourse relation consists of two flat text spans associated with a relation label. To the best of our knowledge there does not exist a prior (computational) attempt which took the discourse parsing task *beyond isolated arguments* and their fixed, span-based restriction. For example, with our proposed recurrent architecture and the convenience of the explicit start and end markers on discourse boundaries, the context to the left of the first discourse argument as well as the context to the right of the second discourse argument could be easily signaled as such, thus broadening the scope of the discourse parser in order to incorporate more, potentially meaningful and indicative, information. This procedure would then make it possible to *extend* the analysis in the parsing task across the over-restricted span boundaries imposed by the annotators of the PDTB/CDTB.[1] Concerning this matter, we suggest that tokens as represented by embeddings could be weighted according to how far apart they are from either the discourse boundary, or from other core features of a discourse relation.

We have already demonstrated in Chapter 3 that these weighting schemes are highly beneficial in pinpointing and boosting the performance of those words which are strongly associated with a specific discourse relation, for example, when we score them according to how deeply they are embedded within a tree of syntactic dependencies. The recurrent architecture from Chapter 4 has provided means to *automatically* compute such weightings and to determine the relative contribution of individual tokens by the attention mechanism. We believe, how-

[1] Note that a proportion of implicit discourse relations in the PDTB and CDTB are non-adjacent. It remains an open question for future research how the intermediate content could be integrated as part of a supplementation for the discourse argument representation.

ever, that a direct, manual supervision in terms of a carefully chosen (explicit) integration of external semantic resources can be of great value to the parsing task at hand. It remains a fruitful research direction to pursue how other semantic parsers and knowledge sources (beyond dependencies) could support the weighting of individual embeddings. For instance, the Supersense tagger (Ciaramita and Altun, 2006) or WordNet senses (Miller, 1995) might be consulted to strengthen the effect of individual words or phrases, supplementing our neural network setting for the task of implicit discourse parsing.

Unsupervised Data Acquisition of Implicit Discourse Relations: As in many other areas of NLP, the scarce amount of (manually) annotated training data poses a serious limitation to the task of implicit discourse parsing, especially for those exotic sense relations with only a handful of annotated instances, which are simply too infrequent to obtain any reliable statistical generalizations.[2] It has been shown that artificially created training instances (for instance, by removing the connective from an explicit relation) in fact decrease parsing performances, because these relations cannot capture the natural semantics of an implicit discourse relation. Various methods have been proposed to augment or support the amount of available training data of implicit discourse relations. Very recently, Wu et al. (2017) proposed a technique based on co-training to select useful features and instances from artificially constructed implicit relations. Earlier, Wang et al. (2012) introduced a selection criterion to collect typical (discarding atypical) training examples by means of a bootstrapping method. Rutherford and Xue (2015) infer implicit relations by assessing the optionality of a discourse connective in a distant supervision approach, and Ji et al. (2015) tackle the problem by domain adaptation techniques between explicit and implicit relations.

Only very few approaches for data augmentation based on *parallel corpora* have been suggested, e.g., by Hong et al. (2014). We do, however, see great potential in this form of unsupervised data acquisition because existing resources can be directly exploited for that purpose. For example, bilingual sentence pairs can be obtained from corpora from the domain of machine translation, for instance from Europarl (Koehn, 2005). These comparable sentences could then serve to extract implicit discourse relations as follows: We conjecture that not every word in the sentence of the source language will be found as a literal translation in the target language, and this observation will also apply to discourse connectives. Provided that in either source or target language a discourse relation is signaled by means of an explicit connective, but in the other it is not, then the sentence pair (or two inter-sentential clauses) could be extracted as an instance of an implicit discourse relation in the respective language.[3] Not only will this technique allow us to obtain a seed set of implicit relations for languages for which currently no manually annotated resources exist, but it also serves as a distant supervision

[2] For example, the rare implicit relations EXPANSION:Exception (0.05%) or TEMPORAL:Asynchronous:Succession (3.12%), cf. the distribution in Section 2.1.4 for an overview.

[3] The explicit connective will serve either as a generalized sense label, or a manually predefined mapping could be consulted, e.g., from *aber* (*but*) to "Contrast".

signal to obtain better feature representations. Following these lines of thought, a couple of very promising techniques to infer such evidence from mono- and multilingual alignments of comparable texts have been proposed already for the domain of implicit semantic role labeling, cf. Roth and Frank (2013) or Sikos et al. (2016), which will be the focus of our next subsection.

11.1.2 Implicit Semantic Role Labeling

Relaxing the Assumptions from the Theoretical Literature: The theoretical literature imposes several restrictions on the identification and resolution of null complements (Ruppenhofer, 2005; Scott, 2006; Németh and Bibok, 2010). For example, we oftentimes find the distinction between definite, indefinite, and constructionally licensed null instantiations (DNIs vs. INIs, and CNIs), as well as a differentiation regarding core and non-core roles; roughly speaking, a categorization of implicit arguments into those which can be resolved in the context, and those which need not necessarily be linked to an antecedent. The developers of practical iSRL tools had a hard time tuning their systems to meet the linguistic requirements of these idiosyncrasies on a standard evaluation set of fiction texts, which contained this distinction as a core basis of the annotations, cf. Ruppenhofer et al. (2010). One of the reasons for the poor performances of the systems can be seen in the fact that the contexts varied widely in which specific DNIs, INIs and their associated predicates appeared, and predicate instances of similar patterns were simply too infrequent to obtain any useful generalizations, e.g., with supervised machine learning (Chen et al., 2010).

Since, even in the theoretical literature this differentiation between existential interpretations and the core and non-core character in specific cases is oftentimes controversial, we argue that, in general, for the purpose of a *practical application*, such a fine-grained adjustment to the different types of unrealized roles is not necessary. In fact, a functional system capable of recognizing any correct association between a predicate and a filler in the non-local context provides added value to conventional information extraction. We claim that in order to accomplish this goal, large-scale generalizations (similar to the approaches outlined in Chapters 6 and 7) can be useful, primarily including *all* roles, i.e. core and non-core roles, assuming all roles are resolvable per default. In what follows, we elaborate on a motivating example.

Supporting Practical Applications with Implicit Roles: User generated content and, in particular, crowd-sourced reviews, e.g., for products, restaurants or vacation trips are produced in massive amounts on the internet every day and provably provide a core source of information for many people to rely on. For instance, on a travel website, users typically add a short description about their personal experiences they had once they arrived at a certain location. In this context, we see great potential in the analysis of implicit semantic roles.

First, iSRL can assist future travelers in question answering: In order to accurately assess the question *How to get from Hong Kong to Macau?* we propose to learn a statistical model from all sentences in the holiday reviews paired with their explicit SRL structure (by simply running a parser to obtain the annotations). Ideally, those instances should be extracted which feature the ARRIVING frame (Baker et al., 1998) and explicitly realize fillers for both the *goal* (core) and the *source* (non-core) frame element by *Macau* and *Hong Kong*, respectively. Sometimes, we will observe that another non-core role, MODE_OF_TRANSPORTATION (MoT) is filled. Explicit realizations can be of various types[4] and the most convenient way of transportation can be simply provided as an answer to the user by generalization over these fillers, either by majority vote or by semantic aggregations as proposed in Chapter 7. Note that, in this context, without the inspection of non-core roles this question answering task could not be solved as easily. Also note that, related questions, such as *What is the best way to get from Hong Kong to Macau?* would rely on other non-core frame elements, e.g., MEANS. A list of highly relevant target frames to the topic at hand (e.g., ARRIVING, TRAVEL, EMOTION) could either be predefined or automatically acquired.

Second, since users of these platforms are generally unrestricted in what they can write, the MoT role, for instance, will not always be filled in tour descriptions. Again, based on the acquired information obtained through iSRL statistics, an intelligent review platform could *detect* that an argument is unrealized, and *prompt* the user to add related information to his description. An idealized prompt in this context could have the form: *You traveled from Hong Kong to Macau. How did you go there?* Since not all reviews are informative to the same extent, this could greatly increase the overall value of individual texts and other readers will profit.

Another stylistic key element employed by various writers of reviews is the elaboration on a specific subject. On a related note, we address this specific implicit discourse relation in the next section.

11.1.3 Interplay between (Implicit) Discourse Structure and Implicit Semantic Roles

Towards an Account for Entity-based Coherence Relations: The coherence relation ELABORATION is a very prominent one and at the same time the most frequent discourse relation in the Rhetorical Structure Theory Discourse Treebank (Carlson et al., 2002), accounting for almost 30% of all relation instances. The discourse relation is defined as presenting additional detail on a subject matter.[5] It has been argued by Knott et al. (2001) that ELABORATION is in fact semantically very similar to *entity-based coherence*. To be more precise, Knott et al. (2001) claim that ELABORATION is an *identity relation between entities* because of its specific properties, which make the relation unique and distinct from other relations. The authors argue that, for instance, in causal or explanation relations no subcomponents are identifiable, therefore these relations primarily hold between propositions. This is, however, not the case for ELABORATION, for which component elements within

[4] For example, *by ship, car, a ferry boat, third-party logistic providers, bus, plane, water*, etc.

[5] Cf. `http://www.sfu.ca/rst/01intro/definitions.html`

those propositions—namely the entities themselves—can be identified.

In the Penn Discourse Treebank (Prasad et al., 2008), entity-based coherence relations, i.e. EntRels, make up roughly 13% of all relations. Note, however, that in Chapter 8 we have already pointed out that EntRels are seriously underrepresented in the PDTB because they are only annotated between adjacent text spans. Crucially, they are defined as an *implicit* relation with entities encountered in the two discourse units whose reference is either realized directly, or indirectly.[6] As an illustration, consider the Example (38). The entity is introduced in the first discourse argument (*he, the London stockbroker*) and mentioned again in the second argument by direct pronominal reference at the beginning of the sentence.

(38) Arg1: Last summer, he chucked his 10-year career as a London stockbroker and headed for the mountains.

Arg2: He didn't stop until he got to Jackson Hole, Wyo.

Implicit discourse sense: EntRel[7]

Almost identical patterns of this form have been the focus of extensive psycholinguistic studies by Rohde et al. (2007), Rohde and Horton (2010), and Kehler and Rohde (2017), respectively, who were particularly interested in how pronoun interpretation (in the second discourse argument) is affected by discourse coherence. A prototypical example referred to by the authors is shown in (39).

(39) John$_{\text{SOURCE}}$ handed a book to Bob$_{\text{GOAL}}$. He ______ .

With examples like the one in (39), participants were tested on how they would interpret the ambiguous pronoun. In a nutshell, the authors found that when the pronoun is interpreted as referring to the source thematic role (*John* in this case), this went along with greater expectations towards an *explanation* (or elaboration) in the ongoing discourse. An opposite trend was visible for discourse relations of *occasion* or *result*, which elicited a goal-preferring interpretation (*Bob* in the example).[8] Furthermore, Rohde and Horton (2010) found that certain verb types (of implicit causality, e.g., *scold*) tend to make comprehenders expect explanations in the subsequent discourse context, and demonstrated that also object properties (e.g., normal vs. abnormal nouns) have an influence on the expectation of follow-up sentences, cf. Rohde et al. (2007).

[6] Cf. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf, p. 23, accessed November 2017, and the overview in Section 2.1.4 of this dissertation.

[7] Document ID wsj_0776 of PDTB training section.

[8] More precisely, under the so-called Question Under Discussion model of discourse interpretation (von Stutterheim and Klein, 1989; Kuppevelt, 1995; Benz and Jasinskaja, 2017), significantly more *why*-questions were evoked when the pronoun was interpreted as the source role. Goal-preferring interpretations are typically associated with an end state (e.g., with transfer of possession verbs) and more commonly gave rise to questions asking *what will happen next?*

To summarize, there exists a noticeable and quantitatively measurable association between different relations of discourse coherence and the way how pronouns as part of entity-based coherence relations are interpreted. Note, however, that the PDTB does *not* associate any of these finer-grained relation senses to their `EntRels`, i.e. entity-relations are only marked as such, even if they have provably supplemental discourse information encoded. In fact, we argue that—in addition to entity coherence—the previous Example (38) captures aspects of temporal information as well. One possible discourse interpretation would postulate that *after* the stockbroker stopped his career, he went to Wyoming (TEMPORAL:Asynchronous:precedence in PDTB notation).[9]

Consider again another motivating example of an `EntRel` from the PDTB.

(40) `Arg1`: Jerome J. Jahn, executive vice president and chief financial officer, said Mr. Rubendall was resigning by "mutual agreement" with the board.

`Arg2`: "He is going to pursue other interests," Mr. Jahn said.

Implicit discourse sense: `EntRel`[10]

In Example (40), two entities are referenced across the two discourse arguments, i.e. *Jahn/Jahn and Rubendall/he*. Given the second reading, we argue once more that a comphrehender would anticipate additional discourse features on top of plain entity-based coherence. In the concrete example—and independent of its idiomatic use—"other interests" are most likely attributable to a *reason* or *result* of resigning.

As pointed out before, Rohde et al. (2007) and Kehler and Rohde (2017) have already successfully demonstrated that discourse coherence can be traced back to local cues responsible for the interpretation, e.g., certain verbs in combination with source and goal thematic roles, and the particular manifestation of nominal objects. However, it should be noted that many naturally occurring entity relations, such as the ones in (38) or (40), do not match the artificially created stimuli involved in these psycholinguistic experiments. The issue that pronoun patterns are idealized in this setting has also been pointed out by Asr and Demberg (2015). In fact, sentences collected from corpora are commonly far more elaborate and to the same extent more complex in their syntactic realization (e.g., including a mixture of direct and indirect reported speech). They usually come with a *multitude of distinct thematic role realizations*, including core and non-core roles, which naturally differ from sentence to sentence. Even though psycholinguistic settings in general are well-suited to test single factors in isolation (e.g., a

[9] Another interpretation of the second discourse argument could be a default one of type elaboration (EXPANSION:Conjunction), or probably both events happened at the same time (TEMPORAL:Synchronous).

[10] Document ID `wsj_0229` of PDTB training section.

bias towards the interpretation of one particular semantic role), they are unfortunately too restricted to properly account for a larger range of local triggers and marked discourse phenomena. We therefore argue that in order to explain *all possible instantiations* of entity-based coherence (as well as other discourse) relations, we need a model which can account for *all thematic role realizations in combination with any given predicate*—ideally also incorporating *implicit semantic roles*. Gerber (2011) effectively concludes his famous work on implicit roles by motivating the urgent need for *joint models of semantic arguments* that are required for true human sentence comprehension both at the sentence *and* discourse level.

We have already demonstrated in Chapter 8 that the establishment of such a model is feasible: Based on a large-scale generalization, we have computed for a particular predicate-role configuration the likelihood of other *unexpressed* roles occurring in this context. We have shown that when such an implicit role is highly expected, then its properties are very similar to the discourse structure in which the predicate is embedded. We have, for instance, computed for *not applauding* a significantly higher probability of a causal role/relation occurring, as opposed to only *applaud*. In fact, the exact opposite effect was measurable for the predicate *resign*. This technique allows us to also partly explain the entity relation in (40), where the subsequent sentence does in fact provide evidence for causality information on a reason or result.

Finally, it should be noted that a large proportion (≈20%) of the annotated `EntRels` in the PDTB *cannot* be explained by coreferentiality, i.e. no coreferential element pair can be found in either of the two discourse arguments (cf. the discussion Section 8.3.4 for details). Again, we conclude that the driving force to account for these relations lies in a deeper form of semantics. As a matter of fact, Roth (2017) has very recently shown that semantic roles are indeed a core indicator for modeling discourse coherence in a machine learning setting. For future work, we see fruitful possibilities in integrating implicit semantic roles here as well. As a practical use case, we describe a slight extension of our second bridge experiment outlined in Chapter 9, which we outline hereafter.

Extending Cross-Argument SRL to Uninstantiated Arguments: In Chapter 9, we have performed an experiment to ground entity-based discourse coherence on the level of semantic roles. The approach was inspired by Centering Theory (Grosz et al., 1995) and the four transition types of continuations, retentions, and (two variants of) shifts were indirectly encoded into the local predicate-argument structure as a result of linking those roles that are coreferential in both discourse arguments. We demonstrated that our approach using event-labels is computationally feasible and particularly attractive because it offers more flexibility in a finer-grained assessment of local coherence as opposed to a single salient entity as reference point in the backward-looking center. However, one specific aspect that we have not dealt with so far concerns the treatment of locally unexpressed roles and how they should be encoded into the cross-argument labels.

As an illustration, consider the following Example (41) from Yeh and Chen (2001) (and their original translations) in which the second discourse argument lacks the presence of a core role, represented by the zero anaphor (∅/*Electronig stocks*).

(41) Arg1: [電子股$_{\texttt{impl-A1}}$] 受 美國高科技股 重挫 影響， *Electronics stocks were affected by high-tech stocks fallen heavily in America.*

Arg2: [∅$_{\texttt{A1}}$] 今日 持續 下跌； [*Electronics stocks*] *continued falling down today.*

Implicit discourse sense: EntRel
Label: $\texttt{0}_{\texttt{i}}\texttt{1}_{\texttt{i}}$XXX (Continue)

According to the PropBank role set, the zero anaphor in the second discourse argument is part of the argument realization of the predicate *continue.01* and corresponds to an unexpressed proto-patient role, i.e. A1, the *thing continuing*.[11] Note also, that the *causer of the continuation*, A0 is an implicit core role whose antecedent can be found in the first discourse argument (*US high-tech stocks*). Without access to implicit links in the second discourse argument, i.e. neither to the zero anaphor, nor to the classical implicit role, the label of this local coherence relation would have the form XXXXX, meaning that no overt semantic roles besides the predicate are present.[12] However, assuming that we had direct access to these implicit relations, we could directly encode this information into the resulting label of the form $\texttt{0}_{\texttt{i}}\texttt{1}_{\texttt{i}}$XXX, where the zero anaphor in Arg2 is coreferential with the thing affected (A1) in Arg1, and the unexpressed causer of the continuation in Arg2 can be linked to the thing affecting (A0) in Arg1. Analogous to the principles formulated in Centering Theory, the exact form of the pattern allows us to deduce that the entity relation roughly corresponds to a continuation transition.

With the aid of our proposed formalism it becomes straightforward to bring together and bundle related outputs from various NLP tools, which can ultimately enhance the resolution process of implicit relations: Current state-of-the-art coreference resolution systems for Chinese (Lee et al., 2013; Clark and Manning, 2016) can serve the function of linking role indices. Crucially, these systems do not resolve anaphoric zero pronouns, yet, in their study, Yeh and Chen (2001) implemented a rule-based system for the resolution of zero anaphora based on Centering Theory, whose output can be directly integrated into our cross-argument labels. More advanced techniques for zero anaphora resolution exist, e.g., proposed by Chen and Ng (2013, 2016). In the latter approach, the authors have recently

[11] Only for illustration purposes, we focus on the predicates in the English translation instead of the original source text, cf. http://verbs.colorado.edu/propbank/framesets-english-aliases/continue.html and http://verbs.colorado.edu/propbank/framesets-english-aliases/affect.html.

[12] Details on how we derived the exact form of the labels can be found in the original description in Chapter 9. Note also that the definition of the online frame set for *continue.01* would classify *falling down* as a second realization of A1, but we omit this case to avoid unnecessary complications for the sake of this example.

described the first neural network parser for zero pronoun resolution in Chinese. However, from a technical point of view, the model is a feedforward architecture which selects candidate constituents from the context, and we believe that the procedure could be structurally supplemented by a sequential access (in the form of a recurrent neural network) which we proposed in Chapter 9. This way, our proposed model would directly perform Chinese zero anaphora resolution on the SRL level, allowing for the harmonization of a majority of related phenomena into one joint representation scheme, i.e. implicit discourse relations (entity-based coherence), zero anaphora (implicit semantic roles), explicit semantic roles, and coreferentiality.

11.2 Concluding Remarks

Both implicit semantic role labeling and implicit discourse parsing will soon move out of the experimental environment and find their way into the applications of everyday life. Intelligent speech assistants, for example, will be enjoying even greater popularity in the next few years and it is no surprise that many of the biggest and most influential software companies want to have a share in the great success story. According to Apple CEO Tim Cook, already at the end of 2016 more than two billion weekly requests were made by users of Apple's famous speech assistant Siri.[13] As a result, enormous amounts of spoken language data are generated and collected every day, and data scientists, computational linguists, and machine learning experts collaboratively seek ways to handle and exploit these valuable sources by optimization of the neural network machinery that operates on top of big data, not only to enhance the voice recognition rate in speech-to-text systems, but—even more significantly—to improve the practical communication skills of the speech assistants with the human conversation partner. It is incontestable that Siri, Alexa, and their friends will become increasingly intelligent within the next few years, and I believe that in the distant future communication with machines, most notably, in spoken dialog systems will slowly but surely approach the quality of human-level conversations.

Unfortunately, such a scenario is still far from being a reality. We know that *Artificial Intelligence* has become a buzzword, in fact, a catchall term to denote in general *any* type of application that incorporates statistical models or machine learning in some way or other. What users of speech assistants at present pretty much can do is to ask for the local time and weather in Tokyo, to request the fastest route to the airport, or to order a pizza with their favorite topping.[14] I agree that these simple applications are indeed impressive to a certain extent because they provide facilitation and automation of both information transfer

[13] http://www.businessinsider.de/apple-q4-2016-earnings-2016-10?r=US&IR=T, all links in this section were accessed December 2017.

[14] For a comparison of different devices and use cases, cf. https://www.androidauthority.com/google-assistant-vs-siri-vs-bixby-vs-amazon-alexa-vs-cortana-best-virtual-assistant-showdown-796205/

and acquisition, yet they are far from truly (artificially) intelligent—at least in their current state.

However, a recent development that we can clearly observe is a rapid increase in the development and use of such voice-based communication devices whose overall number has been growing enormously over the last few years.[15] Compared to 33 million voice-enabled devices in 2017, current estimates predict 2.5 billion devices by the year 2021.[16] In this context, I want to emphasize that, in the foreseeable future, communication with machines *in natural language* will become an essential part of our life. I conjecture that almost every domain will sooner or later be accompanied by its own speech-based assistant, for instance, in the areas of professional life to structure business meetings and synchronize events[17], sporting activities, for school education and teaching purposes, but also in the household (smart home management)[18], in the medical context, as shopping assistants[19], or in the automotive area, for example, to conduct autonomous driving and, of course, for navigation, which has already become a standard in many cars nowadays already.

Based on the latest developments, we can forecast that the mode of interaction with these systems will be highly personalized. The programs will consider and make use of user-specific background information on previous interactions with the system, including meta data also from other textual sources, e.g., from personal emails or instant messages with friends and colleagues, and speech assistants will ultimately self-teach, learn, and improve themselves based on the interaction with the user in natural dialogues. This claim is based on the fact that rich media and user-generated content have recently led to a data explosion as for many people voice- and text-based information exchange—most notably through instant messaging apps—has become an integral part of their lives, and it is fair to say that this type of technology has revolutionized communication.[20] It is incredible how highs and lows of people's lives are meticulously documented on the basis of massive amounts of instant messages that are sent around the globe every single second. On the basis of this simple observation, it is important to bear

[15] And the use of voice-based communication in general, for example, in the form of voice- or audio messages exchanged between users on mobile instant messaging platforms, such as WhatsApp or the Chinese messaging app WeChat; cf. https://www.macrumors.com/2017/11/28/whatsapp-locked-audio-recording-update/, https://chinachannel.co/2016-wechat-data-report/

[16] https://medium.com/snips-ai/how-we-are-solving-the-biggest-issue-of-conversational-assistants-data-f34600048e80, http://voicelabs.co/2017/01/15/the-2017-voice-report/

[17] https://www.fastcompany.com/40502346/alexa-for-business-puts-amazons-voice-assistant-to-work

[18] https://www.cnet.com/news/talk-to-your-house-with-these-voice-activated-smart-home-systems/

[19] https://techcrunch.com/2017/08/22/walmart-and-google-partner-on-voice-based-shopping/

[20] Cf. https://www.theguardian.com/technology/2016/jul/03/from-political-coups-to-family-feuds-how-whatsapp-became-our-favourite-way-to-chat, https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users/

in mind what it means when this type of data will sooner or later be processed semantically by intelligent algorithms which can make sense of the (currently still unstructured) content.

Moreover, as speech assistants become more sophisticated, so will also increase the complexity of the interaction and the possible ways of communication. Crucially, the systems will soon be able to successfully process inputs that go beyond isolated commands and simple requests (*When is today's meeting?*), and ultimately more elaborate questions will be possible (*Why did my boss cancel today's meeting?*). What sounds utopian (and still philosophical) today, could become reality in only a few decades: Users will be able to engage in longer conversations with the systems—first on the basis of a controlled vocabulary, restricted to a specific genre or domain, later on in a more general way. It is already apparent today, that in a few decades from now—due to the demographic development that we are witnessing—an adequate care for the elderly will represent an enormous challenge to our society. Japan, for example, is already preparing for that situation in an exemplary manner.[21] It will surely not appeal to everyone, but, remarkably, so-called "carebots" will serve the purpose of assisting elderly people, not only physically, but in the near future also emotionally. As people will always feel the need for someone to talk to and to confide in somebody, especially when there is nobody else to turn to, sharing personal stories with humanoid robots, and making these robots *understand* and *react* to the stories in an appropriate manner, will be the vision for the future that both Natural Language Understanding and Artificial Intelligence will have to solve.

To summarize, story understanding, in particular tracking entities and events in a coherent text, and switching from simple *when* to more elaborate *why* questions in the interaction with a speech assistant will require more sophisticated techniques, for which exemplary approaches were outlined in this thesis. I am convinced that, in order to realize the vision of truly intelligent systems, implicit information extraction, the recovery of implicit links, and semantic associations will be essential in this context, providing the core basis for advanced inference capabilities and true Natural Language Understanding.

[21] `http://www.businessinsider.com/japan-developing-carebots-for-elderly-care-2015-11?IR=T`

Bibliography

Afantenos, S. D. and Asher, N. (2010). Testing SDRT's Right Frontier. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, pages 1–9, Beijing, China.

Afantenos, S. D., Kow, E., Asher, N., and Perret, J. (2015). Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 928–937, Lisbon, Portugal.

Allen, J. (1995). *Natural Language Understanding*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 2 edition.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

Asr, F. T. and Demberg, V. (2012). Implicitness of Discourse Relations. In *Proceedings of COLING 2012*, pages 2669–2684, Mumbai, India. The COLING 2012 Organizing Committee.

Asr, F. T. and Demberg, V. (2015). Uniform Information Density at the Level of Discourse Relations: Negation Markers and Discourse Connective Omission. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, pages 118–128, London, UK.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Beliga, S., Meštrović, A., and Martinčić-Ipšić, S. (2015). An Overview of Graph-Based Keyword Extraction Methods and Approaches. *Journal of Information and Organizational Sciences*, 39(1):1–20.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155.

Benz, A. and Jasinskaja, K. (2017). Questions Under Discussion: From Sentence to Discourse. *Discourse Processes*, 54(3):177–186.

Bhat, R. A., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A., Ramagurumurthy Vishnu, S., and Xia, F. (2017). *The Hindi/Urdu Treebank Project*, pages 659–697. Springer Netherlands, Dordrecht.

Biran, O. and McKeown, K. (2013). Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Volume 2: Short Papers*, pages 69–73, Sofia, Bulgaria.

Björkelund, A., Hafdell, L., and Nugues, P. (2009). Multilingual Semantic Role Labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado. Association for Computational Linguistics.

Blair-Goldensohn, S., McKeown, K., and Rambow, O. (2007). Building and Refining Rhetorical-Semantic Relation Models. In Sidner, C. L., Schultz, T., Stone, M., and Zhai, C., editors, *HLT-NAACL*, pages 428–435. The Association for Computational Linguistics.

Braud, C. and Denis, P. (2015). Comparing Word Representations for Implicit Discourse Relation Classification. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y., editors, *EMNLP*, pages 2201–2211. The Association for Computational Linguistics.

Braud, C. and Denis, P. (2016). Learning Connective-based Word Representations for Implicit Discourse Relation Identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 203–213.

Bresnan, J. and Kaplan, R. M. (1982). Introduction: Grammars as Mental Representations of Language. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages xvii–lii. MIT Press, Cambridge, MA.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based N-gram Models of Natural Language. *Comput. Linguist.*, 18(4):467–479.

Bruce, B. and Moser, M. G. (1992). Grammar, case. In Shapiro, S. C., editor, *Encyclopedia of artificial intelligence*, page 563–570. John Wiley & Sons, New York, 2 edition.

Candito, M., Amsili, P., Barque, L., Benamara, F., Chalendar, G. D., Djemaa, M., Haas, P., Huyghe, R., Mathieu, Y. Y., Muller, P., Sagot, B., and Vieu, L. (2014). Developing a French FrameNet: Methodology and First results. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Carlson, L., Marcu, D., and Okurowski, M. E. (2002). RST Discourse Treebank. LDC Catalog No.: LDC2002T07, ISBN, 1-58563-223-6.

Carreras, X. and Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 152–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Carston, R. (2006). Code and Inference: The Meaning of Words in Context. In Fabricius-Hansen, C., Behrens, B., and Krave, M. F., editors, *Pre-Proceedings of the SPRIK Conference: Explicit and Implicit Information in Text - Information Structure across Languages*, pages 3–6, Oslo, Norway. University of Oslo.

Chambers, N. and Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 789–797.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, C. and Ng, V. (2013). Chinese Zero Pronoun Resolution: Some Recent Advances. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1360–1365.

Chen, C. and Ng, V. (2016). Chinese Zero Pronoun Resolution with Deep Neural Networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Chen, D. and Manning, C. (2014). A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Chen, D., Schneider, N., Das, D., and Smith, N. A. (2010). Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden. Association for Computational Linguistics.

Chen, J., Zhang, Q., Liu, P., Qiu, X., and Huang, X. (2016). Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Chiarcos, C. (2012). Towards the Unsupervised Acquisition of Discourse Relations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 213–217, Jeju Island, Korea. Association for Computational Linguistics.

Chiarcos, C. and Schenk, N. (2015a). A Minimalist Approach to Shallow Discourse Parsing and Implicit Relation Recognition. In *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 42–49.

Chiarcos, C. and Schenk, N. (2015b). Memory-Based Acquisition of Argument Structures and its Application to Implicit Role Detection. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 178–187.

Chiarcos, C. and Schenk, N. (2018). The ACoLi CoNLL Libraries: Beyond Tab-Separated Values. In *11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Cianflone, A. and Kosseim, L. (2018). Attention for Implicit Discourse Relation Recognition. In *11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ciaramita, M. and Altun, Y. (2006). Broad-coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 594–602, Sydney, Australia. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (2016). Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2256–2262.

Collobert, R. and Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, Helsinki, Finland. ACM.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011a). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. (2011b). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.*, 20(3):273–297.

Croft, W. and Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Daelemans, W. and van den Bosch, A. (2009). *Memory-Based Language Processing*. Cambridge University Press, New York, NY, USA, 1st edition.

Dahl, D. A., Palmer, M. S., and Passonneau, R. J. (1987). Nominalizations in PUNDIT. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 131–139, Stanford, California, USA. Association for Computational Linguistics.

Danescu-Niculescu-Mizil, C., Lee, L., and Ducott, R. (2009). Without a 'Doubt'?: Unsupervised Discovery of Downward-entailing Operators. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 137–145, Boulder, Colorado. Association for Computational Linguistics.

Das, D., Chen, D., Martins, A. F. T., Schneider, N., and Smith, N. A. (2014). Frame-Semantic Parsing. *Computational Linguistics*, 40:1:9–56.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). Probabilistic Frame-semantic Parsing. In *Human Language Technologies: The 2010 Annual Conference* *of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 948–956, Los Angeles, California. Association for Computational Linguistics.

Daume III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Davoodi, E. and Kosseim, L. (2016). On the Contribution of Discourse Structure on Text Complexity Assessment. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 166–174.

Do, Q. N. T., Bethard, S., and Moens, M. (2017). Improving Implicit Semantic Role Labeling by Predicting Semantic Frame Arguments. *CoRR*, abs/1704.02709.

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67:547–619.

duVerle, D. A. and Prendinger, H. (2009). A Novel Discourse Parser Based on Support Vector Machine Classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL 2009, pages 665–673, Suntec, Singapore. Association for Computational Linguistics.

Feizabadi, P. S. and Padó, S. (2015). Combining Seemingly Incompatible Corpora for Implicit Semantic Role Labeling. In *Proceedings of STARSEM*, pages 40–50, Denver, Colorado.

Feng, M., Xiang, B., Glass, M. R., Wang, L., and Zhou, B. (2015). Applying Deep Learning to Answer Selection: A Study and An Open Task. *CoRR*, abs/1508.01585.

Feng, V. W. and Hirst, G. (2012). Text-level Discourse Parsing with Rich Linguistic Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 60–68, Jeju Island, Korea. Association for Computational Linguistics.

Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J. M., Schlaefer, N., and Welty, C. A. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79.

Fillmore, C. J. (1968). The Case for Case. In Bach, E. and Harms, R. T., editors, *Universals in Linguistic Theory*, pages 0–88. Holt, Rinehart and Winston, New York.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In Harnad, S., editor, *Origins and evolution of language and speech*, pages 155–202. Academy of Sciences.

Fillmore, C. J. (1986). Pragmatically Controlled Zero Anaphora. In *Proceedings of Berkeley Linguistics Society*, pages 95–107, Berkeley, CA.

Fisher, R. and Simmons, R. G. (2015). Spectral Semi-Supervised Discourse Relation Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 89–93.

Garvey, C. and Caramazza, A. (1974). Implicit Causality in Verbs. *Linguistic Inquiry*, 5(3):459–464.

Gerber, M. and Chai, J. (2012). Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Comput. Linguist.*, 38(4):755–798.

Gerber, M. and Chai, J. Y. (2010). Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 1583–1592.

Gerber, M. S. (2011). *Semantic role labeling of implicit arguments for nominal predicates*. PhD thesis, Michigan State University.

Ghosh, S., Johansson, R., and Tonelli, S. (2011). Shallow discourse parsing with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011*, pages 1071–1079.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 1–9, Prague, Czech Republic. Association for Computational Linguistics.

Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Comput. Linguist.*, 28(3):245–288.

Ginzburg, J. (2015). *The Interactive Stance: Meaning for Conversation*. Oxford University Press UK.

Giuglea, A.-M. and Moschitti, A. (2006). Semantic Role Labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 929–936, Sydney, Australia. Association for Computational Linguistics.

Givón, T. (1983). *Topic Continuity in Discourse*. John Benjamins Publishing, Amsterdam.

Givón, T. (1995). Coherence in text vs. coherence in mind. In Gernsbacher, M. A. and Givón, T., editors, *Coherence in Spontaneous Text*. John Benjamins Publishing, Amsterdam.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Gorinski, P., Ruppenhofer, J., and Sporleder, C. (2013). Towards Weakly Supervised Resolution of Null Instantiations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 119–130, Potsdam, Germany. Association for Computational Linguistics.

Graff, D. and Chen, K. (2005). Chinese Gigaword. LDC Catalog No.: LDC2003T09, ISBN, 1:58563–58230.

Graff, D. and Cieri, C. (2003). English Gigaword, LDC2003T05. Web Download, Philadelphia: Linguistic Data Consortium.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Grice, H. P. (1975). Logic and Conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, San Diego, CA.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1983). Providing a Unified Account of Definite Noun Phrases in Discourse. In *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, ACL '83, pages 44–50, Cambridge, Massachusetts. Association for Computational Linguistics.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1986). Towards a computational theory of discourse interpretation. Unpublished ms.

Grosz, B. J. and Sidner, C. L. (1986). Attention, Intentions, and the Structure of Discourse. *Comput. Linguist.*, 12(3):175–204.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Comput. Linguist.*, 21(2):203–225.

Hale, J. (2003). The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

Hartshorne, J. K. (2014). What is implicit causality? *Language, Cognition and Neuroscience*, 29(7):804–824.

Hermann, K. M., Das, D., Weston, J., and Ganchev, K. (2014). Semantic Frame Identification with Distributed Word Representations. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL-2014)*, Baltimore, Maryland.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Hernault, H., Bollegala, D., and Ishizuka, M. (2010a). A Semi-Supervised Approach to Improve Classification of Infrequent Discourse Relations Using Feature Vector Extension. In *EMNLP*.

Hernault, H., Prendinger, H., duVerle, D. A., and Ishizuka, M. (2010b). HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3):1–33.

Hickl, A. (2008). Using Discourse Commitments to Recognize Textual Entailment. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 337–344, Manchester, United Kingdom. Association for Computational Linguistics.

Higashinaka, R., Nakano, M., and Aikawa, K. (2003). Corpus-based Discourse Understanding in Spoken Dialogue Systems. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 240–247, Sapporo, Japan. Association for Computational Linguistics.

Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., and Nagata, M. (2013). Single-Document Summarization as a Tree Knapsack Problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1515–1520.

Hobbs, J. R. (1977). Pronoun resolution. *SIGART Newsletter*, 61:28.

Hobbs, J. R. (1985). On the Coherence and Structure of Discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Leland Stanford Junior University, Stanford, California.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Hong, Y., Zhu, S., Yan, W., Yao, J., Zhu, Q., and Zhou, G. (2014). *Expanding Native Training Data for Implicit Discourse Relation Classification*, pages 67–75. Springer Berlin Heidelberg, Berlin, Heidelberg.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Schiffrin, D., editor, *Meaning, Form, and Use in Context: Linguistic Applications*, pages 11–42, Washington. Georgetown University Press.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90 In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, New York, New York. Association for Computational Linguistics.

Hovy, E. H. (2006). Learning by Reading: An Experiment in Text Analysis. In Sojka, P., Kopecek, I., and Pala, K., editors, *TSD*, volume 4188 of *Lecture Notes in Computer Science*, pages 3–12. Springer.

Hu, B., Lu, Z., Li, H., and Chen, Q. (2015). Convolutional Neural Network Architectures for Matching Natural Language Sentences. *CoRR*, abs/1503.03244.

Huang, H.-H. and Chen, H.-H. (2011). Chinese Discourse Relation Recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. (2013). Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 2333–2338, San Francisco, California, USA. ACM.

Iida, R., Inui, K., and Matsumoto, Y. (2007). Zero-anaphora Resolution by Learning Rich Syntactic Pattern Features. 6(4):1:1–1:22.

Jauhar, S. K. and Hovy, E. (2017). Embedded Semantic Lexicon Induction with Joint Global and Local Optimization. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 209–219, Vancouver, Canada. Association for Computational Linguistics.

Ji, Y. and Eisenstein, J. (2014). Representation Learning for Text-level Discourse Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Ji, Y., Haffari, G., and Eisenstein, J. (2016). A Latent Variable Recurrent Neural Network for Discourse-Driven Language Models. In *Proceedings of the 2016*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 332–342, San Diego, California. Association for Computational Linguistics.

Ji, Y. and Smith, A. N. (2017). Neural Discourse Structure for Text Categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.

Ji, Y., Zhang, G., and Eisenstein, J. (2015). Closing the Gap: Domain Adaptation from Explicit to Implicit Discourse Relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2219–2224.

Jian, P., She, X., Zhang, C., Zhang, P., and Feng, J. (2016). Discourse Relation Sense Classification Systems for CoNLL-2016 Shared Task. In *Proceedings of the CoNLL-16 shared task*, pages 158–163, Berlin, Germany. Association for Computational Linguistics.

Jiang, Z. P. and Ng, H. T. (2006). Semantic Role Labeling of NomBank: A Maximum Entropy Approach. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 138–145, Sydney, Australia. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2017). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 3rd (draft) edition.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. In Groenendijk, J., Janssen, T., and Stokhof, M., editors, *Formal Methods in the Study of Language*.

Kamp, H. (1995). Discourse Representation Theory. In Verschueren, J., Östman, J.-O., and Blommaert, J., editors, *Handbook of Pragmatics*, pages 253–257. Benjamins.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Kehler, A. and Rohde, H. (2017). Evaluating an Expectation-Driven QUD Model of Discourse Interpretation. *Discourse Processes*, 54(3):219–238.

Khan, A., Salim, N., and Jaya Kumar, Y. (2015). A Framework for Multi-document Abstractive Summarization Based on Semantic Role Labelling. *Appl. Soft Comput.*, 30(C):737–747.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, pages 252–256, San Diego.

Knott, A., Oberlander, J., O'Donnell, M., Mellish, C., and Mellish, E. M. O. C. (2001). Beyond Elaboration: The Interaction of Relations and Focus in Coherent Text. In *Text Representation: Linguistic and Psycholinguistic Aspects, chapter 7*, pages 181–196, Amsterdam. John Benjamins.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Kolhatkar, V. and Taboada, M. (2017). Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada. Association for Computational Linguistics.

Kong, F., Ng, T. H., and Zhou, G. (2014). A Constituent-Based Approach to Argument Labeling with Joint Inference in Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 68–77, Doha, Qatar. Association for Computational Linguistics.

Krifka, M. (2006). Handouts zum Seminar Textkohärenz und Textbedeutung. Humboldt-Universität Berlin: Anaphorische Beziehung in Texten: Diskursreferenten, Zugänglichkeitshierarchien, Centering Theory.

Kubler, S., McDonald, R., Nivre, J., and Hirst, G. (2009). *Dependency Parsing*. Morgan and Claypool Publishers.

Kuppevelt, J. V. (1995). Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1):109–147.

Laparra, E. (2015). *Implicit semantic roles in discourse*. PhD thesis.

Laparra, E. and Rigau, G. (2012). Exploiting Explicit Annotations and Semantic Types for Implicit Argument Resolution. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012.*, Palermo, Italy. IEEE Computer Society.

Laparra, E. and Rigau, G. (2013a). ImpAr: A Deterministic Algorithm for Implicit Semantic Role Labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1180–1189, Sofia, Bulgaria. Association for Computational Linguistics.

Laparra, E. and Rigau, G. (2013b). Sources of Evidence for Implicit Argument Resolution. In *10th International Conference on Computational Semantics (IWCS'13). Pages 1-11.*

Lascarides, A. and Asher, N. (1993). Temporal Interpretation, Discourse Relations and Commonsense Entailment. *Linguistics and Philosophy*, 16(5):437–493.

Lascarides, A. and Asher, N. (2007). Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure. In Bunt, H. and Muskens, R., editors, *Computing Meaning: Volume 3*, pages 87–124. Kluwer Academic Publishers.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, 39(4):885–916.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Lei, Z., Wang, N., Li, R., and Wang, Z. (2013). Definite Null Instantiation Detection in FrameNet. *Journal of Chinese Information Processing*, 27(3):107.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*.

Levin, B. (2013). Argument Structure. In Aronoff, M., editor, *Oxford Bibliographies in Linguistics*. Oxford University Press, New York.

Levin, B. (2014). Semantic Roles. In Aronoff, M., editor, *Oxford Bibliographies in Linguistics*. Oxford University Press, New York.

Levin, B. and Hovav, M. R. (2005). *Argument realization*. Cambridge University Press.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014)*, pages 302–308, Baltimore, MD.

Levy, R. and Jaeger, T. F. (2006). Speakers Optimize Information Density Through Syntactic Reduction. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 849–856, Canada. MIT Press.

Li, J. and Jurafsky, D. (2017). Neural Net Models of Open-domain Discourse Coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Li, J., Li, R., and Hovy, E. (2014). Recursive Deep Models for Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar. Association for Computational Linguistics.

Li, J. J. and Nenkova, A. (2014). Reducing Sparsity Improves the Recognition of Implicit Discourse Relations. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, pages 199–207.

Li, R., Wu, J., Wang, Z., and Chai, Q. (2015). Implicit Role Linking on Chinese Discourse: Exploiting Explicit Roles and Frame-to-Frame Relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1263–1271.

Lin, Z., Kan, M.-Y., and Ng, H. T. (2009). Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 343–351, Singapore. Association for Computational Linguistics.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

Liu, D. and Gildea, D. (2010). Semantic Role Features for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 716–724, Beijing, China. Association for Computational Linguistics.

Liu, K. (2011). Research on Chinese FrameNet Construction and Application Technologies. *Journal of Chinese Information Processing*, 25(6):46.

Liu, Y. and Li, S. (2016). Recognizing Implicit Discourse Relations via Repeated Reading: Neural Networks with Multi-Level Attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1224–1233.

Liu, Y., Li, S., Zhang, X., and Sui, Z. (2016). Implicit Discourse Relation Classification via Multi-Task Neural Networks. *CoRR*, abs/1603.02776.

Loper, E., ting Yi, S., and Palmer, M. (2007). Combining lexical resources: Mapping between propbank and verbnet. In *In Proceedings of the 7th International Workshop on Computational Linguistics*.

Lopyrev, K. (2015). Generating News Headlines with Recurrent Neural Networks. *CoRR*, abs/1512.01712.

Louis, A., Joshi, A., and Nenkova, A. (2010). Discourse Indicators for Content Selection in Summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.

Lyu, C., Lu, Y., Ji, D., and Chen, B. (2015). Deep Learning for Textual Entailment Recognition. In *27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, Vietri sul Mare, Italy, November 9-11, 2015*, pages 154–161.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Manning, C. D. (2015a). Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4):701–707.

Manning, C. D. (2015b). NLP and Deep Learning 1: Human Language & Word Vectors. Deep Learning Summer School, Montreal 2015. `http://videolectures.net/deeplearning2015_manning_language_vectors/`.

Manning, C. D. (2015c). NLP and Deep Learning 2: Compositional Deep Learning. Deep Learning Summer School, Montreal 2015. `http://videolectures.net/deeplearning2015_manning_deep_learning/`.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland.

Marcu, D. (1999). A Decision-based Approach to Rhetorical Parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 365–372, College Park, Maryland. Association for Computational Linguistics.

Marcu, D. and Echihabi, A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Plainsboro, NJ. Association for Computational Linguistics.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330.

McKinlay, A. (2013). *Modelling Entity Instantiations*. University of Leeds (School of Computing). PhD thesis.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). Annotating Noun Argument Structure for NomBank. In

Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

Mihaylov, T. and Frank, A. (2016a). Discourse Relation Sense Classification Using Cross-argument Semantic Similarity Based on Word Embeddings. In *Proceedings of the CoNLL-16 shared task*, pages 100–107, Berlin, Germany. Association for Computational Linguistics.

Mihaylov, T. and Frank, A. (2016b). Discourse Relation Sense Classification Using Cross-argument Semantic Similarity Based on Word Embeddings. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mikolov, T., Yih, S. W.-t., and Zweig, G. (2013c). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn Discourse Treebank. In *Proceedings of the Language Resources and Evaluation Conference. Lisbon, Portugal.*

Mirza, P. and Tonelli, S. (2016). CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 64–75.

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent Models of Visual Attention. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2204–2212. Curran Associates, Inc.

Modi, A. and Titov, I. (2014). Learning Semantic Script Knowledge with Event Embeddings. In *Proceedings of the 2nd International Conference on Learning Representations (Workshop track)*, Banff, Canada.

Mooney, R. J. and DeJong, G. (1985). Learning Schemata for Natural Language Processing. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence. Los Angeles, CA, USA, August 1985*, pages 681–687.

Moor, T., Roth, M., and Frank, A. (2013). Predicate-specific annotations for implicit role binding: corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 369–375, Potsdam, Germany.

Moreda, P., Llorens, H., Saquete, E., and Palomar, M. (2011). Combining Semantic Information in Question Answering Systems. *Journal of Information Processing and Management*, 47(6):870–885.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016a). A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., and Allen, J. F. (2017). LSDSem 2017 Shared Task: The Story Cloze Test. *LSDSem 2017*, page 46.

Mostafazadeh, N., Vanderwende, L., Yih, W.-t., Kohli, P., and Allen, J. (2016b). Story Cloze Evaluator: Vector Space Representation Evaluation by Predicting What Happens Next. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 24–29, Berlin, Germany. Association for Computational Linguistics.

Mou, L., Men, R., Li, G., Xu, Y., Zhang, L., Yan, R., and Jin, Z. (2016). Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Mukherjee, S. and Bhattacharyya, P. (2012). Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 1847–1864.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher: John Benjamins Publishing Company.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate O (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376.

Nguyen, K., Tannier, X., Ferret, O., and Besançon, R. (2015). Generative Event Schema Induction with Entity Disambiguation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 188–197.

Nisioi, S., Stajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 85–91.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2015)*, pages 3–16, Cairo, Egypt. LNCS 9041, Springer.

Németh, E. T. and Bibok, K. (2010). *The Role of Data at the Semantics-pragmatics Interface*. Mouton series in pragmatics. De Gruyter Mouton.

Pacheco, L. M., Lee, I.-T., Zhang, X., Zehady, K. A., Daga, P., Jin, D., Parolia, A., and Goldwasser, D. (2016). Adapting Event Embedding for Implicit Discourse Relation Recognition. In *Proceedings of the CoNLL-16 shared task*, pages 136–142. Association for Computational Linguistics.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput. Linguist.*, 31(1):71–106.

Palmer, M. S., Dahl, D. A., Schiffman, R. J., Hirschman, L., Linebarger, M., and Dowding, J. (1986). Recovering Implicit Information. In *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, ACL '86, pages 10–19, New York, New York. Association for Computational Linguistics.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A Decomposable Attention Model for Natural Language Inference. *CoRR*, abs/1606.01933.

Park, J. and Cardie, C. (2012). Improving Implicit Discourse Relation Recognition Through Feature Set Optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 108–112, Seoul, South Korea. Association for Computational Linguistics, Association for Computational Linguistics.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1310–1318.

Patterson, G. and Kehler, A. (2013). Predicting the Presence of Discourse Connectives. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 914–923.

Peñas, A. and Hovy, E. (2010). Semantic Enrichment of Text with Background Knowledge. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR '10, pages 15–23, Los Angeles, California. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Pichotta, K. and Mooney, R. J. (2016). Using Sentence-Level LSTM Language Models for Script Inference. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL 2009, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

Pitler, E. and Nenkova, A. (2009). Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 13–16.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. K. (2008). Easily Identifiable Discourse Relations. In *COLING 2008, 22nd International Conference on Computational Linguistics, Posters Proceedings, 18-22 August 2008, Manchester, UK*, pages 87–90.

Polanyi, L. (1985). A theory of discourse structure and discourse coherence. In Elifort, W. H., Kroeber, P. D., and Peterson, K. L., editors, *Papers from the General Session of the 21st. Regional Meeting of the Chicago Linguistics Society*, pages 306–322. Chicago, IL.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5):601 – 638.

Poria, S., Cambria, E., and Gelbukh, A. F. (2015). Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis. In *Proceedings of the 2015 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pages 2539–2544.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., and Stein, B. (2014). Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., and Toms, E., editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York. Springer.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*, pages 2961–2968, Marrakech, Morocco.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. L. (2007). The Penn Discourse Treebank 2.0 annotation manual.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., and Webber, B. (2006). The Penn Discourse Treebank 1.0. annotation manual. Technical report, University of Pennsylvania Technical Report IRCS-0601.

Qin, L., Zhang, Z., and Zhao, H. (2016). Shallow Discourse Parsing Using Convolutional Neural Network. In *Proceedings of the CoNLL-16 shared task*, pages 70–77. Association for Computational Linguistics.

Qin, L., Zhang, Z., Zhao, H., Hu, Z., and Xing, E. (2017). Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Rahimtoroghi, E., Wu, J., Wang, R., Anand, P., and Walker, M. A. (2017). Modelling Protagonist Goals and Desires in First-Person Narrative. *CoRR*, abs/1708.09040.

Rahman, A. and Ng, V. (2011). Narrowing the Modeling Gap: A Cluster-Ranking Approach to Coreference Resolution. *J. Artif. Intell. Res.*, 40:469–521.

Rao, J., He, H., and Lin, J. (2016). Noise-Contrastive Estimation for Answer Selection with Deep Neural Networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1913–1916, Indianapolis, Indiana, USA. ACM.

Regneri, M., Koller, A., and Pinkal, M. (2010). Learning script knowledge from web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala.

Rohde, H. and Horton, W. (2010). Why or what next? Eye movements reveal expectations about discourse direction. Talk at the 23rd Annual CUNY Conference on Human Sentence Processing. New York, NY.

Rohde, H., Kehler, A., and Elman, J. L. (2007). Pronoun Interpretation as a Side Effect of Discourse Coherence. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 617–622, Nashville, TN.

Rönnqvist, S., Schenk, N., and Chiarcos, C. (2017). A Recurrent Neural Model with Attention for the Recognition of Chinese Implicit Discourse Relations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–262, Vancouver, Canada. Association for Computational Linguistics.

Roth, M. (2014). *Inducing Implicit Arguments via Cross-document Alignment: A Framework and its Applications*. PhD thesis, Department of Computational Linguistics, Heidelberg University.

Roth, M. (2017). Role Semantics for Better Models of Implicit Discourse Relations. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Roth, M. and Frank, A. (2013). Automatically Identifying Implicit Arguments to Improve Argument Linking and Coherence Modeling. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 306–316, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roth, M. and Lapata, M. (2016). Neural Semantic Role Labeling with Dependency Path Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Ruppenhofer, J. (2005). Regularities in Null Instantiation. Ms, University of Colorado.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Ruppenhofer, J., Gorinski, P., and Sporleder, C. (2011). In search of missing arguments: A linguistic approach. In Angelova, G., Bontcheva, K., Mitkov, R., and Nicolov, N., editors, *RANLP*, pages 331–338. RANLP 2011 Organising Committee.

Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., and Palmer, M. (2010). SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 45–50, Los Angeles, California. Association for Computational Linguistics.

Rush, A. M., Chopra, S., and Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y., editors, *EMNLP*, pages 379–389. The Association for Computational Linguistics.

Rutherford, A., Demberg, V., and Xue, N. (2017). A Systematic Study of Neural Discourse Models for Implicit Discourse Relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291, Valencia, Spain. Association for Computational Linguistics.

Rutherford, A. and Xue, N. (2014). Discovering Implicit Discourse Relations Through Brown Cluster Pair Representation and Coreference Patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden. Association for Computational Linguistics.

Rutherford, A. and Xue, N. (2015). Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado. Association for Computational Linguistics.

Rutherford, A. and Xue, N. (2016). Robust Non-Explicit Neural Discourse Parser in English and Chinese. In *Proceedings of the CoNLL-16 shared task*, pages 55–59. Association for Computational Linguistics.

Rutherford, A. T., Demberg, V., and Xue, N. (2016). Neural Network Models for Implicit Discourse Relation Classification in English and Chinese without Surface Features. *CoRR*, abs/1606.01990.

Sagae, K. (2009). Analysis of Discourse Structure with Syntactic Dependencies and Data-Driven Shift-Reduce Parsing. In *International Conference on Parsing Technologies (IWPT-09)*, Paris, France.

Saito, M., Yamamoto, K., and Sekine, S. (2006). Using Phrasal Patterns to Identify Discourse Relations. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*.

Sak, H., Senior, A., Rao, K., Beaufays, F., and Schalkwyk, J. (September 2015). Google voice search: faster and more accurate. `https://research.googleblog.`

com/2015/09/google-voice-search-faster-and-more.html. Accessed: 2016-10-11.

Sammons, M., Vydiswaran, V., and Roth, D. (2012). Recognizing Textual Entailment. In Bikel, D. M. and Zitouni, I., editors, *Multilingual Natural Language Applications: From Theory to Practice*, chapter 6, pages 209–258. IBM Press.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ.

Schenk, N. and Chiarcos, C. (2016). Unsupervised Learning of Prototypical Fillers for Implicit Semantic Role Labeling. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1473–1479.

Schenk, N. and Chiarcos, C. (2017). Resource-Lean Modeling of Coherence in Commonsense Stories. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 68–73, Valencia, Spain. Association for Computational Linguistics.

Schenk, N., Chiarcos, C., Donandt, K., Rönnqvist, S., Stepanov, E., and Riccardi, G. (2016). Do We Really Need All Those Rich Linguistic Features? A Neural Network-Based Approach to Implicit Sense Labeling. In *Proceedings of the CoNLL-16 shared task*, pages 41–49, Berlin, Germany. Association for Computational Linguistics.

Schenk, N., Chiarcos, C., and Sukhareva, M. (2015). Towards the Unsupervised Acquisition of Implicit Semantic Roles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 570–578, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Schuler, K. K. (2005). *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, Philadelphia, PA, USA. AAI3179808.

Schuster, M. and Paliwal, K. (1997). Bidirectional Recurrent Neural Networks. *Trans. Sig. Proc.*, 45(11):2673–2681.

Scott, K. (2006). When less is more: Implicit arguments and relevance theory. In *UCL Working Papers in Linguistics.*, pages 1–25.

Sha, L., Chang, B., Sui, Z., and Li, S. (2016). Reading and Thinking: Re-read LSTM Unit for Textual Entailment Recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2870–2879, Osaka, Japan. The COLING 2016 Organizing Committee.

Shen, D. and Lapata, M. (2007). Using Semantic Roles to Improve Question Answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic. Association for Computational Linguistics.

Sikos, J., Versley, Y., and Frank, A. (2016). Implicit semantic roles in a multilingual setting. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (* SEM 2016)*, pages 45–54, Berlin, Germany. *SEM.

Silberer, C. and Frank, A. (2012). Casting Implicit Role Linking as an Anaphora Resolution Task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 1–10, Montréal, Canada. Association for Computational Linguistics.

Smet, H. D. (2005). A Corpus of Late Modern English Texts. *International Computer Archive of Modern and Medieval English (ICAME)*, 29:69–82.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems 24*.

Socher, R. and Manning, C. D. Deep Learning for Natural Language Processing (without Magic). Keynote at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013). `http://nlp.stanford.edu/courses/NAACL2013/`.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, WA. Association for Computational Linguistics.

Sporleder, C. and Lascarides, A. (2008). Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, 14(3):369–416.

Stede, M. (2011). *Discourse Processing*, volume 15 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

Stepanov, E., Riccardi, G., and Bayer, O. A. (2015). The UniTN Discourse Parser in CoNLL 2015 Shared Task: Token-level Sequence Labeling with Argument-specific Models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 25–31, Beijing, China. Association for Computational Linguistics.

Stone, P. J. and Hunt, E. B. (1963). A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, Detroit, Michigan. ACM.

Subba, R. and Di Eugenio, B. (2009). An Effective Discourse Parser That Uses Rich Linguistic Information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 566–574, Boulder, Colorado. Association for Computational Linguistics.

Sukhareva, M. and Chiarcos, C. (2014). Diachronic Proximity vs. Data Sparsity in Cross-lingual Parser Projection. A Case Study on Germanic. In *COLING-2014 Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial-2014)*, Dublin, Ireland.

Sukhareva, M. and Chiarcos, C. (2016). Combining Ontologies and Neural Networks for Analyzing Historical Language Varieties. A Case Study in Middle Low German. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, Portorož, Slovenia.

Sun, M. and Chai, J. Y. (2007). Discourse processing for context question answering based on linguistic knowledge. *Knowledge-Based Systems*, 20(6):511–526.

Sun, W. (2010). Semantics-Driven Shallow Parsing for Chinese Semantic Role Labeling. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 103–108, Uppsala, Sweden. Association for Computational Linguistics.

Surdeanu, M., McClosky, D., Smith, M. R., Gusev, A., and Manning, C. D. (2011). Customizing an Information Extraction System to a New Domain. In *Workshop on Relational Models of Semantics*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112, Montreal, Canada. MIT Press.

Taboada, M. (2009). Implicit and explicit coherence relations. *Discourse, of course. Amsterdam: John Benjamins*, pages 125–138.

Taboada, M. and Mann, W. C. (2006a). Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588.

Taboada, M. and Mann, W. C. (2006b). Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Tao, L. (1996). Topic Discontinuity and Zero Anaphora in Chinese Discourse: Cognitive Strategies in Discourse Processing. pages 487–513.

Tonelli, S. and Delmonte, R. (2010). VENSES++: Adapting a Deep Semantic Processing System to the Identification of Null Instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 296–299, Uppsala, Sweden. Association for Computational Linguistics.

Tonelli, S. and Delmonte, R. (2011). Desperately Seeking Implicit Arguments in Text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 54–62, Portland, Oregon, USA. Association for Computational Linguistics.

Trandabăţ, D. (2011). Using Semantic Roles to Improve Summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG-2011)*, pages 164–169, Nancy, France.

Trivedi, R. and Eisenstein, J. (2013). Discourse Connectors for Latent Subjectivity in Sentiment Analysis. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 808–813, Atlanta, Georgia. Association for Computational Linguistics.

Turian, J., Ratinov, L.-A., and Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-Supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

UzZaman, N. and Allen, J. F. (2010). Extracting Events and Temporal Expressions from Text. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 1–8. IEEE.

Van Dijk, T. A. (1997). The study of discourse. *Discourse as structure and process*, 1:1–34.

von Stutterheim, C. and Klein, W. (1989). Referential Movement in Descriptive and Narrative Discourse. In DIETRICH, R. and GRAUMANN, C. F., editors, *Language Processing in Social Context*, volume 54 of *North-Holland Linguistic Series: Linguistic Variations*, pages 39 – 76. Elsevier.

Wang, C. and Fan, J. (2014). Medical Relation Extraction with Manifold Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 828–838.

Wang, J. and Lan, M. (2016). Two End-to-end Shallow Discourse Parsers for English and Chinese in CoNLL-2016 Shared Task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40. Association for Computational Linguistics.

Wang, X., Li, S., Li, J., and Li, W. (2012). Implicit Discourse Relation Recognition by Selecting Typical Training Examples. In *Proceedings of COLING 2012*, pages 2757–2772, Mumbai, India. The COLING 2012 Organizing Committee.

Webber, B. L. (2004). D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779.

Webber, B. L. (2013). What excludes an Alternative in Coherence Relations? In *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany*, pages 276–287.

Weiss, G. and Bajec, M. (2016). Discourse Sense Classification from Scratch using Focused RNNs. In *Proceedings of the CoNLL-16 shared task*, pages 50–54. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Wolf, F. and Gibson, E. (2005). Representing Discourse Coherence: A Corpus-Based Study. *Comput. Linguist.*, 31(2):249–288.

Wolf, F., Gibson, E., Fisher, A., and Knight, M. (2003). A procedure for collecting a database of texts annotated with coherence relations.

Wu, C., Shi, X., Su, J., Chen, Y., and Huang, Y. (2017). Co-training for Implicit Discourse Relation Recognition Based on Manual and Distributed Features. *Neural Process. Lett.*, 46(1):233–250.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057. JMLR Workshop and Conference Proceedings.

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., and Rutherford, A. (2015). The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*, Beijing, China.

Xue, N., Ng, H. T., Pradhan, S., Webber, B., Rutherford, A., Wang, C., and Wang, H. (2016). The CoNLL-2016 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*, Berlin, Germany. Association for Computational Linguistics.

Xue, N. and Palmer, M. (2009). Adding semantic roles to the chinese treebank. *Natural Language Engineering*, 15(1):143–172.

Yee, E., Chrysikou, E. G., and Thompson-Schill, S. L. (2013). The cognitive neuroscience of semantic memory. *The Oxford handbook of cognitive neuroscience*, 1:353–374.

Yeh, C.-L. and Chen, Y.-J. (2001). An Empirical Study of Zero Anaphora Resolution in Chinese Based on Centering Model. In *Proceedings of Research on Computational Linguistics Conference XIV*, pages 237–251, Tainan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

You, L., Liu, T., and Liu, K. (2007). Chinese FrameNet and OWL representation. In *Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007)*, pages 140–145.

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). Relation Classification via Convolutional Deep Neural Network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Zeng, W., Luo, W., Fidler, S., and Urtasun, R. (2016). Efficient Summarization with Read-Again and Copy Mechanism. *CoRR*, abs/1611.03382.

Zhang, B., Su, J., Xiong, D., Lu, Y., Duan, H., and Yao, J. (2015). Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2230–2235.

Zhang, B., Xiong, D., and Su, J. (2016a). Neural Discourse Relation Recognition with Semantic Memory. *CoRR*, abs/1603.03873.

Zhang, B., Xiong, D., and Su, J. (2016b). Variational Neural Discourse Relation Recognizer. *CoRR*, abs/1603.03876.

Zhou, D., Zhang, X., and He, Y. (2017). Event extraction from Twitter using Non-Parametric Bayesian Mixture Model with Word Embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 808–817, Valencia, Spain. Association for Computational Linguistics.

Zhou, J. and Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China. Association for Computational Linguistics.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Zhou, Y., Jill Lu, J. Z., and Xue, N. (2014). Chinese Discourse Treebank 0.5, LDC2014T21. Web Download, Philadelphia: Linguistic Data Consortium.

Zhou, Y. and Xue, N. (2012). PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77, Jeju Island, Korea. Association for Computational Linguistics.

Zhou, Z.-M., Xu, Y., Niu, Z.-Y., Lan, M., Su, J., and Tan, C. L. (2010). Predicting Discourse Connectives for Implicit Discourse Relation Recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1507–1514, Beijing, China. Association for Computational Linguistics.

Álvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., and Rigau, G. (2008). Complete and Consistent Annotation of WordNet using the Top Concept Ontology. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.