From Sartre to Frege in Three Steps: A* Search for Enriching Semantic Text Similarity Measures

Davide Colla
University of Turin,
Computer Science Department
davide.colla@unito.it

Marco Leontino
University of Turin,
Computer Science Department
marco.leontino@unito.it

Enrico Mensa
University of Turin,
Computer Science Department
enrico.mensa@unito.it

Daniele P. Radicioni
University of Turin,
Computer Science Department
daniele.radicioni@unito.it

Abstract

English. In this paper we illustrate a preliminary investigation on semantic text similarity. In particular, the proposed approach is aimed at complementing and enriching the categorization results obtained by employing standard distributional resources. We found that the paths connecting entities and concepts from documents at stake provide interesting information on the connections between document pairs. Such semantic browsing device enables further semantic processing, aimed at unveiling contexts and hidden connections (possibly not explicitly mentioned in the documents) between text documents.[1]

1 Introduction

In the last few years many efforts have been spent to extract information contained in text documents, and a large number of resources have been developed that allow exploring domain-based knowledge, defining a rich set of specific semantic relationships between nodes (Vrandecic and Krötzsch, 2014; Auer et al., 2007; Navigli and Ponzetto, 2012). Being able to extract and to make available the semantic content of documents is a challenging task, with beneficial impact on different applications, such as document categorisation (Carducci et al., 2019), keyword extraction (Colla et al., 2017), question answering, text summarisation, semantic texts comparison, on building explanations/justifications for similarity judgements (Colla et al., 2018) and more. In this paper we present an approach aimed at extracting meaningful information contained in text documents, also based on background information contained in an encyclopedic resource such as Wikidata (Vrandecic and Krötzsch, 2014).

Although our approach has been devised on a specific application domain (PhD theses in philosophy), we argue that it can be easily extended to further application settings. The approach focuses on the ability to extract relevant pieces of information from text documents, and to map them onto the nodes of a knowledge graph, obtained from semantic networks representing encyclopedic and lexicographic knowledge. In this way it is possible to compare different documents based on their graphical description, which has a direct anchoring to their semantic content.

We propose a system to assess the similarity between textual documents, hybridising the propositional approach (such as traditional statements expressed through RDF triples) with a distributional description (Harris, 1954) of the nodes contained in the knowledge graph, that are represented with word embeddings (Mikolov et al., 2013; Camacho-Collados et al., 2015; Speer et al., 2017). This step allows to obtain similarity measures (based on vector descriptions, and on pathfinding algorithms) and explanations (represented as paths over a semantic network) more focused on the semantic definition of concepts and entities involved in the analysis.

2 Related Work

Surveying the existing approaches requires to briefly introduce the most widely used resources along with their main features.

Resources

BabelNet (BN) is a wide-coverage multilingual semantic network, originally built by integrating

WordNet (Miller, 1995) and Wikipedia (Navigli and Ponzetto, 2010). NASARI is a vectorial resource whose senses are represented as vectors associated to BabelNet synsets (Camacho-Collados et al., 2015). Wikidata is a knowledge graph based on Wikipedia, whose goal is to overcome problems related to information access by creating new ways for Wikipedia to manage its data on a global scale (Vrandecic and Krötzsch, 2014).

2.1 Approaches to semantic text similarity

Most literature in computing semantic similarity between documents can be arranged into three main classes.

Word-based similarity. Word-based metrics are used to compute the similarity between documents based on their terms; examples of features analysed are common morphological structures (Islam and Inkpen, 2008) and words overlap (Huang et al., 2011) between the texts. In one of the most popular theories on similarity (the Tversky's contrast model) the similarity of a word pair is defined as a direct function of their common traits (Tversky, 1977). This notion of similarity has been recently adjusted to model human similarity judgments for short texts: the Symmetrical Tversky Ratio Model (Jimenez et al., 2013), and employed to compute semantic similarity between word- and sense-pairs (Mensa et al., 2017; Mensa et al., 2018).

Corpus-based similarity. Corpus-based measures try to identify the degree of similarity between words using information derived from large corpora (Mihalcea et al., 2006; Gomaa and Fahmy, 2013).

Knowledge-based similarity. Knowledge-based measures try to estimate the degree of semantic similarity between documents by using information drawn from semantic networks (Mihalcea et al., 2006). In most cases only the hierarchical structure of the information contained in the network is considered, without considering the relation types within nodes (Jiang and Conrath, 1997; Richardson et al., 1994); some authors consider the "is-a" relation (Resnik, 1995), but leaving unexploited the more domain-dependent ones. Moreover, only concepts are usually considered, omitting the Named Entities.

An emerging paradigm is that of *knowledge graphs.* Knowledge graph extraction is a challenging task, particularly popular in recent years (Schuhmacher and Ponzetto, 2014). Several approaches have been developed, e.g., aimed at extracting knowledge graphs from textual corpora, attaining a network focused on the type of documents at hand (Pujara et al., 2013). Such approaches may be affected by scalability and generalisation issues. In the last years many resources representing knowledge in a structured form have have been proposed that build on encyclopedic resources (Auer et al., 2007; Suchanek et al., 2007; Vrandecic and Krötzsch, 2014).

As regards as semantic similarity, a framework has been proposed based on entity extraction from documents, providing mappings to knowledge graphs in order to compute semantic similarities between documents (Paul et al., 2016). Their similarity measures are mostly based on the network structure, without introducing other instruments such as embeddings, that are largely acknowledged as relevant in semantic similarity. Hecht et al. (2012) propose a framework endowed with explanatory capabilities from similarity measures based on relations between Wikipedia pages.

3 The System

In this Section we illustrate the generation process of the knowledge graph from Wikidata, which will be instrumental to build paths across documents. Such paths are then used, at a later time, to enrich the similarity scores computed during the classification.

3.1 Knowledge Graph Extraction

The first step consists of the extraction of a knowledge graph related to the given reference domain. Wikidata is then searched for concepts and entities related to the domain being analysed. By starting from the extracted elements, which constitute the basic nodes of the knowledge graph, we still consider Wikidata and look for relevant semantic relationships towards other nodes, not necessarily already extracted in the previous step. The types of relevant relationships depend on the treated domain. Considering the philosophical domain, we selected a set of 30 relations relevant to compare the documents. For example, we considered the relation *movement* that represents the literary, artistic, scientific or philosophical movement,the relation *studentOf* that represents the person who has taught the considered philosopher, and the relation *influencedBy* that represents the person's

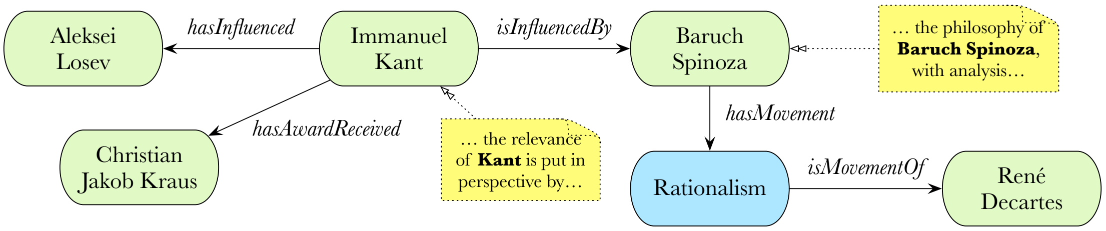


Figure 1: A small portion of the knowledge graph extracted from Wikidata, related to the philosophical domain; nodes represent BabelSynsets (concepts or NEs), rectangles represent documents.

idea from which the considered philospher's idea has been influenced. In this way, we obtain a graph where each node is a concept or entity extracted from Wikidata; such nodes are connected with edges labeled with specific semantic relations.

The obtained graph is then mapped onto BabelNet. At the end of the first stage, the knowledge graph represents the relevant domain knowledge (Figure 1) encoded through BabelNet nodes, that are connected through the rich set of relations available in Wikidata. Each text document can be linked to the knowledge graph, thereby allowing to make semantic comparisons by analysing the possible paths connecting document pairs.

Without loss of generality, we considered the philosophical domain, and extracted a knowledge graph containing $22,672$ nodes and $135,910$ typed edges; Wikidata entities were mapped onto BabelNet approximately in the 90% of cases.

3.2 Information extraction and semantic similarity

The second step consists in connecting the documents to the obtained knowledge graph. We harvested a set of $475,383$ UK doctoral theses in several disciplines through the Electronic Theses Online Service (EThOS) of the British National Library.[2] At first, concepts and entities related to the reference domain were extracted from the considered documents, with a special focus on two different types of information, such as *concepts* and *Named Entities*. *Concepts* are keywords or multiword expressions representing meaningful items related to the domain (such as, e.g., 'philosophy-of-mind', 'Rationalism', *etc.*) while *Named Entities* are persons, places or organisations (mostly universities, in the present setting) strongly related to the considered domain. Named entities are extracted using the Stanford CoreNLP NER module (Manning et al., 2014) improved with extraction rules based on morphological and syntactical patterns, considering for example sequences of words starting with a capital letter or associated to a particular Part-Of-Speech pattern. Similarly, we extract relevant concepts based on particular PoS patterns (such as `NOUN-PREPOSITION-NOUN`, thereby recognizing, for example, *philosophy of mind*).

We are aware that we are not considering the problem of word sense disambiguation (Navigli, 2009; Tripodi and Pelillo, 2017). The underlying assumption is that as long as we are concerned with a narrow domain, this is a less severe problem: e.g., if we recognise the person *Kant* in a document related to philosophy, probably the person cited is the philosopher whose name is *Immanuel Kant* (please refer to Figure 1), rather than the less philosophical Gujarati poet, playwright and essayist Kavi Kant.[3]

By mapping concepts and Named Entities found in a document onto the graph, we gain a set of *access points* to the knowledge graph. Once acquired the access points to the knowledge graph for a pair of documents, we can compute the semantic similarity between documents by analysing the paths that connect them.

3.3 Building Paths across Documents

The developed framework is used to compute paths between pairs of senses and/or entities featuring two given documents. Each edge in the knowledge graph has associated a semantic relation type (such as, e.g., "*hasAuthor*", "*influencedBy*", "*hasMovement*"). Each path intervening between two documents is in the form

$$DOC_1 \xrightarrow{ACCESS} SaulKripke \xrightarrow{influencedBy} LudwigWittgenstein \xrightarrow{influencedBy} BertrandRussell \xrightarrow{influencedBy} BaruchDeSpinoza \xleftarrow{ACCESS} DOC_2$$

[2] https://ethos.bl.uk.

[3] https://tinyurl.com/y3s9lsp7.

In this case we can argue in favor of the relatedness of the two documents based on the chain of relationships illustrating that *Saul Kripke* (from document d_1) has been *influenced-by* Ludwig Wittgenstein, that has been *influenced-by* Bertrand Russel, that in turn has been *influenced-by* Baruch De Spinoza, mentioned in d_2. The whole set of paths connecting elements from a document d_1 to a document d_2 can be thought of as a form of evidence of the closeness of the two documents: documents with numerous shorter paths connecting them are intuitively more related. Importantly enough, such paths over the knowledge graph do not contain general information (e.g., Kant was a man), but rather they are highly domain-specific (e.g., Oskar Becker had as doctoral student Jürgen Habermas).

A* Search

The computation of the paths is performed via a modified version of the $A^\star$ algorithm (Hart et al., 1968). In particular, paths among access nodes are returned in order, from the shortest to the longest one. Given the huge dimension of the network, and since we are guaranteed to retrieve shortest paths first, we stop the search after one second of computation time.

4 Experimentation

In this Section we report the results of a preliminary experimentation: given a dataset of PhD theses, we first explore the effectiveness of standard distributional approaches to compute the semantic similarity between document pairs; we then elaborate on how such results can be complemented and enriched through the computation of paths between entities therein.

Experimental setting We extracted 4 classes of documents (100 for each class) from the EThOS dataset. For each record we retrieved the title and abstract fields, that were used for subsequent processing. We selected documents containing 'Antibiotics', 'Molecular', 'Hegel' or 'Ethics' either in their title (in 15 documents per class) or in their abstract (15 documents per class). Each class is featured on average by 163.5 tokens (standard deviation $\sigma = 39.3$), including both title and abstract. The underlying rationale has been that of selecting documents from two broad areas, each one composed by two different sets of data, having to do with medical disciplines and molecular biology in the former case, and with Hegelianism and the broad theme of ethics in the latter case. Intra-domain classes (that is both 'Antibiotics'-'Molecular' and 'Hegel'-'Ethics') are not supposed to be linearly separable, as it mostly occurs in real problems. Of course, this feature makes more interesting the categorization problem. The dataset was used to compute some descriptive stats (such as inverse document frequency), characterizing the whole collection of considered documents.

From the aforementioned set of 400 documents we randomly chose a subset of 20 documents, 5 documents for each of the 4 classes from those containing the terms (either 'Antibiotics', 'Molecular', 'Hegel' or 'Ethics') in the title. This selection strategy was aimed at selecting more clearly individuated documents, exhibiting a higher similarity degree within classes than across classes.[4]

4.1 Investigation on Text Similarity with Standard Distributional Approaches

GLoVE and Word Embedding Similarity

The similarity scores were computed for each document pair with a Word Embedding Similarity approach (Agirre et al., 2016). In particular, each document d has been provided with a vector description averaging the GloVe embeddings t_i (Pennington et al., 2014) for all terms in the title and abstract:

$$\overrightarrow{N_d} = \frac{1}{|T_d|} \sum_{t_i \in T_d} \vec{t_i}, \tag{1}$$

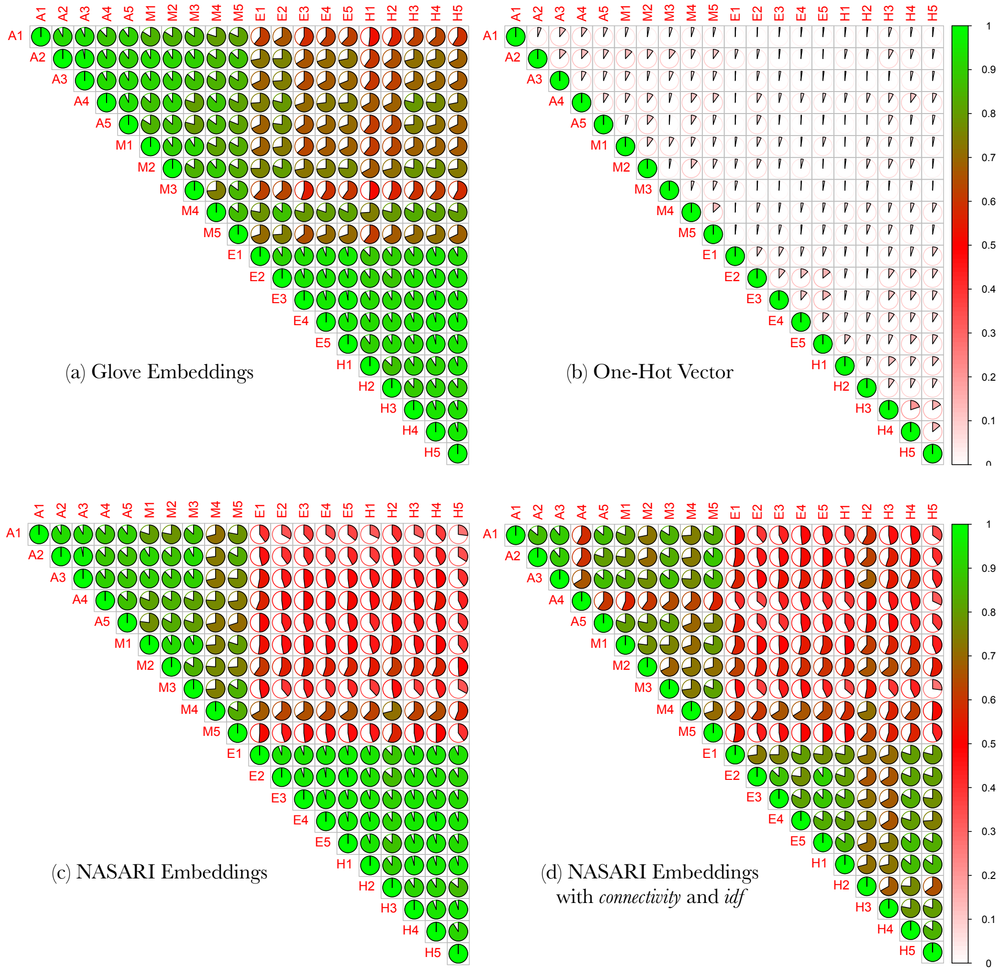
where each $\vec{t_i}$ is the GloVe vector for the term t_i. Considering two documents d_1 ad d_2, each one associated to a particular vector $\overrightarrow{N_{d_i}}$, we compare them using the cosine similarity metrics:

$$sim(\overrightarrow{N_{d_1}}, \overrightarrow{N_{d_2}}) = \frac{\overrightarrow{N_{d_1}} \cdot \overrightarrow{N_{d_2}}}{\|\overrightarrow{N_{d_1}}\| \|\overrightarrow{N_{d_2}}\|}. \tag{2}$$

The obtained similarities between each document pair are reported in Figure 2(a).[5] The computed distances show that overall this approach is sufficient to discriminate the scientific doctoral theses from the philosophical ones. In particular, the top green triangle shows the correlation scores among antibiotics documents, while the bottom triangle reports the correlation scores among philo-

[4] In future work we will verify such assumptions by involving domain experts in order to validate and/or refine the heuristics employed in the document selection.

[5] The plot was computed using the *corrplot* package in R.



(a) Glove Embeddings

(b) One-Hot Vector

(c) NASARI Embeddings

(d) NASARI Embeddings with *connectivity* and *idf*

Figure 2: Comparison between correlation scores. Documents have scientific subject ('A' for 'Antibiotics', 'M' for 'Molecular' biology), and philosophic subject ('E' for 'Ethics', 'H' for 'Hegel').

sophical documents. The red square graphically illustrates the poor correlation between the two classes of documents. On the other side, the subclasses (Hegelism-Ethics and Antibiotics-Molecular) could not be separated. Provided that word embeddings are known to conflate all senses in the description of each term (Camacho-Collados and Pilehvar, 2018), this approach performed surprisingly well in comparison to a baseline based on a one-hot vector representation, only dealing with term-based features (Figure 2(b)).

NASARI and Sense Embedding Similarity

We then explored the hypothesis that semantic knowledge can be beneficial for better separating documents: after performing word sense disambiguation (the BabelFy service was employed (Moro et al., 2014)), we used the NASARI embedded version to compute the vector $\overrightarrow{N_d}$, as the average of all vectors associated to the senses contained in S_d, basically employing the same formula as in Equation 1. We then computed the similarity matrix, displayed in Figure 2(c). It clearly emerges that also NASARI is well suited to solve a classification task when domains are well separated. However, also in this case the adopted approach does not seem to discriminate well within the two main classes: for instance, the square with vertices E1-H1; E5-H1; E5-H5; E1-H5 should be reddish, indicating a lower average similarity between documents pertaining the Hegel and Ethics classes. We experimented in a set of widely varied conditions and parameters, obtaining slightly better similarity scores by weighting NASARI vectors with senses *idf*, and senses connectivity (c, obtained from BabelNet):

$$\overrightarrow{N_d} = \frac{1}{|S_d|} \sum_{s_i \in S_d} \vec{s_i} \cdot \log\left(\frac{|S_d|}{H(s_i)}\right) \cdot \left(1 - \frac{1}{c}\right), \quad (3)$$

where $H(s_i)$ is the number of documents containing the sense s_i. The resulting similarities scores are provided in Figure 2(d).

Documents are in fact too close, and presumably the adopted representation (merging all senses in each document) is not as precise as needed. In this setting, we tried to investigate the documents similarity based on the connections between their underlying sets of senses. Such connections were computed on the aforementioned graph.

4.2 Enriching Text Similarity with Paths across Documents

In order to examine the connections between the considered documents we focused on the philosophical portion of our dataset, and exploited the knowledge graph described in Section 3. The computed paths are not presently used to refine the similarity scores, but only as a suggestion to characterize possible connections between document pairs. The extracted paths contain precious information that can be easily integrated in downstream applications, by providing specific information that can be helpful for domain experts to achieve their objectives (e.g., in semantically browsing text documents, in order to find influence relations across different philosophical schools).

As anticipated, building paths among the fundamental concepts of the documents allows grasping important ties between the documents topics. For instance, one of the extracted paths (between the author 'Hegel' and the work 'Sense and Reference' (Frege, 1948)) shows the connections between the entities at stake as follows. G.W.F. Hegel *hasMovement* Continental Philosophy, which is in turn the *movementOf* H.L. Bergson, who has been *influencedBy* G. Frege, who finally *hasNotableWork* Sense and Reference. The semantic specificity of this information provides precious insights that allow for a proper consideration of the relevance of the second document w.r.t. the first one. It is worth noting that the fact that Hegel is a continental philosopher is trivial –tacit knowledge– for philosophers, and was most probably left implicit in the thesis abstract, while it can be a relevant piece of information for a system requested to assess the similarity of two philosophical documents. Also, this sort of path over the extracted knowledge graph enables a form of semantic browsing that benefits from the rich set of Wikidata relations paired with the valuable coverage ensured by BabelNet on domain-specific concepts and entities.

The illustrated approach allows the uncovering of insightful and specific connections between documents pairs. However, this preliminary study also pointed out some issues. One key problem is the amount of named entities contained in the considered documents (e.g., E5 only has one access point, while E3 has none). Another issue has to do with the inherently high connectivity of some nodes of the knowledge graph (hubness). For instance, the nodes *Philosophy*, *Plato* and *Aristotle* are very connected, which results in the extraction of some trivial and uninteresting paths among the specific documents. The first issue could be tackled by also considering the main concepts of a document if no entity can be found, whilst the second one could be mitigated by taking into account the connectivity of the nodes as a negative parameter while computing the paths.

5 Conclusions

In this paper we have investigated the possibility of enriching semantic text similarity measures via symbolic and human readable knowledge. We have shown that distributional approaches allow for a satisfactory classification of documents belonging to different topics, however, our preliminary experimentation showed that they are not able to capture the subtle aspects characterizing documents in close areas. As we have argued, exploiting paths over graphs to explore connections between document pairs may be beneficial in making explicit domain-specific links between documents.

As a future work, we could refine the methodology related to the extraction of the concepts in the Knowledge Graph, defining approaches based on specific domain-related ontologies. Two relevant works, to these ends, are the *PhilOnto* ontology, that represents the structure of philosophical literature (Grenon and Smith, 2011), and the *InPho* taxonomy (Buckner et al., 2007), combining automated information retrieval methods with knowledge from domain experts. Both resources will be employed in order to extract a more concise, meaningful and discriminative Knowledge Graph.

Acknowledgments

The authors are grateful to the EThOS staff for their prompt and kind support. Marco Leontino has been supported by the REPOSUM project, BONG_CRT_17_01 funded by Fondazione CRT.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Cameron Buckner, Mathias Niepert, and Colin Allen. 2007. Inpho: the indiana philosophy ontology. *APA Newsletters-newsletter on philosophy and computers*, 7(1):26–28.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a novel approach to a semantically-aware representation of items. In *Proceedings of NAACL*, pages 567–577.

Giulio Carducci, Marco Leontino, Daniele P Radicioni, Guido Bonino, Enrico Pasini, and Paolo Tripodi. 2019. Semantically aware text categorisation for metadata annotation. In *Italian Research Conference on Digital Libraries*, pages 315–330. Springer.

Davide Colla, Enrico Mensa, and Daniele P Radicioni. 2017. Semantic measures for keywords extraction. In *Conference of the Italian Association for Artificial Intelligence*, pages 128–140. Springer.

Davide Colla, Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2018. Tell me why: Computational explanation of conceptual similarity judgments. In *Proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Special Session on Advances on Explainable Artificial Intelligence*, Communications in Computer and Information Science (CCIS), Cham. Springer International Publishing.

Gottlob Frege. 1948. Sense and reference. *The philosophical review*, 57(3):209–230.

Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18.

Pierre Grenon and Barry Smith. 2011. Foundations of an ontology of philosophy. *Synthese*, 182(2):185–204.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, SSC-4(2):100–107.

Brent Hecht, Samuel H Carton, Mahmood Quaderi, Johannes Schöning, Martin Raubal, Darren Gergle, and Doug Downey. 2012. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 415–424. ACM.

Cheng-Hui Huang, Jian Yin, and Fang Hou. 2011. A text similarity measurement combining word semantic information with tf-idf method. *Jisuanji Xuebao(Chinese Journal of Computers)*, 34(5):856–864.

Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):10.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Sergio Jimenez, Claudia Becerra, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. 2013. Softcardinality-core: Improving text overlap with distributional measures for semantic textual similarity. In *Proceedings of *SEM 2013*, volume 1, pages 194–201.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 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.

Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2017. Merali at semeval-2017 task 2 subtask 1: a cognitively inspired approach. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 236–240, Vancouver, Canada, August. Association for Computational Linguistics.

Enrico Mensa, Daniele P Radicioni, and Antonio Lieto. 2018. Cover: a linguistic resource combining common sense and lexicographic information. *Language Resources and Evaluation*, 52(4):921–948.

Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Christian Paul, Achim Rettinger, Aditya Mogadala, Craig A Knoblock, and Pedro Szekely. 2016. Efficient graph-based document similarity. In *European Semantic Web Conference*, pages 334–349. Springer.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2013. Knowledge graph identification. In *International Semantic Web Conference*, pages 542–557. Springer.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Ray Richardson, A Smeaton, and John Murphy. 1994. Using wordnet as a knowledge base for measuring semantic similarity between words.

Michael Schuhmacher and Simone Paolo Ponzetto. 2014. Knowledge-based graph document modeling. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 543–552. ACM.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Rocco Tripodi and Marcello Pelillo. 2017. A game-theoretic approach to word sense disambiguation. *Computational Linguistics*, 43(1):31–70.

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57(10).