Mining Meaning from Text by Harvesting Frequent and Diverse Semantic Itemsets

Luigi Di Caro and Guido Boella

Department of Computer Science
University of Turin, Italy
{dicaro,boella}@di.unito.it

Abstract. In this paper, we present a novel and completely-unsupervised approach to unravel meanings (or senses) from linguistic constructions found in large corpora by introducing the concept of *semantic vector*. A semantic vector is a space-transformed vector where features represent fine-grained semantic information units, instead of values of co-occurrences within a collection of texts. More in detail, instead of seeing words as vectors of frequency values, we propose to first explode words into a multitude of tiny semantic information retrieved from existing resources like WordNet and ConceptNet, and then clustering them into frequent and diverse patterns. This way, on the one hand, we are able to model linguistic data with a larger but much more dense and informative semantic feature space. On the other hand, being the model based on basic and conceptual information, we are also able to generate new data by querying the above-mentioned semantic resources with the features contained in the extracted patterns. We experimented the idea on a dataset of 640 millions of triples *subject-verb-object* to automatically inducing senses for specific input verbs, demonstrating the validity and the potential of the presented approach in modeling and understanding natural language.

Keywords: Natural Language Understanding, Distributional Semantics, Diverse Itemset Mining

1 Introduction

Most Computational Linguistics applications may need semantic information to improve their effectiveness. Semantic resources are often constructed with automatic approaches, since manual building of ontologies is not feasible on large scale [24, 42, 19, 37, 31, 5, 7].

Distributional Semantics (DS) is nowadays one of the frontiers in this field [4, 22, 8, 2, 14]. DS derives from the *Distributional Hypothesis* introduced by Z. Harris [23], where Vector Space Models (VSMs) represent its main expression [39]. The current availability of huge corpora like ukWac [17] makes these approaches particularly efficient. Data Mining (DM) techniques leveraging on VSMs have been successfully applied on text since many decades on Topic Extraction-related

In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Nancy, France, 2014.

tasks [10, 1, 13]. Specifically, terms become interconnected by similarity scores forming concept-like entities, i.e., words clusters sharing similar contexts [9]. DS refines traditional DM on text, since it considers language as a grammar-based type of data. However, DS still sees linguistically-refined tokens as the basic bricks in VSMs, suffering of an intrinsic limitation: a wide range of words and grammar constructions is actually rarely used. Even in very large corpora, there is little chance of finding statistically-significant patterns that can be used to carve out meanings out of them. This is known as the long tail problem [45]. Moreover, DM starts from linguistic items to develop semantic information, without reusing it for further analysis.

This work is based on an interdisciplinary approach that relies on Conceptual Spaces [20], a theory developed by P. Gardenfors in the Cognitive Science field, through which concepts are represented by vectors whose features are cognitive axes that humans naturally use to give meaning to their perceptions. In this sense, rather than considering VSMs of linguistic symbols we will consider VSMs of extensive semantic information associated with them, derived from different sources. Our methodology leverages on the wealth of resources available on the web concerning semantics, like Linked Open Data (e.g., DBPedia[1], Freebase[2], etc.), linguistic resources (e.g., WordNet [29], ConceptNet [44], BabelNet [32], etc.), Semantic Web technologies (e.g., FRED [15], TPALO [33], etc.), and automatic parsing of large corpora like Wikipedia to map linguistic contexts into semantic features.

An initial proof-of-concept of the proposal is given by recent research in which the transformation of terms into top-level hypernyms carried to improvement in several computational tasks, as in [28, 18]. While this is in line with this paper, this transformation only involves terminological abstractions by means of IS-A substitutions. In fact, this contribution represents a large generalization that takes into account a wider spectrum of conceptual relationships. The outcomes of this work are threefold:

- a new methodology that introduces the concept of *semantic vectors*
- a novel technique for mining frequent and *diverse* itemsets based on *set cover problem* [41], implemented with an heuristic approach.
- a model that generalizes over existing linguistic constructions with low resource requirements that is also able to generate new linguistic data

The paper first presents the motivations and the goals of this work. Then, the approach is explained in terms of methodology and algorithms. An evaluation phase is then presented, showing the data and the pursued objectives. A final part of conclusions and future work ends the paper.

[1] http://dbpedia.org/About

[2] http://www.freebase.com/

2 Background and Related Work

In Computational Linguistics, recent advances and successful experiences of statistical distributional approaches can be found in different tasks. The IBM Watson Question Answering system[3] is maybe the most recent and well-known direct result. This also explains the fortunate and growing trend of available semantic resources often constructed with automatic approaches, since manual building of ontologies is not feasible on large scale. Currently, many systems actually try to automatically extract semantic knowledge from texts by means of three possible generic approaches: distributional analysis, pattern-based approaches, and Machine Learning techniques.

Nowadays, semantic information extraction is currently approached by distributional analysis of linguistic items over specific contexts [34] or by starting from seeds and patterns to build ontologies from scratch [31]. In some cases, linguistic items are substituted by super-senses (i.e., top-level hypernyms) [28]. However, such generalization should be applied taking into account a wider notion of semantic context introduced by related cognitive theories [21], that has not been addressed by current computational approaches.

Distributional Semantics (DS) is nowadays one of the frontiers[4] within the Computational Linguistics field [3]. DS derives from the *distributional hypothesis* introduced by Z. Harris in 1954 [23]. Vector Space Models (VSMs) [39], proposed by Gerald Salton in the seventies, are the main expression of this idea. Data Mining (DM) techniques fully leveraging on VSMs and Latent Semantic Analysis (LSA) [16] have been successfully applied on text since many decades on topic extraction-related tasks, often producing concept-like entities, i.e., words clusters sharing similar contexts [9].

Current research in DS focuses on the exploration of the different impact of parameters such as *context type* (i.e., text regions vs. linguistic items), *window* (context extension), *frequency weighting strategy* (e.g., number of occurrences, Pointwise Mutual Information, etc.), *dimensionality reduction* (e.g., Latent Semantic Analysis, Random Indexing, etc.), and *similarity measure* (e.g., Cosine similarity, Jaccard's coefficient, etc.). Then, it produces co-occurrences matrices (or tensors) that model the semantics of the tokens by means of weights distributions.

DS refines traditional DM on text, since it considers language as a grammar-based type of data instead of simple unstructured paragraphs. However, DS still sees linguistically-refined tokens (words, lemmas, part-of-speech, etc.) as the basic bricks in VSMs, suffering of an intrinsic limitation: a wide range of words and grammar constructions is actually rarely used.

On the contrary, this work concerns a radical departure from this direction, releasing the assumption made by all approaches to rely on linguistic items (either terms or more context-aware tokens). The current methodology still starts

[3] http://www.ibm.com/smarterplanet/us/en/ibmwatson/

[4] See also the ERC project COMPOSES leaded by Marco Baroni. http://clic.cimec.unitn.it/composes/

from syntax and strings of text to extract semantics, while it would be more reasonable to have an automated approach which also leverages on the existing semantic resources it produces as further input. The idea at the basis of the proposed approach is to conceptually restructure the problem of DS under the light of research in Cognitive Science. The above-mentioned theory of Conceptual Spaces introduced by Peter Gardenfors is about a concept representation motivated by the notions of conceptual similarity and prototype theory [38]. A *conceptual space* is a multi-dimensional feature space where points denote objects, and regions define concepts. Its bases are composed by quality dimensions, which denote basic features in which concepts and objects can be compared, such as weight, color, taste and so on. Symbolic representations are particularly weak at modeling concept learning, which is paramount for understanding many cognitive phenomena. Concept learning is closely tied to the notion of similarity, which is also poorly served by the symbolic approach.

Taking inspiration from this vision of language, as basic bricks of DS we substitute linguistic items with a representation of their meaning in terms of sets of quality dimensions. In detail, each word can be represented by a set of semantic relationships and properties that define a specific concept by means of physical and behavioral facts. For instance, a cat has legs, claws, paws, eyes, etc. (properties); then, it usually chases mouses and it sees in the dark (behaviour); it is an animal and a feline (taxonomical information), and it can have many other relations like the places where it can be habitually found.

The Conceptual Spaces (CS) framework developed in the Cognitive Sciences field by [20] is based on a vectorial representation of concepts whose features are cognitive axes through humans naturally give meaning to their perceptions. CS is directly connectable with VSMs since it is a particular type of VSMs where features represent the conceptual level. Our approach is about injecting semantics into tokens towards a concept-level feature set. One of the most important brick in almost all Computational Linguistics tasks is the computation of similarity scores between texts at different levels: terms, sentences and discourses. As recently discussed in the literature [36], semantic similarity needs to be cross-level.

In the DS current view, the triple *subject-verb-object* extracted from the sentence "*the cat climbs a tree*" is equally seen as the triple extracted from "*the monkey climbs a tree*", since the two subjects share the same linguistic context. In this work, instead, the two situations will be differentiated and therefore more deeply understood: in the first case, it will be able to correlate the fact of having claws with the ability of climbing a tree; in the second case, this will happen for the presence of prehensile feet. This is due to the introduction of semantics within the process of distributional analysis itself. In fact, they share physical body parts with a similar kind of functionality. Since only specific semantic information are useful at a time, this new approach can also filter out non-relevant information (for instance, the fact that both are mammals with fur and teeth does not suggest the ability to climb a tree). Nowadays, the extraction of these features can be done due to the huge availability of semantic resources.

Once linguistic items are replaced by semantic representations, it becomes possible to reuse the methodology itself having as input the larger basis of semantic information created by the system, thus creating a positive feedback cycle and enlarging the possibilities of the system. We call this concept as *semantic loop*, and, to the best of our knowledge, it is the first attempt to go beyond single-processing systems that connect syntax with semantics towards recursive processing of extracted semantics. For example, the action of "*seeing*" can show a correlation with the fact of having eyes. Nowadays, the link between actions and properties of subject and objects are not used while they actually provide significant information for deeper language understanding.

This paper presents an approach that also relies on the concept of *diversity*. Diversity has been taken into account mostly in Information Retrieval (IR) scenarios, where systems become aware of the need of obtaining search results that cover different aspects of the data [12, 11]. However, this concept can be also useful in different contexts like clustering [30] and recommendation [40]. In spite of this, within the Pattern Mining (PM) and Association Rules (AR) areas, to the best of our knowledge, diversity has not been faced yet. Since our system architecture needs to manage the output of these techniques with the additional goal of producing frequent patterns that are able to cover different aspects of the input, we also revisited them in this sense.

This shift in the basic bricks opens new research questions and challenges concerning Data Mining methodologies: the problem of correlating atomic linguistic items becomes to correlate sets of features, where only some of them are actually significant. Thus, the new challenges become to understand:

- which features need to be filtered out
- which features can be combined to approximate concepts (according to Conceptual Spaces)

The advantages of the proposed research direction are the following:

- the integration of semantic information within the internal steps of the current methodology can create a virtuous loop through which semantic resources can be automatically extended.
- linguistic items are fragmented into minimal conceptual information that enables statistical approaches to overcome the problem of low-frequency words occurrences. In fact, even in very large corpora, there is little chance of finding statistically-significant patterns that can be used to carve out meanings out of text. This is known as the *long tail* problem. Statistical approaches are usually not able to propagate existing information belonging to frequent data to such *long tail*. One of the aim of this proposal is to define a linguistic framework in which both rare and frequent words are fragmented into more basic facts on which reason on, avoiding low-frequency issues.
- the use of multilingual resources will have an impact on the creation of more powerful semantic resources, that will be more independent by word-to-word translations. Within the DS field, a minimal step in this direction has already been done by means of transformations of words into general concepts

or super-senses. However, this only involves terminological abstractions by means of IS-A relationship substitutions. In fact, our proposal represents a generalization of this since it considers a wider spectrum of conceptual relationships. For example, a person can assume the role of living entity, doctor or student in the context of breathing, making surgical interventions, and studying mathematics respectively. The point is that only specific properties are activated by the context at a time, so we avoid to assign fixed top-level hypernyms for all the cases. In addition to this, the simple generalization of a linguistic item does not extend the current analysis of correlations between atomic tokens.

The outcomes of such novel approach can be many:

- a new methodology that introduces the concept of *semantic loop*, i.e., iterative use of extracted semantics as input for further extension of semantic resources
- new semantic resources, created by the use of the proposed methodology
- revisitations of Data Mining techniques for dealing with a new and more complex type of data with respect to standard VSMs applied on text
- the proposed contribution can also have impact on how semantic knowledge can be re-used or inherited from data in different languages. For instance, in case there is no translation for two words in two different languages, it will be possible to leverage their semantic information to link them automatically. Only translation at concept level it will be needed (i.e., translation of the new feature space). Thus, the semantic loop can work also for alignment of different languages.

3 Approach

Our proposal concerns an automatic methods to build a large-scale semantic framework based on a concept-level distributional analysis of the semantics contained in plain texts. Our methodology avoids manual constructions of ontologies which is known to be unfeasible. On the contrary, the method goes towards a direct and extensive exploitation of the wealth of available resources regarding semantics. In particular, it leverages different types of knowledge that can be used to transform words (intended as lemmas or generic linguistic items, from now on) into sets of extended and fine-grained semantic information. The resulting explosion of such heterogeneous knowledge, coming from different sources and methods, create a new challenge: how to align, filter, and merge it in order to feed Vector Space models with semantics, as opposite to lexical entities.

3.1 Semantic Basis

In this paper, we started focusing on ConceptNet [43], a semantic crowdsourced knowledge. In detail, the Open Mind Common Sense project developed by MIT collected unstructured common-sense knowledge by asking people to contribute over the Web. ConceptNet, a semantic graph created from a parsing of such knowledge, is its final outcome. In contrast with linguistic resources like WordNet [29], ConceptNet contains semantic information more related to common-sense facts. For this reason, it has a wider spectrum of semantic relationships but a much more sparse coverage due to the non-methodological approach that was used to build it. For instance, among the more unusual types of relationships (24 in total), it contains information like "*ObstructedBy*" (i.e., referring to what would prevent it from happening), "*CausesDesire*" (i.e., what does it make you want to do), and "*MotivatedByGoal*" (i.e., why would you do it). In addition, it also has classic relationships like "*is_a*" and "*part_of*" as in most linguistic resources. An at-a-glance view of these semantic relations is shown in Table 1.

Table 1. The relations in ConceptNet, with example sentences in English.

Relation	**Example sentence**
IsA	NP is a kind of NP.
LocatedNear	You are likely to find NP near NP.
UsedFor	NP is used for VP.
DefinedAs	NP is defined as NP.
HasA	NP has NP.
SymbolOf	NP represents NP.
CapableOf	NP can VP.
ReceivesAction	NP can be VP.
Desires	NP wants to VP.
HasPrerequisite	NPjVP requires NPjVP.
CreatedBy	You make NP by VP.
MotivatedByGoal	You would VP because you want VP.
PartOf	NP is part of NP.
CausesDesire	NP would make you want to VP.
Causes	The effect of VP is NPjVP.
MadeOf	NP is made of NP.
HasFirstSubevent	The first thing you do when you VP is NPjVP.
HasSubevent	One of the things you do when you VP is NPjVP.
AtLocation	Somewhere NP can be is NP.
HasLastSubevent	The last thing you do when you VP is NPjVP.
HasProperty	NP is AP.

In spite of this, the approach can work with other resources. For example, another type of knowledge that can have an high impact on our semantic integration comes from Linked Open Data (LOD). One of the most used LOD resources in Computational Linguistics is DBPedia, a dataset containing data directly extracted from Wikipedia. It contains more than 3 million concepts described by 1 billion triples, including descriptions in several languages. Other knowledge bases are UMBEL (i.e., a 20k subjects ontology derived from OpenCyc), GeoNames (i.e., descriptions of geographical features), and several others. Then, WordNet [29] is a large lexical database of English nouns, verbs, adjectives and adverbs that can further extend the semantic basis. All the words are therein grouped into sets of synonyms (also called synsets), each expressing a distinct concept. WordNet contains also a set of relationships that link the synsets. To make some examples, synsets can be used to extrapolate "*same_as*" properties from synonyms, then hypernyms can be mapped into "*is_a*" taxonomical information, while meronyms can be seen as "*part_of*" features.

3.2 Data for Distributional Analysis

In order to experiment the validity of the approach, we had the need of computing a distributional model starting from a large collection of texts. However, instead of parsing corpora from scratch, we used a dataset of *subject-verb-object* (SVO) triples generated as part of the NELL project[5]. This dataset contains a set of 604 million triples extracted from the entire dependency-parsed corpus *ClueWeb09* (about 230 billion tokens)[6]. The dataset also provides the frequency of each triple in the parsed corpus. We integrated a Named Entity Recognition module to transform proper names into generic semantic classes, like people and organizations[7].

3.3 Algorithm

In this section, we explain the details of the approach. In particular, the algorithm is composed by three different phases: (1) the data pre-processing step with the generation of two transactional databases (transactions of items, as in the fields of Frequent Itemset Mining and Association Rules [6]) that we also call *semantic vectors*; (2) the extraction of frequent, closed, and *diverse* itemsets (we will briefly introduce the meaning of all these names in the next paragraphs); and finally (3) the creation of *semantic verb models*, that generalize and automatically induce *senses* from entire linguistic constructions at sentence-level.

Transactional Databases Generation The first step of the algorithm regards the generation of the *semantic vectors*, i.e., vectors whose features represent conceptual and semantic facts rather than document- or context-occurrences. Since the aim of the system is to capture senses from data, we start from the root of the meaning, that is the verb. So, for a specific input verb v, we parse all the SVO triples in the datasets that have a frequency higher than a set threshold[8], and we only take those who are morphological variations of v. Then, for each one of these triples, we query ConceptNet with the subject-term and

[5] http://rtw.ml.cmu.edu/resources/svo/

[6] http://lemurproject.org/clueweb09/

[7] We used the Stanford NLP toolkit available at http://www-nlp.stanford.edu/.

[8] In our experiments we considered SVO triples that occur at least 100 times in the whole ClueWeb09 corpus, in order to remove noisy data.

the object-term, retrieving all their semantic information that will later build the new semantic space. Table 2 shows an example of the information collected in this phase.

Table 2. An example of subject- and object-terms semantic transformation for one triple of the verb "*to learn*" (*student-learns-math*). This represents one row of the two transactional databases.

Subject-term	Subject semantic features	Object-term	Object semantic features
student	*CapableOf*-study, *AtLocation*-at_school, *IsA*-person, *Desires*-learn, *PartOf*-class, *CapableOf*-read_book, ...	**math**	*IsA*-subject, *HasProperty*-useful_in_business, *UsedFor*-model_physical_world, *ReceivesAction*-teach_in_class, ...

Then, we associate each semantic information to a unique *id* and construct two transactional databases: one for the semantic information of the subjects, and one for the objects. An example of result of the first phase is shown in Table 3.

Table 3. An example of the two transactional databases created for the verb "to learn" and the ID-label association table.

Transactional DB of the subjects	Transactional DB of the objects
1 34 67 90	2 4 6 23 67 87 122 198
3 4 12 36 59 88 90 91	42 54 67 87 122 124
34 67 45	2 6 54 67 87
...	...

ID	Associated Semantic information
1	isa-young_person
2	atlocation-classroom
3	atlocation-at_school
4	capableof-learn
...	...

Diverse Itemsets Mining Once the transactional databases are built for a specific verb "*v*", we use techniques belonging to the field of Frequent Itemset Mining to extract *frequent patterns*, i.e, semantic features that frequently co-occur in our transactional databases.

The description of the problem is the following: let $I = i_1, i_2, ..., i_n$ be a set of *items* (i.e., our semantic features) and D be a multiset of transactions, where each transaction t is a set of items such that $t \subseteq I$. For any $X \subseteq I$, we say that a transaction t contains X if $X \subseteq t$. The set X is called *itemset*. The set of all $X \subseteq I$ (the powerset of I) naturally forms an itemset *lattice*. The count of an itemset X is the number of transactions in D that contain X. The *support* of an itemset X is the proportion of transactions in D that contain X.

An itemset X is called *frequent* if its support is greater than or equal to some given percentage threshold s, where s is called *minimum support*.

When the database contains a significant number of large frequent itemsets, mining all of them can be very expensive, since the space of itemsets to generate can be huge. However, if any subset of a frequent itemset is frequent, it can be sufficient to discover only all the maximal frequent itemsets (MFIs). A frequent itemset X is called maximal if there does not exist a frequent itemset Y such that $X \subseteq Y$. Mining frequent itemsets can thus be reduced to mining a "border" in the itemset lattice. All itemsets above the border are infrequent and those that are below the border are all frequent. Another type of frequent itemset, called *closed frequent itemset* (CFI), was proposed in [35]. A frequent itemset X is *closed* if none of its proper supersets have the same support.

In our experimentation, we used the library called SPMF for finding closed frequent itemsets[9], applying the CHARM algorithm [46]. This is done for both the transactional databases (subject and object databases associated to the verb '*v*'). Since our aim is to capture all the linguistic *senses*, i.e., the different meanings connectable to the use of a specific verb, we also need to obtain itemsets that cover all the items that are found in frequent itemsets. In other words, we want to extract *diverse itemsets*, i.e., a minimal set of frequent and closed itemsets that cover all the frequent items. The concept of *diversity* has been mostly used in Information Retrieval tasks, and to the best of our knowledge there is no attempt in capturing "kind of" diverse itemsets in the current literature.

In order to produce these novel types of frequent itemsets, we viewed the problem as a *set cover problem* [41], implementing an heuristic-based approach to face it. Given a set of elements $U = i_1, i_2, ..., i_m$ (called the universe) and a set S of n sets whose union equals the universe, the set cover problem is to identify the smallest subset of S whose union equals the universe. The only parameter of the algorithm is the percentage of diversity *div* that the candidate itemsets must have with respect to the ones already selected. The main cycle of the algorithm is then over the closed itemsets, starting from the ones with the highest cardinality (i.e., the ones that cover most of the items). For each candidate itemset, if its current percentage of diversity overtakes *div*, it is added to the result set. In our experiments, we set its initial value to 0.5 (candidate itemsets must have a half of their items that are not already present in the selected itemsets). In case the insertion phase ends without having covered all the items that are contained in the input closed itemsets, the value decreases of a certain factor *alpha* (set to 0.1, in our experiments). This way, the algorithm assures its termination.

Verb Model Construction In the final phase, once obtained the frequent and diverse itemsets for both the two transactional databases, we connect all the

[9] http://www.philippe-fournier-viger.com/spmf/index.php

subject-itemsets with all the object-itemsets, weighting the connection according to the their co-occurrences in the same triples of the original dataset.

The *semantic verb model* constructed for a specific verb "v" is thus a set of weighted connections between *frequent and diverse* semantic features belonging to the subjects of "v" and *frequent and diverse* semantic features of the objects of "v". On the one hand, this is a way to summarize the semantics suggested by the verb. On the other hand, it is also a result that can be used to generate new data by querying existing semantic resources with such semantic subject- and object-itemsets. Still, this can be done without looking for words similar to frequent subjects and objects, but by finding new subjects and objects that, even if not similar in general, have certain semantic information that fill a specific context.

The resulting models are automatically calculated, and they are very concise, since in all the large and sparse semantic space only few features are relevant to certain meanings (headed by the verb). This is also in line with what stated in [26] where the authors claimed that semantics is actually structured by low-dimensionality spaces that are covered up in high-dimensional standard vector spaces.

4 Experiments and Results

In this section we present the result of the approach on different cases. In particular, we extracted all the triples in the dataset containing different verbs like *to play*, *to eat*, *to sing*, and so forth. Then, for each of these verbs we executed the algorithm and extracted the models, i.e., sets of weighted pairs of *diverse* subject- and object-itemsets. Table 4 shows some examples of the automatically extracted semantic information.

In the experiments, we wanted to evaluate the quality of the constructed models and their ability to generalize over the input data also taking into account their size in comparison with classic word-based vector spaces.

On the one hand, the approach is able to model the meanings expressed by complete verbal phrases with minimal resource requirements, as shown in Figure 1. In fact, starting from hundreds of verbal instances, the method produces itemsets with a feature space much smaller then common word spaces in which words and chunks are represented by vector spaces of the order of thousands of features. For instance, in the presented example, with a minimum support of 0.05 (i.e., 5%), the resulting model is constituted by 4 diverse itemsets for the objects and 24 for the subjects, with an average itemset cardinality of 18.5 and 12.6 respectively, covering more than 50% of the semantic features of all the input triples.

On the other hand, we calculated the *coverage* of the extracted models, that is the percentage of triples *subject-verb-object* in the input data in which at least one item is included in the extracted diverse itemsets. These results are shown in Figure 2. Notice that the coverage of the diverse itemsets is always equals to

Table 4. Examples of the main semantic information that are automatically induced for subjects and objects, tested on various verbs.

Verb	Subject semantic features	Object semantic features
to pay	*isa*-person	*relatedto*-money
to read	*isa*-person *notcapableof*-fly *desires*-clothe *capableof*-think *capableof*-love *capableof*-talk_to_each_other *desires*-privacy *partof*-society *capableof*-voice_opinion *hasa*-name	*usedfor*-read *atlocation*-library *atlocation*-newspaper
to visit	*isa*-person	*atlocation*-city *aTlocation*-museum *partof*-web_site *usedfor*-entertainment
to eat	*isa*-mammal *capableOf*-fear_death *capableOf*-cook_dinner *capableOf*-run *capableOf*-eat *capableof*-pay_bill *atLocation*-earth ...	*atlocation*-oven *usedfor*-eat *atlocation*-store *hasproperty*-delicious *atlocation*-tree *isa*-food *atlocation*-restaurant ...
to play	*notcapableof*-fly *isa*-mammal *capableof*-think *atlocation*-earth *desires*-laugh *capableof*-hear_noise *capableof*-experience_joy *partof*-society ...	*atlocation*-movie_theater *hasproperty*-fun *atlocation*-theatre *isa*-story *usedfor*-entertainment *hasproperty*-entertain *capableof*-tell_story *usedfor*-learn ...
to sing	*isa*-person *capableof*-think *capableof*-love *atLocation*-earth	*partof*-album *usedfor*-pleasure_yourself *atlocation*-record *usedfor*-have_fun *hasproperty*-melodic *usedfor*-express_feel_and_emotion *isa*-composition_of_music *createdby*-composer *atlocation*-on_cd *usedfor*-entertainment ...

the coverage of the closed itemsets, even if the formers are less than (or equal to) the latters.

To the best of our knowledge, this is the first attempt to model entire linguistic constructions *subject-verb-object* in terms of *senses* at sentence-level automatically carved out from the data by deeply analyzing co-occurrent fine-grained semantic information instead of lexical and syntactic chunks. We think that further efforts on this direction can importantly change the vision and the horizon of current Natural Language Understanding goals as well as the management of large collections of textual data with concise, generative, and semantics-based models.

5 Conclusions

This contribution represents a first effort to pass from standard word-vectors to *semantic vectors*. This causes the raise of new challenges, like the alignment

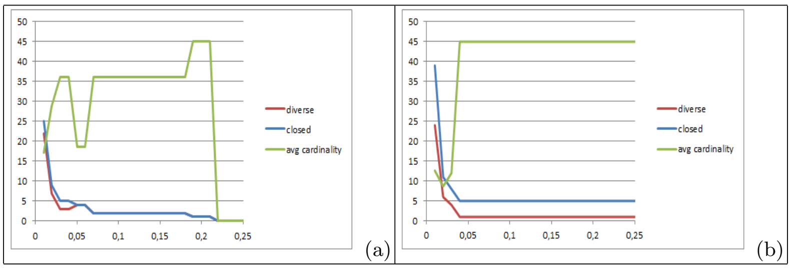

Fig. 1. Size of closed (blue line) and diverse (red line) itemsets w.r.t. minimum support, and average number of itemset cardinality (green line). The plot on the left (a) is for the subjects of the example verb "*to sing*", while the plot on the right (b) is for its direct objects.

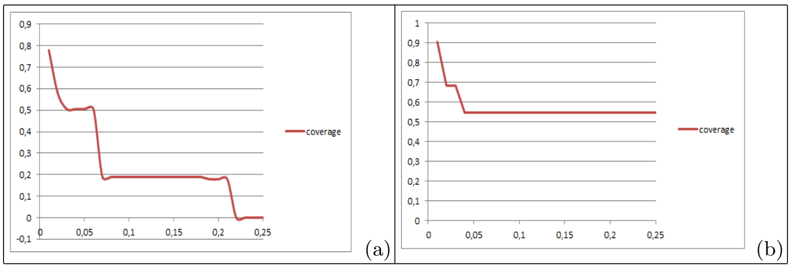

Fig. 2. Coverage w.r.t. minimum support. The plot on the left (a) is for the subjects of the example verb "*to sing*", while the plot on the right (b) is for its direct objects.

and the filtering of heterogeneous semantic information. Still, such shift in the basic bricks also concerns Data Mining techniques, since the problem of correlating linguistic items becomes to correlate sets of semantic features, where only some of them are actually significant. In this paper, we presented an approach that connect Natural Language Processing techniques (Lexico-syntactic analysis, syntactic parsing[10] and Named Entity Recognition) with Pattern Mining approaches like Frequent Itemset Mining and the cover set problem.

To produce *semantic vectors*, we started by using ConceptNet, one of the largest semantic resource currently available. In spite of this, in future work we will also come back to lexico-syntactic parsing of large corpora like Wikipedia for the extraction of further semantic information directly from text.

The impact of this new research direction can be extremely high. The main question this proposal wants to engender is the following: what if computational systems can directly reason on semantics instead of syntax? Future NLP

[10] We refer here to the used *subject-verb-object* input structures.

technologies could move away from language through more complex meaning understanding, also dealing with unseen and low-frequency words.

By reducing commonly-huge vector spaces based on linguistic items into synthetic conceptual matrices, we also attack the Big Data problem for textual databases. For example, if we think at the term "*color*", a linguistic-based vectorial representation would contain hundreds of terms that usually co-occur with it, such as "*pastel*", "*dark*", "*light*", "*red*", "*brilliant*", and so forth. In Wikipedia, for instance, we found more than 500 adjectival lemmas that co-occur with this term. On the other hand, the concept of "color" can be potentially represented by few dimensions. For instance, the HSV scheme uses only three dimensions: brightness, hue, and saturation.

We evaluated the approach by its ability to reduce the space and generalize over the input data. In future work, we will also measure the approach on tasks like Ontology Learning and Question Answering. This paper also introduces a the concept of *semantic loop*, i.e., the recursive use of extracted semantics as input for further extensions. The use of this methodology can create new and extended semantic resources.

Finally, we will leverage techniques for data compression like Multi Dimensional Scaling (MDS) [27], Principal Component Analysis (PCA) [25] and tensors decompositions to actually transform combinations of properties into reduced-spaces capturing the more significant part of the data (due to their ability to approximate information while preserving the maximum level of their expressivity). Cognitive psychology has deeply used such techniques in a wide variety of applications where the explanation of cognitive processes can be derived directly from them.

References

1. Alvanaki, F., Sebastian, M., Ramamritham, K., Weikum, G.: Enblogue: emergent topic detection in web 2.0 streams. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. pp. 1271–1274. ACM (2011)
2. Baroni, M.: Composition in distributional semantics. Language and Linguistics Compass 7(10), 511–522 (2013)
3. Baroni, M., Bernardi, R., Zamparelli, R.: Frege in space: A program for compositional distributional semantics. Submitted, draft at http://clic. cimec. unitn. it/composes (2013)
4. Baroni, M., Lenci, A.: Distributional memory: A general framework for corpus-based semantics. Computational Linguistics 36(4), 673–721 (2010)
5. Biemann, C.: Ontology learning from text: A survey of methods. In: LDV forum. vol. 20, pp. 75–93 (2005)
6. Borgelt, C.: Frequent item set mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6), 437–456 (2012)
7. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology learning from text: An overview. Ontology learning from text: Methods, evaluation and applications 123, 3–12 (2005)
8. Cabrera, J.M., Escalante, H.J., Montes-y Gómez, M.: Distributional term representations for short-text categorization. In: Computational Linguistics and Intelligent Text Processing, pp. 335–346. Springer (2013)

9. Candan, K., Di Caro, L., Sapino, M.: Creating tag hierarchies for effective navigation in social media. In: Proceedings of the 2008 ACM workshop on Search in social media. pp. 75–82. ACM (2008)
10. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining. p. 4. ACM (2010)
11. Cataldi, M., Di Caro, L., Schifanella, C.: Immex: Immersive text documents exploration system. In: Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on. pp. 1–6. IEEE (2011)
12. Cataldi, M., Schifanella, C., Candan, K.S., Sapino, M.L., Di Caro, L.: Cosena: a context-based search and navigation system. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems. p. 33. ACM (2009)
13. Chen, Y., Amiri, H., Li, Z., Chua, T.S.: Emerging topic detection for organizations from microblogs. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 43–52. ACM (2013)
14. Croce, D., Storch, V., Annesi, P., Basili, R.: Distributional compositional semantics and text similarity. In: Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on. pp. 242–249. IEEE (2012)
15. Draicchio, F., Gangemi, A., Presutti, V., Nuzzolese, A.G.: Fred: From natural language text to rdf and owl in one click. In: The Semantic Web: ESWC 2013 Satellite Events, pp. 263–267. Springer (2013)
16. Dumais, S.T.: Latent semantic analysis. Annual review of information science and technology 38(1), 188–230 (2004)
17. Ferraresi, A., Zanchetta, E., Baroni, M., Bernardini, S.: Introducing and evaluating ukwac, a very large web-derived corpus of english. In: Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google. pp. 47–54 (2008)
18. Flati, T., Navigli, R.: Spred: Large-scale harvesting of semantic predicates. In: Proceedings of 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria (2013)
19. Fortuna, B., Mladenič, D., Grobelnik, M.: Semi-automatic construction of topic ontologies. Semantics, Web and Mining pp. 121–131 (2006)
20. Gärdenfors, P.: Conceptual spaces: The geometry of thought. MIT press (2004)
21. Gibson, J.: The concept of affordances. Perceiving, acting, and knowing pp. 67–82 (1977)
22. Grefenstette, E., Sadrzadeh, M.: Experimental support for a categorical compositional distributional model of meaning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1394–1404. Association for Computational Linguistics (2011)
23. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
24. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics-Volume 2. pp. 539–545. Association for Computational Linguistics (1992)
25. Jolliffe, I.: Principal component analysis. Wiley Online Library (2005)
26. Karlgren, J., Holst, A., Sahlgren, M.: Filaments of meaning in word space. In: Advances in Information Retrieval, pp. 531–538. Springer (2008)
27. Kruskal, J.B., Wish, M.: Multidimensional scaling, vol. 11. Sage (1978)
28. Lenci, A.: Carving verb classes from corpora. Word Classes. A cura di Raffaele Simone e Francesca Masini. Amsterdam-Philadelphia: John Benjamins p. 7 (2010)
29. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)

30. Morik, K., Kaspari, A., Wurst, M., Skirzynski, M.: Multi-objective frequent termset clustering. Knowledge and information systems 30(3), 715–738 (2012)
31. Navigli, R., Velardi, P., Faralli, S.: A graph-based algorithm for inducing lexical taxonomies from scratch. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three. pp. 1872–1877. AAAI Press (2011)
32. Navigli, R., Ponzetto, S.P.: Babelnet: Building a very large multilingual semantic network. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 216–225. Association for Computational Linguistics (2010)
33. Nuzzolese, A.G., Gangemi, A., Presutti, V., Draicchio, F., Musetti, A., Ciancarini, P.: Tìpalo: A tool for automatic typing of dbpedia entities. In: The Semantic Web: ESWC 2013 Satellite Events, pp. 253–257. Springer (2013)
34. Padó, S., Lapata, M.: Dependency-based construction of semantic space models. Computational Linguistics 33(2), 161–199 (2007)
35. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: Database TheoryICDT99, pp. 398–416. Springer (1999)
36. Pilehvar, M.T., Jurgens, D., Navigli, R.: Align, disambiguate and walk: A unified approach for measuring semantic similarity. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013) (2013)
37. Ponzetto, S., Strube, M.: Deriving a large scale taxonomy from wikipedia. In: Proceedings of the national conference on artificial intelligence. vol. 22, p. 1440. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2007)
38. Rosch, E.: Principles of categorization. Concepts: core readings pp. 189–206 (1999)
39. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (Nov 1975), `http://doi.acm.org/10.1145/361219.361220`
40. Sigurbjörnsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th international conference on World Wide Web. pp. 327–336. ACM (2008)
41. Slavík, P.: A tight analysis of the greedy algorithm for set cover. In: Proceedings of the twenty-eighth annual ACM symposium on Theory of computing. pp. 435–441. ACM (1996)
42. Snow, R., Jurafsky, D., Ng, A.: Learning syntactic patterns for automatic hypernym discovery. Advances in Neural Information Processing Systems 17 (2004)
43. Speer, R., Havasi, C.: Representing general relational knowledge in conceptnet 5. In: LREC. pp. 3679–3686 (2012)
44. Speer, R., Havasi, C.: Conceptnet 5: A large semantic network for relational knowledge. In: The Peoples Web Meets NLP, pp. 161–176. Springer (2013)
45. Wu, F., Hoffmann, R., Weld, D.S.: Information extraction from wikipedia: Moving down the long tail. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 731–739. ACM (2008)
46. Zaki, M.J., Hsiao, C.J.: Charm: An efficient algorithm for closed itemset mining. In: SDM. vol. 2, pp. 457–473. SIAM (2002)