

From a Lexical to a Semantic Distributional Hypothesis

Luigi Di Caro[†], Guido Boella[†], Alice Ruggeri[†],
Loredana Cupi[†], John Adebayo Kolawole[‡], Livio Robaldo^{*}

[†] University of Torino

^{*} University of Luxembourg

[‡] University of Bologna

{dicaro, boella, ruggeri}@di.unito.it

loredana.cupi@unito.it, kolawolejohn.adebayo@unibo.it

livio.robaldo@uni.lu

Abstract

English. Distributional Semantics is based on the idea of extracting *semantic* information from *lexical* information in (multilingual) corpora using statistical algorithms. This paper presents the challenging aim of the *SemBurst* research project¹ which applies distributional methods not only to words, but to sets of semantic information taken from existing semantic resources and associated with words in syntactic contexts. The idea is to inject semantics into vector space models to find correlations between statements (rather than between words). The proposal may have strong impact on key applications such as Word Sense Disambiguation, Textual Entailment, and others.

Italiano. *La semantica distribuzionale si basa sull'idea di estrarre automaticamente informazione semantica attraverso algoritmi statistici applicati ad occorrenze lessicali in grandi corpora. Questo articolo presenta l'idea del progetto SemBurst che applica metodi distribuzionali non solo alle parole, ma ad insiemi di informazioni semantiche tratte da risorse semantiche disponibili e associate alle parole nei relativi contesti sintattici. Lo scopo è quello infatti di iniettare semantica negli spazi vettoriali per trovare correlazioni tra informazioni semantiche (piuttosto che tra elementi lessicali). Questo nuovo approccio potrà avere un alto impatto su applicazioni chiave come Word Sense Disambiguation, Textual Entailment, e altri.*

1 Introduction and Background

One of the main current research frontiers in Computational Linguistics is represented by studies and techniques usually associated with the label Distributional Semantics (DS), which are focused on the exploitation of distributional analyses of words in syntactic compositions. Their importance is demonstrated by recent ERC projects (COMPOSES and DisCoTex²) and by a growing research interest in the scientific community³. The proposal presented in this paper is about going far beyond this state of the art.

DS uses traditional Data Mining (DM) techniques on text, considering language as a grammar-based type of data, instead of simple unstructured sequences of tokens. It quantifies semantic (in truth lexical) similarities between linguistically-refined tokens (words, lemmas, parts-of-speech, etc.), based on their distributional properties in large corpora. DM relies on Vector Space Models (VSMs), a representation of textual information as vectors of numeric values (Salton et al., 1975). DM techniques such as Latent Semantic Analysis (LSA) have been successfully applied to text for information indexing and extraction tasks, using matrix decompositions such as Singular Value Decomposition (SVD) to reconstruct the latent structure behind the distributional hypothesis (Deerwester et al., 1990). It usually works by evaluating the relatedness of different terms, forming word clusters sharing similar contexts. Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) and Salient Semantic Analysis (SSA) (Hassan and Mihalcea, 2011) revisits these methods in the way they define the conceptual layer. With LSA a word's hidden

¹Semantic Burst: Embodying Semantic Resources in Vector Space Models, financed by Compagnia di San Paolo - cod. 2014.L1.272.

²European Research Council projects nr. 283554 (COMPOSES) and nr. 306920 (DisCoTex).

³Clark, S. Vector space models of lexical meaning. A draft chapter of the Wiley-Blackwell Handbook of Contemporary Semantics second edition.

concept is based on its surrounding words, with ESA it is based on Wikipedia entries, and with SSA it is based on hyperlinked words in Wikipedia entries. These approaches represent only a partial step towards the use of semantic information as input for Distributional Analysis.

While distributional representations excel at modelling lexical semantic phenomena such as word similarity and categorization (conceptual aspect), Formal Semantics in Computational Linguistics focuses on the representation of the meaning in a set theoretic way (functional aspect), providing a systematic treatment of compositionality and reasoning. Recent interest in the combination of Formal Semantics and Distributional Semantics have been proposed (Lewis and Steedman, 2013) (Turney, 2012) (Garrette et al., 2014), that employ approaches based on the lexical level. However, 1) the problem of compositionality of lexical distributional vectors is still open and the proposed solutions are limited to combination of vectors, 2) reasoning on classic distributional representations is not possible, since they are VSMs at the lexical level only, 3) the connection of DS with traditional Formal Semantics is not straightforward (Turney, 2012) (Garrette et al., 2014) since DS is limited to a semantics of similarity which is able to support retrieval but not other aspects such as reasoning; and 4) DS does not scale up to phrases and sentences due to data sparseness and growth in model size (Turney, 2012), restraining the use of tensors.

2 A Semantic Distributional Hypothesis

This proposal is based on the idea of applying distributional analysis not only to words but also to sets of semantic features taken from semantic resources. The idea is that the semantic information injected into an input text corpus will act as a catalyst to facilitate the creation of further semantic information and to find correlations with semantic features of other words in their syntactic context. For instance, the word “cat” in “*the cat bites the mouse*” will be replaced by physical facts (it has claws, paws, eyes, whiskers, etc.), behavioural information (it chases mice, it is capable of climbing up a tree, etc.), taxonomical information (it is a feline, it is a predator, etc.), habitats, etc. This will create a new multi-dimensional semantic search space where distributional analysis will be used to clean up and correlate statements rather than words, for example, finding the relation

between a carnivore-subject and a meat-object in the sentence “*The cat bites the mouse*” or between a cat’s claws and the act of climbing in the sentence “*The cat climbs the tree*”.

2.1 Feasibility

The proposed shift to *semantics* as input for distributional analysis is now feasible due to the large number of semantic resources available such as BabelNet (Navigli and Ponzetto, 2010), ConceptNet (Speer and Havasi, 2012), FrameNet (Baker et al., 1998), DBPedia, etc. However, these resources are sometimes incomplete, contradictory, ambiguous, and difficult to integrate together, so they cannot be used in Formal Semantics. Formal Semantics handles reasoning, quantification, and compositionality of meaning using set-theoretic models, and therefore requires data consistency. The aim of this proposal is to overcome these problems by applying the distributional hypothesis to the partial and contradictory semantic information that can be associated with words contained in large corpora and structured in syntactic contexts, in the same way it has been successfully applied to words in the last few decades. For example, if a corpus contains ambiguities and other noise, this does not prevent distributional analysis on words, because the calculations use the most significant data. Analogously, in case of a few ambiguities and contradictions in the semantic resources, a distributional approach using several resources and advances in Data Mining will manage to derive the most probable relations between statements.

2.2 Research Objectives

The presented approach is intended to reduce the existing gap between Distributional Semantics and Formal Semantics by creating a novel type of semantics, still distributional, but working on a semantic rather than lexical input. The idea is articulated in the following sub-objectives: 1) to acquire and integrate semantic information from different resources; 2) to create not only distributional word representations, but also distributional representations of semantic features with tensors. By moving to the semantic level, it will help overcome the problem of sparseness in classic word-based tensors. Since semantic information represents knowledge shared by multiple words, this proposal will allow to consider more complex syntactic structures to be considered than currently practiced. Then, it aims to 3) deal with compo-

sitionality at a more appropriate level - no longer as a fusion of lexical distribution vectors, but as a fusion of semantic features and 4) will enable reasoning on the semantic representation built via distributional vectors of semantic features. Further semantic resources will be created, which can be re-injected several times as input into the distributional analysis, thus 5) creating a positive loop of expanding knowledge. The proposal can also 6) consider multilingual contexts where semantic resources are not available, 7) finally reframing tasks as later described in Section 3.6.

3 Project Architecture

3.1 Data Acquisition

The first step required by this proposal is to aggregate linguistic and semantic resources such as ConceptNet, FrameNet, WordNet, BabelNet, etc. The result will be a semantic database (*SDB*) of lexical and semantic information. This will require integration of data from different sources with problems such as alignment, conflict resolution, and granularity mismatch. The second step regards the expansion of an input corpus (result of a selection from existing available corpora) with the semantic information contained in *SDB* for each of its words. Let us assume a word w_i in *SDB* can be associated with a set $\sigma_i = \{ \langle rel_a, c_1 \rangle, \langle rel_b, c_2 \rangle, \dots, \langle rel_k, c_n \rangle \}$ of semantic features of the type $\langle rel, c_j \rangle$ to mean that word w_i has a relation rel with concept c_j in some semantic resource (e.g., $w_i = cat$ and $\sigma_i = \{ \langle isA, MAMMAL \rangle, \langle capableOf, JUMP \rangle, \dots \}$). This word-by-facts replacement can be iterated multiple times over the concepts in σ_i (e.g., $c_j = MAMMAL$ in σ_i can enrich σ_i itself to σ_k with σ_{MAMMAL} such that $\sigma_k = \sigma_i \cup \{ \langle isA, ANIMAL \rangle, \langle capableOf, BREATH \rangle, \dots \}$). Given a sentence S , the idea is to enrich each w_i with σ_i so that we build a different and richer input for distributional analysis than in traditional approaches (see Figure 1).

3.2 Distributional Analysis of Semantics

The second part of the proposal concerns the use of advanced DM techniques such as tensor-based representations (of semantics, rather than words) by embodying syntactic roles (subjects, modifiers, verbs, and arguments) into its dimensions (see Figure 1). The complexity of algorithms for tensors is a major challenge in this level, although recent re-

search has shown that background information can improve this issue (Schifanella et al., 2014). Advanced data analysis techniques on tensors allow operations that are suitable for the aim of this ongoing research project. In particular, the problem of correlating lexical items will be reframed as the problem of correlating sets of semantic features within syntactic structures, using similarity and correlation measures over tensors to align, merge and filter data items.

3.3 Compositionality and reasoning

The proposal allows to address the compositionality problem at a semantic level. Let us consider the adjective-noun collocation “*dead parrot*”. Parrots are pets, but dead parrots are not. This is an example of complicated compositionality (Kruszewski and Baroni, 2014). Unlike e.g., “*blue parrot*”, the adjective overrides typical features of the noun it is associated with. Currently, Distributional Semantics uses to model compositionality by merging word distribution vectors (Mitchell and Lapata, 2008; Grefenstette and Sadrzadeh, 2011), hopefully lowering the frequency of collocations where the phrase “*dead parrot*” occurs as a pet. In our approach, instead, we reframe the problem as: how can distributional analysis handle the fact that the semantic feature $\langle hasProperty, NOT-ALIVE \rangle$ associated with the word “*dead*” overrides the necessary feature of the role of pet ($\langle isA, PET \rangle$), i.e., $\langle hasProperty, ALIVE \rangle$ played by “*parrot*”? Moreover, we can apply reasoning on the resulting semantic representation of “*dead parrot*”: since the property NOT-ALIVE in semantic resources is associated with $\langle hasProperty, NO-MOVE \rangle$, we can also predict that, for example, “*the dead parrot flies*” is not a proper sentence since FLY in $\langle capableOf, FLY \rangle$ is associated with $\langle isA, MOVE \rangle$.

3.4 Extension of Semantic Resources

A distributional analysis over the acquired semantic information can create novel semantic resources with the following radically new aspects. First, semantics will assume the form of combinations of statements within syntactic contexts, thus generalizing over concepts which could not be found even in very large corpora. Assume, for instance, that “*cat*” is not associated with the semantic feature $\langle has, CLAWS \rangle$: we can add this feature to the word “*cat*” if it occurs in contexts where the distinguishing feature for climb-

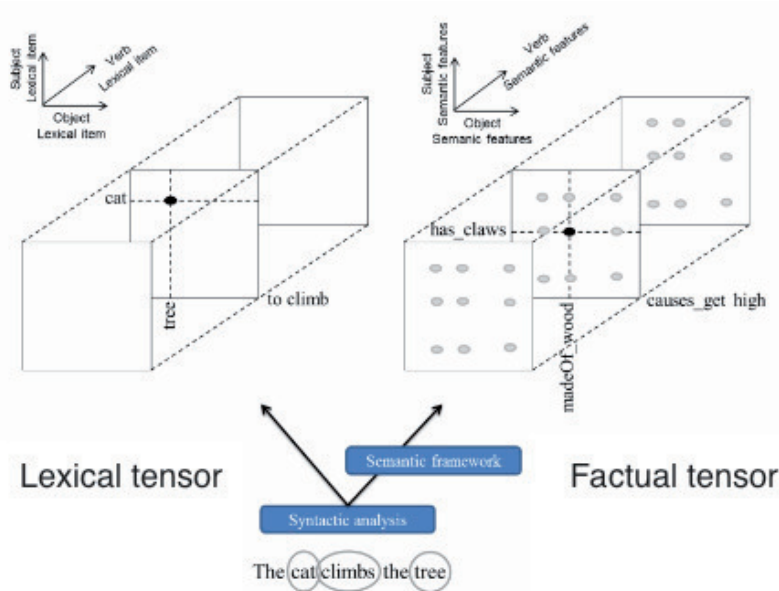


Figure 1: Distributional representation of natural language based on statements rather than lexical items.

ing is using claws (“the * climbs the mast”, “the * climbs the curtains”, etc.); moreover, the extended resources will be used again thus creating a positive loop of semantic feedback.

3.5 Multilingual Mapping

Multilingualism can be better managed since semantic features represent conceptual rather than lexical information units. When semantic resources are missing in one language, the proposed approach will use those of the English language, using automated translation from the target language to English. Ambiguities and errors will be introduced, but analyses on large numbers will hopefully manage the situation, allowing the creation of semantic knowledge for new languages.

3.6 Exploitation

Word Sense Disambiguation (WSD). Instead of linking words to word senses (a priori defined in resources such as WordNet) by exploring word-based contexts, we will replace each word with all the semantic features of all its uses in the corpora, clustering features and disambiguating by matching the word features with those of other words in the syntactic structure using the result of the semantic analysis (see Section 3.2).

Parsing. Syntactic parsing is a procedure that requires semantic information (e.g., to understand which phrase in the parse tree a modifier should be associated with). This approach will alleviate ambiguity problems at syntactic level by using the

semantics extracted by the distributional approach over the semantic features.

Information Retrieval (IR). By using the proposed approach, computational systems can process complex queries and improve precision and recall of relevant documents. The aim is to go beyond the state of the art in query expansion by combining similar semantic features in accordance with the syntactic structure, rather than using bag-of-words approach, synonyms and paraphrases.

Textual Entailment (TE). Current research in TE attempts to solve the problem of implicit meaning in texts by lexical inference (e.g., *selling* implies *owning*), using resources (e.g., WordNet), distributional semantics and similarity measures. However, these techniques still operate at lexical level. This proposal operates at a semantic rather than lexical level which brings out the implicit meanings sought by other means in TE research.

Generation and Summarization. This proposal will enable the generation of lexical compositions reflecting plausible combinations of semantic features instead of lexical substitutions. This will open a completely new horizon of summarization results.

4 Conclusions

This paper presents a recently-funded project on a research frontier in Computational Linguistics. It includes a brief survey on the topic and the essential keys of the proposal with its impact.

References

- Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.
- Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.
- Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.
- Dan Garrette, Katrin Erk, and Raymond Mooney. 2014. A formal approach to linking logical form and vector-space lexical semantics. In *Computing Meaning*, pages 27–48. Springer.
- Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.
- Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *AAAI*.
- Germán Kruszewski and Marco Baroni. 2014. Dead parrots make bad pets: Exploring modifier effects in noun phrases. *Lexical and Computational Semantics (* SEM 2014)*, page 171.
- Mike Lewis and Mark Steedman. 2013. Combining distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.
- Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.
- 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.
- Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Claudio Schifanella, K Selçuk Candan, and Maria Luisa Sapino. 2014. Multiresolution tensor decompositions with mode hierarchies. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):10.
- Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.
- Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, pages 533–585.