Representing Verbs with Rich Contexts: an Evaluation on Verb Similarity

Emmanuele Chersoni  
Aix-Marseille University  
France  
emmanuelecchersoni@gmail.com

Enrico Santus  
The Hong Kong Polytechnic University  
Hong Kong  
esantus@gmail.com

Alessandro Lenci  
University of Pisa  
Italy  
alessandro.lenci@unipi.it

Philippe Blache  
CNRS - Aix-Marseille University  
France  
blache@lpl-aix.fr

Chu-Ren Huang  
The Hong Kong Polytechnic University  
Hong Kong  
churen.huang@polyu.edu.hk

Abstract

Several studies on sentence processing suggest that the mental lexicon keeps track of the mutual expectations between words. Current DSMs, however, represent context words as separate features, which causes the loss of important information for word expectations, such as word order and interrelations. In this paper, we present a DSM which addresses the issue by defining verb contexts as joint dependencies. We test our representation in a verb similarity task on two datasets, showing that joint contexts are more efficient than single dependencies, even with a relatively small amount of training data.

1 Introduction

Distributional Semantic Models (DSMs) rely on the Distributional Hypothesis (Harris, 1954; Firth, 1957; Sahlgren, 2008), stating that words occurring in similar contexts have similar meanings. On such theoretical grounds, words co-occurrences extracted from corpora are used to build semantic representations, in the form of vectors, which have become very popular in the NLP community. Proximity between word vectors is taken as an index of meaning similarity, and vector cosine is generally adopted to measure such proximity, even though other measures have been proposed (Weeds et al., 2004).

Most of DSMs adopt a bag-of-words approach, that is they turn a text span (i.e. a word window) into a set of words and they register separately the co-occurrence of each word with a given target. A major problem of this representation is that information concerning word order and interrelations gets lost. This is why works like Ruiz-Casado et al. (2005), Agirre et al. (2009), and Melamud et al. (2014) proposed to introduce richer contexts in distributional spaces, i.e. using entire word windows as features. These richer contexts proved to be helpful to semantically represent verbs, which are characterized by complex argument structures and highly context-sensitive meanings. Nonetheless, they suffer of data sparsity, therefore requiring either larger corpora or complex smoothing processes.

In this paper, we propose a syntactically savvy notion of joint contexts. To test our representation, we implement several DSMs and we evaluate them in a verb similarity task on two datasets. The results show that, even using a corpus of limited size, our syntactic joint contexts are more robust with respect to data sparseness and perform better than single dependencies in a wider range of parameter settings.
The paper is organized as follows. In Section 2, we provide a psycholinguistic and a computational background, describing recent models based on word windows. In Section 3, we describe our reinterpretation of joint contexts with syntactic dependencies. Evaluation settings and results are presented in Section 4.

2 Related Work

A number of studies in sentence processing suggest that verbs activate expectations on their typical argument nouns and vice versa (McRae et al., 1998; McRae et al., 2005) and nouns do the same with other nouns occurring as co-arguments in the same events (Hare et al., 2009; Bicknell et al., 2010). Experimental subjects seem to exploit a rich event knowledge to activate or inhibit dynamically the representations of the potential arguments. This phenomenon, generally referred to as thematic fit (McRae et al., 1998; Matsuki et al., 2011), supports the idea of a mental lexicon arranged as a web of mutual expectations.

Some recent works, like (Baroni and Lenci, 2010; Lenci, 2011; Sayeed and Demberg, 2014; Greenberg et al., 2015), modeled thematic fit estimations by means of dependency-based or of thematic roles-based DSMs. However, these semantic spaces are built similarly to traditional DSMs as they split verb arguments into separate vector dimensions. By using syntactic-semantic links, they encode the relation between an event and each of its participants; but they do not encode directly the relation between participants co-occurring in the same event.

On the other hand, some research efforts in the NLP community aimed at introducing richer contextual features in DSMs, mostly based on word windows. A first example was the composite-feature model by Ruiz-Casado et al. (2005), who extracted word windows through a Web Search engine. They motivated such choice to reduce data sparseness. A composite feature for the target word watches is Alicia always roman- tic movies, extracted from the sentence I heard that Alicia always watches romantic movies with Antony (the placeholder represents the target position). Thanks to this approach, Ruiz-Casado and colleagues achieved 82.50 in the TOEFL synonym detection test, outperforming the Latent Semantic Analysis (LSA; see (Landauer and Dumais, 1997; Landauer et al., 1998)) and several other methods.

Agirre et al. (2009) adopted an analogous approach, relying on a huge learning corpus (1.6 Teraword) to build composite-feature vectors. Their model outperformed a traditional DSM on the similarity subset of the WordSim-353 test set (Finkelstein et al., 2001).

Melamud et al. (2014) introduced a probabilistic similarity scheme for modeling the so-called joint context. By making use of the Kneser-Ney language model (Kneser and Ney, 1995) and of a probabilistic distributional measure, they were able to overcome data sparsity, outperforming a wide variety of DSMs on two similarity tasks, evaluated on VerbSim (Yang and Powers, 2006) and on a set of 1,000 verbs extracted from WordNet (Fellbaum, 1998). On the basis of their results, the authors claimed that composite-feature models are particularly advantageous for measuring verb similarity.

Finally, in a recent paper, Melamud et al. (2015) represented contexts as substitute vectors, which are second-order potential filler words for the first-order target word slots (which are in turn word windows, such as ‘I ___ my phone’, whose substitute vector may contain: love, lost, upgraded etc.). Using such representation, the authors outperformed state-of-the-art models in both predicting and ranking paraphrases of words in context.

3 Syntactic Joint Contexts

A joint context, as defined in Melamud et al. (2014), is a word window of order $n$ around a target word. The target is replaced by a placeholder, and the value of the feature for a word $w$ is the probability of $w$ to fill the placeholder position. Assuming $n=3$, a word like love would be represented by a collection of contexts such as the new students ___ the school campus, my cousin Bob ___ to play basketball etc. Such representation introduces data sparseness, which has been addressed by previous studies either by adopting huge corpora or by relying on n-gram language models to approximate the probabilities of long sequences of words.

However, features based on word windows do not
guarantee to include all the most salient event participants. Moreover, they could include unrelated words, also differentiating contexts substantially describing the same event (e.g. consider Luis the red ball and Luis the blue ball).

For these reasons, we introduce the notion of syntactic joint context, further abstracting from linear word windows by using syntactic dependencies. Our assumption is that knowledge about typical event participants can be inferred by observing the most frequent argument combinations. Below, we describe the procedure we followed to define our syntactic joint contexts.

- First, we extracted a list of verb-argument dependencies from a parsed corpus. For each target, we extracted all the direct dependencies from the sentence in which it occurs. For instance, in Finally, the dictator acknowledged his failure, we will have: target = 'acknowledge-v'; subject = 'dictator-n'; and object = 'failure-n'.

- Second, for each sentence, we generated a joint context feature by joining all the dependencies for the grammatical relations of interest. From the example above, we would generate the feature dictator-n.subj+ + failure-n.obj.

For our experiments, the grammatical relations that we used are subject, object and complement, where complement is meant as a generic relation grouping together all complements introduced by a preposition. During our experimentations, we decided to represent those complements just through the preposition and to omit the noun arguments, in order to reduce data sparsity.

Finally, our distributional representation for a target word is a vector of joint features, such as astronaut-n.subj+ + on-p.complement+, athlete-n.subj+ + medal-n.obj.

4 Evaluation

4.1 Corpus and DSMs

We trained our DSMs on the RCV1 corpus, which contains approximately 150 million words (Lewis et al., 2004). The corpus was tagged with the tagger described in (Dell’Orletta, 2009) and dependency-parsed with DeSR (Attardi et al., 2009). RCV1 was chosen for two reasons: i) to allow a direct comparison with the results reported by Melamud et al., 2014; ii) to show that our joint context-based representation can deal with data sparseness even with a training corpus of limited size.

All DSMs adopt Positive Pointwise Mutual Information (PPMI; (Church and Hanks, 1990)) as a context weighting scheme and vary according to three main parameters: i) type of contexts; ii) number of dimensions; iii) application of Singular Value Decomposition (SVD; see (Landauer and Dumais, 1997, Landauer et al., 1998)).

For what concerns i), beyond the joint context-based DSMs, we have also developed DSMs with single dependencies, to be used as baselines. ii) refers instead to the number of contexts that have been used as vector dimensions. Several values were explored (i.e. 10K, 50K and 100K), selecting the contexts according to their frequency. Finally, iii) concerns the application of SVD to reduce the matrix. We report only the results for a number of latent dimensions ranging from 200 to 400, since the performance drops significantly out of this interval.

All the models have been implemented by means of the DISSECT toolkit (Dinu et al., 2013).

4.2 Similarity Measures

As a similarity measure, we used vector cosine, which is by far the most popular in the literature (Turney and Pantel, 2010). Melamud et al. (2014) have proposed the Probabilistic Distributional Similarity (PDS), which is based on the intuition that two words, w1 and w2, are similar if they are likely to occur in each other’s contexts. The authors define it such that it assigns a high similarity score when both \( p(w1 \mid \text{contexts of } w2) \) and \( p(w2 \mid \text{contexts of } w1) \) are high. We tried to test variations of this measure with our syntactic joint contexts, but we were not able to achieve satisfying results. Therefore, we have decided to report here only the scores with the cosine.

4.3 Datasets

The DSMs are evaluated on two test sets: VerbSim (Yang and Powers, 2006) and the verb subset of the
popular SimLex-999 (Hill et al., 2015). The former includes 130 verb pairs, while the latter includes 222 verb pairs. We focus on verb similarity because verb meaning is highly context sensitive and, therefore, verb representation should benefit more from the introduction of joint features (Melamud et al., 2014).

Both datasets are annotated with similarity judgements, so we measured the Spearman correlation between them and the scores assigned by the model. The Verbsim dataset allows for direct comparison with Melamud et al. (2014), since they also evaluated their model on this test set, achieving a Spearman correlation score of 0.616 and outperforming all the baseline methods.

The verb subset of Simlex-999, at the best of our knowledge, has never been used as a benchmark dataset for verb similarity. The Simlex dataset is known for being quite challenging: as reported by Hill et al. (2015), the average performances of similarity models on this dataset are much lower than on alternative benchmarks like WordSim (Finkelstein et al., 2001) and MEN (Bruni et al., 2014). Considering also that verbs are generally the toughest part-of-speech for distributional modeling, SimLex verbs should represent a nontrivial task.

We exclude from the evaluation datasets all the target words occurring less than 100 times in our corpus. Consequently, we have coverage for 107 pairs in the VerbSim dataset (82.3, the same of Melamud et al. (2014)) and for 214 pairs in the SimLex verbs dataset (96.3).

### 4.4 Results

Table 1 reports the Spearman correlation scores for the vector cosine on several DSMs, implementing single dependencies and syntactic joint contexts. At a glance, we can see the discrepancy between the results obtained in the two datasets, as Simlex verbs confirms to be very difficult to model. However, in both datasets we can recognize a trend related to the number of contexts taken into account: the higher the number, the higher the performance. Single dependencies and joint context perform very similarly, and no one has an edge on the other. It is noticeable that the score obtained on VerbSim by the joint context model with 100K dimensions goes very close to the result reported by Melamud et al. (2014) (0.616).

A clear advantage of the joint contexts can be noticed instead in Table 2 and Table 3, when SVD is applied to reduce the matrix. Independently of $k$, the joint contexts almost always outperform the models using single dependencies. Overall, the performance of the joint contexts seems to be more stable across several parameter configurations, whereas single dependencies are subject to bigger performance drops. Exceptions can be noticed only for the VerbSim dataset, and only with a low number of dimensions.

The best performances are obtained with SVD reduction and $k=200$. In VerbSim, we achieve 0.650, which is above the result of Melamud et al. (2014). In the verb subset of SimLex-999, we obtain 0.283, which is very close to the score obtained by Hill et al. (2015) by applying word2vec (Mikolov et al., 2013) on the full dataset (0.282). Such performance confirms the validity of the representation, leaving the door open to future investigations.

| Model        | VerbSim | SimLex-999 |
|--------------|---------|------------|
| Single - 10k | 0.561   | 0.090      |
| Joint - 10k  | 0.568   | 0.105      |
| Single - 50k | 0.592   | 0.115      |
| Joint - 50k  | 0.592   | 0.105      |
| Single - 100k| 0.587   | 0.132      |
| Joint - 100k | 0.607   | 0.114      |

Table 1: Spearman correlation scores for VerbSim and for the verb subset of SimLex-999. The first column states whether the DSM is implementing single dependencies or joint contexts. The number of contexts taken into account is also reported.

| Model        | $k = 200$ | $k = 300$ | $k = 400$ |
|--------------|-----------|-----------|-----------|
| Single - 10k | 0.168     | 0.172     | 0.165     |
| Joint - 10k  | 0.190     | 0.177     | 0.181     |
| Single - 50k | 0.218     | 0.228     | 0.222     |
| Joint - 50k  | 0.256     | 0.250     | 0.227     |
| Single - 100k| 0.225     | 0.218     | 0.199     |
| Joint - 100k | 0.283     | 0.256     | 0.222     |

Table 2: Spearman correlation scores for the verb subset of SimLex-999, after the application of SVD with different values of $k$. The first column states whether the DSM is implementing single dependencies or joint contexts and the number of contexts.
Table 3: Spearman correlation scores for VerbSim, after the application of SVD with different values of \( k \). The first column states whether the DSM is implementing single dependencies or joint contexts and the number of contexts.

4.5 Conclusions

In this paper, we have presented our proposal for a new type of vector representation based on joint features, which should emulate more closely the general knowledge about event participants that seems to be the organizing principle of our mental lexicon. A core issue of previous studies was the data sparseness challenge, and we coped with it by means of a more abstract, syntactic notion of joint context.

The models using joint dependencies proved to be competitive, and were able to perform comparably to - and in some cases even better than - some of the best results reported in the literature. More importantly, they were able to outperform the models based on single dependencies across several parameter settings, especially after the application of SVD. Similarly, Agirre et al. (2009) showed that large word windows had a higher discriminative power than independent features, but they did it by using a huge training corpus. The fact that we achieved our results with a small corpus like RCV1 strengthen our belief that syntactic dependencies are a possible solution for the data sparsity problem of joint feature spaces.

The performance in the verb similarity task motivates us to further test this representation on a larger range of tasks, such as word sense disambiguation, textual entailment and classification of semantic relations. Moreover, our proposal opens interesting perspectives for computational psycholinguistics, especially for modeling those semantic phenomena that are inherently related to the activation of event knowledge and argument structure information (e.g. thematic fit).

References

[Agirre et al.2009] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pašca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In Proceedings of the 2009 conference of the NAACL-HLT, pages 19–27. Association for Computational Linguistics.

[Attardi et al.2009] Giuseppe Attardi, Felice Dell’Orletta, Maria Simi, and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In Proceedings of EVALITA, 9.

[Baroni and Lenci2010] Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4):673–721.

[Bicknell et al.2010] Klinton Bicknell, Jeffrey L. Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of event knowledge in processing verbal arguments. Journal of Memory and Language, 63(4):489–505.

[Bruni et al.2014] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. J. Artif. Intell. Res.(JAIR), 49(1-47).

[Church and Hanks1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. Computational linguistics, 16(1):22–29.

[Dell’Orletta2009] Felice Dell’Orletta. 2009. Ensemble system for part-of-speech tagging. In Proceedings of EVALITA, 9.

[Dinu et al.2013] Georgiana Dinu, Nghia The Pham and Marco Baroni. 2013. DISSECT-DISTRIBUTIONAL SEMANTICS Composition Toolkit. Proceedings of the System Demonstrations of ACL 2013, East Stroudsburg PA, USA.

[Fellbaum1998] Christiane Fellbaum. 1998. WordNet. Wiley Online Library.

[Finkelstein et al.2001] Lev Finkelstein, Evgeny Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In Proceedings of the 10th international conference on World Wide Web, pages 406–414. ACM.

[Firth 1957] John Rupert Firth. 1957. Papers in linguistics, 1934–1951. Oxford University Press.

[Greenberg et al.2015] Clayton Greenberg, Asad Sayeed, and Vera Demberg. 2015. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In Proceedings of the 2015 conference of the NAACL-HLT, Denver, USA.

[Hare et al.2009] Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Activating event knowledge. Cognition, 111(2):151–167.
[Harris1954] Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

[Hill et al.2015] Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

[Kneser and Ney1995] Reinhard Kneser and Hermann Ney. 1995. Improved back-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

[Landauer and Dumais1997] Thomas K Landauer and Susan Dumais. 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2).

[Landauer et al.1998] Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

[Lenci2011] Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66. Association for Computational Linguistics.

[Lewis et al.2004] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397.

[Matsuki et al.2011] Kazunaga Matsuki, Tracy Chow, Mary Hare, Jeffrey L Elman, Christoph Scheepers and Ken McRae. 2011. Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology*, 37(4).

[McRae et al.1998] Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

[McRae et al.2004] Ken McRae, Mary Hare, Jeffrey L Elman, and Todd Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33(7):1174–1184.

[Melamud et al.2015] Oren Melamud, Ido Dagan, Jacob Goldberger. 2015. Modeling Word Meaning in Context with Substitute Vectors. In *Proceedings of the 2015 conference of the NAACL-HLT, Denver, USA*.

[Melamud et al.2014] Oren Melamud, Ido Dagan, Jacob Goldberger, Idan Szpektor, and Deniz Yuret. 2014. Probabilistic modeling of joint-context in distributional similarity. In *CoNLL*, pages 181–190.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.378*.

[Ruiz-Casado et al.2005] Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Using context-window overlapping in synonym discovery and ontology extension. In *Proceedings of RANLP*, pages 1–7.

[Sahlgren2008] Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

[Sayed and Demberg2014] Asad Sayeed and Vera Demberg. 2014. Combining unsupervised syntactic and semantic models of thematic fit. In *Proceedings of the first Italian Conference on Computational Linguistics (CLiC-it 2014)*.

[Turney and Pantel2010] Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

[Weeds et al.2004] Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1015. Association for Computational Linguistics.

[Yang and Powers2006] Dongqiang Yang and David MW Powers. 2006. *Verb similarity on the taxonomy of WordNet*. Masaryk University.