A Markovian approach to distributional semantics with application to semantic compositionality

Edouard Grave, Guillaume Obozinski, Francis Bach

A Markovian approach to distributional semantics with application to semantic compositionality

Édouard Grave
EECS Department
UC Berkeley
grave@berkeley.edu

Guillaume Obozinski
LIGM – Université Paris-Est
École des Ponts – ParisTech
guillaume.obozinski
@imagine.enpc.fr

Francis Bach
Inria – Sierra project-team
École Normale Supérieure
francis.bach@ens.fr

Abstract

In this article, we describe a new approach to distributional semantics. This approach relies on a generative model of sentences with latent variables, which takes the syntax into account by using syntactic dependency trees. Words are then represented as posterior distributions over those latent classes, and the model allows to naturally obtain in-context and out-of-context word representations, which are comparable. We train our model on a large corpus and demonstrate the compositionality capabilities of our approach on different datasets.

1 Introduction

It is often considered that words appearing in similar contexts tend to have similar meaning (Harris, 1954). This idea, known as the *distributional hypothesis* was famously summarized by Firth (1957) as follow: "you shall know a word by the company it keeps." The distributional hypothesis has been applied in computational linguistics in order to automatically build word representations that capture their meaning. For example, simple distributional information about words, such as co-occurence counts, can be extracted from a large text corpus, and used to build a vectorial representation of words (Lund and Burgess, 1996; Landauer and Dumais, 1997). According to the distributional hypothesis, two words having similar vectorial representations must have similar meanings. It is thus possible and easy to compare words using their vectorial representations.

In natural languages, sentences are formed by the *composition* of simpler elements: words. It is thus reasonable to assume that the meaning of a sentence is determined by combining the meanings of its parts and the syntactic relations between them. This principle, often attributed to the German logician Frege, is known as *semantic compositionality*. Recently, researchers in computational linguistics started to investigate how the principle of compositionality could be applied to distributional models of semantics (Clark and Pulman, 2007; Mitchell and Lapata, 2008). Given the representations of individual words, such as *federal* and *agency*, is it possible to combine them in order to obtain a representation capturing the meaning of the noun phrase *federal agency*?

Most approaches to distributional semantics represent words as vectors in a high-dimensional space and use linear algebra operations to combine individual word representations in order to obtain representations for complex units. In this article, we propose a probabilistic approach to distributional semantics. This approach is based on the generative model of sentences with latent variables, which was introduced by Grave et al. (2013). We make the following contributions:

- Given the model introduced by Grave et al. (2013), we describe how in-context and ouf-of-context words can be represented by posterior distributions over latent variables (section 4).
- We evaluate out-of-context representations on human similarity judgements prediction tasks and determine what kind of semantic relations are favored by our approach (section 5).
- Finally, we evaluate in-context representations on two similarity tasks for short phrases (section 6).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1447–1456, Dublin, Ireland, August 23-29 2014.

2 Related work

Most approaches to distributional semantics are based on vector space models (VSM), in which words are represented as vectors in a high-dimensional space. These vectors are obtained from a large text corpus, by extracting distributional information about words such as the contexts in which they appear. A corpus is then represented as a word-by-context co-occurence matrix. Contexts can be defined as documents in which the target word appear (Deerwester et al., 1990; Landauer and Dumais, 1997) or as words that appear in the neighbourhood of the target word, for example in the same sentence or in a fixed-size window around the target word (Schutze, 1992; Lund and Burgess, 1996).

Next to vector space models, other approaches to distributional semantics are based on probabilistic models of documents, such as probabilistic latent semantic analysis (pLSA) introduced by Hofmann (1999) and which is inspired by latent semantic analysis, or latent Dirichlet allocation (LDA), introduced by Blei et al. (2003). In those models, each document is viewed as a mixture of k topics, where each topic is a distribution over the words of the vocabulary.

The previous models do not take into account the linguistic structure of the sentences used to build word representations. Several models have been proposed to address this limitation. In those models, the contexts are defined by using the syntactic relations between words (Lin, 1998; Curran and Moens, 2002; Turney, 2006; Padó and Lapata, 2007; Baroni and Lenci, 2010). For example, two words are considered in the same context if there exists a syntactic relation between them, or if there is a path between them in the dependency graph.

One of the first approaches to semantic compositionality using vector space models was proposed by Mitchell and Lapata (2008). In this study, individual word representations are combined using linear algebra operations such as addition, componentwise multiplication, tensor product or dilation. Those different composition operations are then used to disambiguate intransitive verbs given a subject (Mitchell and Lapata, 2008) or to compute similarity scores between pairs of small phrases (Mitchell and Lapata, 2010).

Another approach to semantic compositionality is to learn the function used to compose individual word representations. First, a semantic space containing representations for both individual words and phrases is built. For example, the words *federal*, *agency* and the phrase *federal agency* all have a vectorial representation. Then, a function mapping individual word representations to phrase representations can be learnt in a supervised way. Guevara (2010) proposed to use partial least square regression to learn this function. Similarly, Baroni and Zamparelli (2010) proposed to learn a matrix $\mathbf{A}$ for each adjective, such that the vectorial representation $\mathbf{p}$ of the adjective-noun phrase can be obtained from the vectorial representation $\mathbf{b}$ of the noun by the matrix-vector multiplication:

$$\mathbf{p} = \mathbf{A}\mathbf{b}.$$

Socher et al. (2012) later generalized this model by proposing to represent each node in a parse tree by a vector capturing the meaning and a matrix capturing the compositional effects. A composition function, inspired by artificial neural networks, is recursively applied in the tree to compute those representations.

Following the theoretical framework introduced by Coecke et al. (2010), Grefenstette and Sadrzadeh (2011) proposed to represent relational words (such as verbs) by tensors and theirs arguments (such as nouns) by vectors. Composing a relational word with its arguments is then performed by taking the pointwise product between the tensor and the Kronecker product of the vectors representing the arguments. Jenatton et al. (2012) and Van de Cruys et al. (2013) proposed two approaches to model subject-verb-object triples based on tensor factorization.

Finally, research in computation of word meaning in context is closely related to distributional semantic compositionality. Erk and Padó (2008) proposed a structured vector space model in which a word is represented by multiple vectors, capturing its meaning but also the selectional restrictions it has for the different arguments. Those different vectors can then be combined to compute a word representation in context. This model was later generalized by Thater et al. (2010). Dinu and Lapata (2010) introduced a probabilistic model for computing word representations in context. In their approach, words are represented as probability distributions over latent senses.

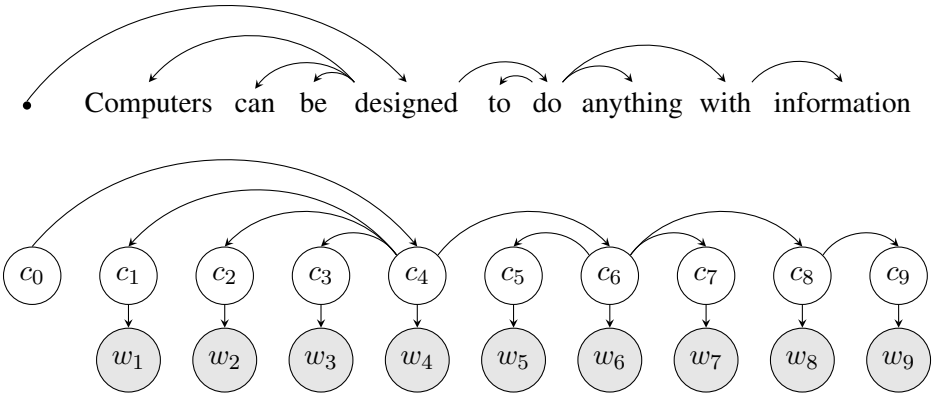

Figure 1: Example of a dependency tree and its corresponding graphical model.

3 Model of semantics

In this section we briefly review the generative model of sentences introduced by Grave et al. (2013), and which serves as the basis of our approach to distributional semantics.

3.1 Generative model of sentences

We denote the tokens of a sentence of length K by the K-uple $\mathbf{w} = (w_1, ..., w_K) \in \{1, ..., V\}^K$, where V is the size of the vocabulary and each integer represents a word. We suppose that each token w_k is associated to a corresponding semantic class $c_k \in \{1, ..., C\}$, where C is the number of semantic classes. Finally, the syntactic dependency tree corresponding to the sentence is represented by the function $\pi : \{1, ..., K\} \mapsto \{0, ..., K\}$, where $\pi(k)$ represents the parent of word k and 0 is the root of the tree (which is not associated to a word).

Given a tree π, the semantic classes and the words of a sentence are generated as follows. The semantic class of the root of the tree is set to a special start symbol, represented by the integer 0.[1] Then, the semantic classes corresponding to words are recursively generated down the tree: each semantic class c_k is drawn from a multinomial distribution $p_T(c_k \mid c_{\pi(k)})$, conditioned on the semantic class $c_{\pi(k)}$ of its parent in the tree. Finally, each word w_k is also drawn from a multinomial distribution $p_O(w_k \mid c_k)$, conditioned on its corresponding semantic class c_k. Thus, the joint probability distribution on words and semantic classes can be factorized as

$$p(\mathbf{w}, \mathbf{c}) = \prod_{k=1}^{K} p_T(c_k \mid c_{\pi(k)}) p_O(w_k \mid c_k),$$

where the variable $c_0 = 0$ represents the root of the tree. The initial class probability distribution $p_T(c_k \mid c_0 = 0)$ is parameterized by the probability vector $\mathbf{q}$, while the transition probability distribution between classes $p_T(c_k \mid c_{\pi(k)})$ and the emission probability distribution $p_O(w_k \mid c_k)$ are parameterized by the stochastic matrices $\mathbf{T}$ and $\mathbf{O}$ (*i.e.*, matrices with non-negative elements and unit-sum columns). This model is a hidden Markov model on a tree (instead of a chain). See Fig. 1 for an example of a sentence and its corresponding graphical model.

3.2 Corpus and learning

We train the generative model of sentences on the ukWac corpus (Baroni et al., 2009). This corpus, which contains approximately 1.9 billions tokens, was POS-tagged and lemmatized using TreeTagger (Schmid, 1994) and parsed using MaltParser (Nivre et al., 2007). Each word of our vocabulary is a pair of lemma and its part-of-speech. We perform smoothing by only keeping the V most frequent pairs, the infrequent ones being replaced by a common token. The parameters $\theta = (\mathbf{q}, \mathbf{T}, \mathbf{O})$ of the model are learned using the algorithm described by Grave et al. (2013). The number of latent states C and the number of lemma/POS pairs V were set using the development set of Bruni et al. (2012).

[1] We recall that the semantic classes corresponding to words are represented by integers between 1 and C.

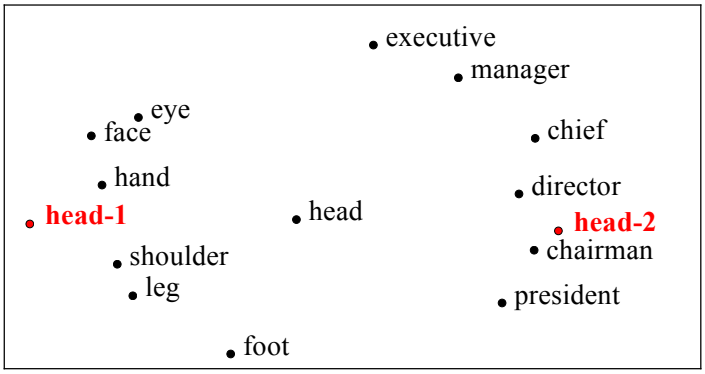

Figure 2: Comparison of out-of-context (black) and in-context (red) word representations. The two-dimensional visualization is obtained by using multidimensional scaling (Borg, 2005). See text for details.

4 Word representations

Given a trained hidden Markov model, we now describe how to obtain word representations, for both in-context and out-of-context words. In both cases, words will be represented as a probability distribution over the latent semantic classes.

In-context word representation. Obtaining a representation of a word in the context of a sentence is very natural using the model introduced in the previous section: we start by parsing the sentence in order to obtain the syntactic dependency tree. We then compute the posterior distribution of semantic classes c for that word, and use this probability distribution to represent the word. More formally, given a sentence $\mathbf{w} = (w_1, ..., w_K)$, the kth word of the sentence is represented by the vector $\mathbf{u}^k \in \mathbb{R}^C$ defined by

$$u_i^k = \mathbb{P}(C_k = i \mid W = \mathbf{w}).$$

The vector $\mathbf{u}^k$ is the posterior distribution of latent classes corresponding to the kth word of the sentence, and thus, sums to one. It is efficiently computed using the message passing algorithm (a.k.a. forward-backward algorithm for HMM).

Out-of-context representation. In order to obtain word representations that are independent of the context, we compute the previously introduced in-context representations on a very large corpus, and for each word type, we average all the in-context representations for all the occurrences of that word type in the corpus. More formally, given a large set of pairs of tokens and their in-context representations $(w_k, \mathbf{u}^k) \in \mathbb{N} \times \mathbb{R}^C$, the representation of the word type a is the vector $\mathbf{v}^a \in \mathbb{R}^C$, defined by

$$\mathbf{v}^a = \frac{1}{Z_a} \sum_{k \,:\, w_k = a} \mathbf{u}^k,$$

where Z_a is the number of occurrences of the word type a. The vector $\mathbf{v}^a$ is thus the posterior distribution of semantic classes averaged over all the occurrences of word type a.

Comparing in-context and out-of-context representations. Since in-context and out-of-context word representations are defined on the same space (the simplex of dimension C) it is possible to compare in-context and out-of-context representations easily. As an example, we have plotted in Figure 2 the out-of-context representation for the words *head, president, chief, chairman, director, executive, eye, face, shoulder, hand, leg, etc.* and the in-context representations for the word *head* in the context of the two following sentences:

1. *The nurse stuck her* head *in the room to announce that Dr. Reitz was on the phone.*
2. *A well-known Wall Street figure may join the Cabinet as* head *of the Treasury Department.*

Distance	RG65	WS353
Cosine	0.68	0.50
Kullback-Leibler	0.69	0.47
Jensen-Shannon	0.72	0.50
Hellinger	0.73	0.51
Agirre et al. (BoW)	0.81	0.65

Distance	SIM.	REL.
Cosine	0.68	0.34
Kullback-Leibler	0.64	0.31
Jensen-Shannon	0.69	0.33
Hellinger	0.70	0.34
Agirre et al. (BoW)	0.70	0.62

Table 1: Left: Spearman's rank correlation coefficient ρ between human and distributional similarity, on the RG65 and WORDSIM353 datasets. Right: Spearman's rank correlation coefficient ρ between human and distributional similarity on two subsets (similarity *v.s.* relatedness) of the WORDSIM353 dataset.

The two-dimensional visualization is obtained by using multidimensional scaling (Borg, 2005). First of all, we observe that the words are clustered in two groups, one containing words belonging to the *body part* class, the other containing words belonging to the *leader* class, and the word *head*, appears between those two groups. Second, we observe that the in-context representations are shifted toward the cluster corresponding to the disambiguated sense of the ambiguous word *head*.

5 Out-of-context evaluation

In this section, we evaluate out-of-context word representations on a similarity prediction task and determine what kind of semantic relations are favored by our approach.

5.1 Similarity judgements prediction

In word similarity prediction tasks, pairs of words are presented to human subjects who are asked to rate the relatedness between those two words. These human similarity scores are then compared to distributional similarity scores induced by our models, by computing the correlation between them.

Methodology. We use the RG65 dataset, introduced by Rubenstein and Goodenough (1965) and the WORDSIM353 dataset, collected by Finkelstein et al. (2001). These datasets comprise 65 and 353 word pairs respectively. Human subjects rated the relatedness of those word pairs. We use the Spearman's rank correlation coefficient ρ to compare human and distributional score distributions.

Comparison of similarity measures. Since words are represented by posterior distributions over latent semantic classes, we have considered distances (or divergences) that are adapted to probability distributions to compute the similarity between word representations: the symmetrised Kullback-Leibler divergence, the Jensen-Shannon divergence, and the Hellinger distance. We use the opposite of these dissimilarity measures in order to obtain similarity scores. We also included the cosine similarity measure as a baseline, as it is widely used in the field of distributional semantics.

We report results on both datasets in Table 1. Unsurprisingly, we observe that the dissimilarity measures giving the best results are the one tailored for probability distribution, namely the Jensen-Shannon divergence and the Hellinger distance. The Kullback-Leibler divergence is too sensitive to fluctuations of small probabilities and thus does not perform as well as other similarity measures between probability distributions. In the following, we will use the Hellinger distance. It should be noted that the results reported by Agirre et al. (2009) were obtained using a corpus containing 1.6 terawords, making it 1,000 times larger than ours. They also report results for various corpus sizes, and when using a corpus whose size is comparable to ours, their result on WORDSIM353 drops to 0.55.

Relatedness *v.s.* similarity. As noted by Agirre et al. (2009), words might be rated as related for different reasons since different kinds of semantic relations exist between word senses. Some words, such as *telephone* and *communication* might even be rated as related because they belong to the same semantic field. Thus, they proposed to split the WORDSIM353 dataset into two subsets: the first one comprising words that are *similar*, *i.e.*, synonyms, antonyms and hyperonym-hyponym and the second

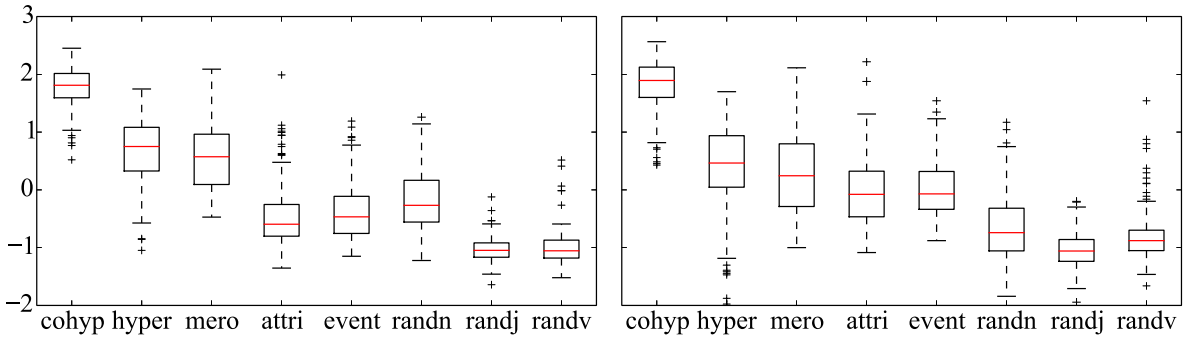

Figure 3: Similarity score distributions for various semantic relations on the BLESS dataset, without using the transition matrix (left) and with using the transition matrix (right) for comparing adjectives and verbs with nouns.

one comprising words that are *related*, *i.e.*, meronym-holonym and topically related words. We report results on these two subsets in Table 1. We observe that our model capture *similarity* ($\rho = 0.70$) much better than *relatedness* ($\rho = 0.34$). This is not very surprising since our model takes the syntax into account.

5.2 Semantic relations captured by our word representations

As we saw in the previous section, different semantic relations between words are not equally captured by our word representations. In this section, we thus investigate which kind of semantic relations are favored by our approach.

The BLESS dataset. The BLESS dataset (Baroni and Lenci, 2011) comprises 200 concrete concepts and eight relations. For each pair of concept-relation, a list of related words, referred to as relatum, is given. Five semantic relations are considered: *co-hyponymy*, *hypernymy*, *meronymy*, *attribute* and *event*. The *attribute* relation means that the relatum is an adjective expressing an attribute of the concept, while the *event* relation means that the relatum is a verb designing an activity or an event in which the concept is involved. The dataset also contains three *random* relations (*randn*, *randj* ans *randv*), obtained by the association of a random relatum, for different POS: noun, adjective and verb.

Methodology. We follow the evaluation proposed by the authors: for each pair of concept-relation, we keep the score of the most similar relatum associated to that pair of concept-relation. Thus, for each concept, we have eight scores, one for each relation. We normalize these eight scores (mean: 0, std: 1), in order to reduce concept-specific effects. We then report the score distributions for each relation as box plots in Figure 3 (left).

Results. We observe that the co-hyponymy relation is the best captured relation by a large margin. It is followed by the hypernymy and meronymy relations. The random noun relation is prefered over the attribute and the event relations. This happens because words with different part-of-speeches tend to appear in different semantic classes. It is thus impossible to compare words with different parts-of-speeches and thus to capture relation such as the event or the attribute relation as defined in the BLESS dataset. It is however possible to make a more principled use of the model to overcome this issue.

Comparing adjectives with nouns and nouns with verbs. In syntactic relations between nouns and adjectives, the noun is the head word and the adjective is the dependent. Similarly, in syntactic relations between nouns and verbs, most often the verb is the head and the noun is the dependent. Given a vector $\mathbf{v}_a$ representing an adjective and a vector $\mathbf{v}_n$ representing a noun, it is thus natural to left multiply them by the transition matrix of the model to obtain a vector $\mathbf{u}_a$ comparable to nouns and a vector $\mathbf{u}_n$ comparable to verbs:

$$\mathbf{u}_a = \mathbf{T}^\top \mathbf{v}_a \quad \text{and} \quad \mathbf{u}_n = \mathbf{T}^\top \mathbf{v}_n.$$

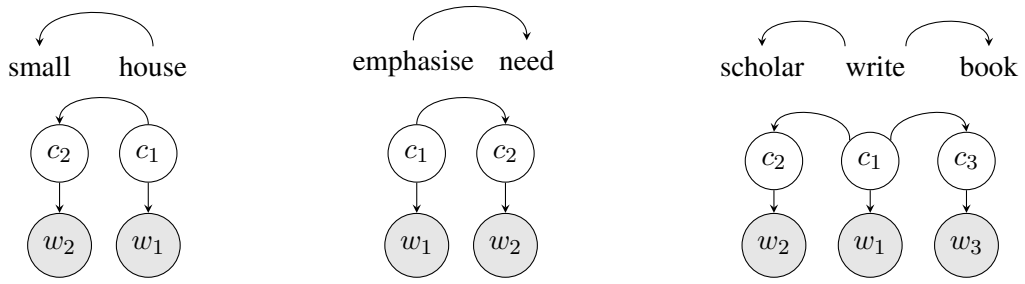

Figure 4: Graphical models used to compute in-context word representations for the compositional tasks.

We report in Figure 3 (right) the new score distributions obtained when adjective and noun representations are transformed before being compared to nouns and verbs. We observe that, when using these transformations, the attribute and event relations are better captured than the random relations. This demonstrates that the transition matrix $\mathbf{T}$ captures selectional preferences.

6 Compositional semantics

So far, we have only evaluated how well our representations are able to capture the meaning of words taken as individual and independent units. However, natural languages are highly compositional, and it is reasonable to assume that the meaning of a sentence or a phrase can be deduced from the meanings of its parts and the syntactic relations between them. This assumption is known as the principle of semantic compositionality.

In this section, we thus evaluate our representations on semantic composition tasks. More precisely, we determine if using in-context word representations helps to compute the similarity between short phrases such as adjective-noun, verb-object, compound-noun or subject-verb-object phrases. We use two datasets of human similarity scores, introduced respectively by Mitchell and Lapata (2010) and Grefenstette and Sadrzadeh (2011).

6.1 Methodology

We compare different ways to obtain a representation of a short phrase given our model. First, as a baseline, we represent a phrase by the out-of-context representation of its head word. In that case, there is no composition at all. Second, following Mitchell and Lapata (2008), we represent a phrase by the sum of the out-of-context representations of the words forming that phrase. Third, we represent a phrase by the in-context representation of its head word. Finally, we represent a phrase by the sum of the two in-context representations of the words forming that phrase. The graphical models used to compute in-context word representations are represented in Fig 4. The probability distribution $p(c_1)$ of the head's semantic class is set to the uniform distribution (and not to the initial class distribution $p_T(c_k \mid c_0 = 0)$).

6.2 Datasets

The first dataset we consider was introduced by Mitchell and Lapata (2010), and is composed of pairs of adjective-noun, compound-noun and verb-object phrases, whose similarities were evaluated by human subjects on a $1-7$ scale. We compare our results with the one reported by (Mitchell and Lapata, 2010). The second dataset we consider was introduced by Grefenstette and Sadrzadeh (2011). Each example of this dataset consists in a triple of subject-verb-object, forming a small transitive sentence, and a landmark verb. Human subjects were asked to evaluate the similarity between the verb and its landmark in the context of the small sentence. Following Van de Cruys et al. (2013), we compare the contextualized verb with the non-contextualized landmark, meaning that the landmark is always represented by its out-of-context representation. We do so because it is believed to better capture the compositional ability of our model and it works better in practice. We compare our results with the one reported by Van de Cruys et al. (2013).

	AN	NN	VN
head (out-of-context)	0.44	0.26	0.41
add (out-of-context)	0.50	0.45	0.42
head (in-context)	0.49	0.42	**0.43**
add (in-context)	**0.51**	0.46	0.41
M&L (vector space model)	0.46	**0.49**	0.38
Humans	0.52	0.49	0.55

	SVO
head (out-of-context)	0.25
add (out-of-context)	0.25
head (in-context)	**0.41**
add (in-context)	0.40
Van de Cruys et al.	0.37
Humans	0.62

Table 2: Spearman's rank correlation coefficients between human similarity judgements and similarity computed by our models on the Mitchell and Lapata (2010) dataset (left) and on the Grefenstette and Sadrzadeh (2011) dataset (right). AN stands for adjective-noun, NN stands for compoundnoun and VN stands for verb-object.

6.3 Discussion

Before discussing the results, it is interesting to note that our approach provides a way to evaluate the importance of disambiguation for compositional semantics. Indeed, the in-context representations proposed in this paper are a way to disambiguate their out-of-context equivalents. It was previously noted by Reddy et al. (2011) that disambiguating the vectorial representations of words improve the performance on compositional tasks.

Mitchell and Lapata (2010) dataset. We report results on the Mitchell and Lapata (2010) dataset in Table 2 (left). Overall, in-context representations achieves better performance than out-of-context ones. For the adjective-noun pairs and the verb-noun pairs, using only the in-context representation of the head word works almost as well (AN) or even better (VN) than adding the representations of the two words forming a pair. This means that for those particular tasks, disambiguation plays an important role. On the other hand, this is not the case for the noun-noun pairs. On that task, most improvement over the baseline comes from the *add* operation.

Grefenstette and Sadrzadeh (2011) dataset. We report results in Table 2 (right). First, we observe that in-context representations clearly outperform out-of-context ones. Second, we note that adding the subject, object and verb representations does not improve the result over only using the representation of the verb. These two conclusions are not really surprising since this task is mainly a disambiguation task, and disambiguation is achieved by using the in-context representations. We also note that our approach yields better results than those obtained by Van de Cruys et al. (2013), while their method was specifically designed to model subject-verb-object triples.

7 Conclusion and future work

In this article, we introduced a new approach to distributional semantics, based on a generative model of sentences. This model is somehow to latent Dirichlet allocation as structured vector space models are to latent semantic analysis. Indeed, our approach is based on a probabilistic model of sentences, which takes the syntax into account by using dependency trees. Similarly to LDA, our model can be viewed as a topic model, the main difference being that the topics are generated using a Markov process on a syntactic dependency tree instead of using a Dirichlet process.

The approach we propose seems quite competitive with other distributional models of semantics. In particular, we match or outperform state-of-the-art methods on semantic compositionality tasks. Thanks to its probabilistic nature, it is very easy to derive word representations for various tasks: the same model can be used to compute in-context word representations for adjective-noun phrases, subject-verb-object triples or even full sentences, which is not the case of the tensor based approach proposed by Van de Cruys et al. (2013).

Currently, the model of sentences does not use the dependency labels, which is the most significant limitation that we would like to address in future work. We also plan to explore spectral methods (Anandkumar et al., 2012) to provide better initialization for learning the parameters of the model. Indeed, we believe this could speed up learning and yields better results, since the expectation-maximization algorithm is quite sensitive to bad initialization. Finally, the code corresponding to this article will be available on the first author webpage.

Acknowledgments

Edouard Grave is supported by a grant from INRIA (Associated-team STATWEB). Francis Bach is partially supported by the European Research Council (SIERRA Project)

References

E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. 2012. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*.

M. Baroni and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

M. Baroni and A. Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

M. Baroni and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*.

I. Borg. 2005. *Modern multidimensional scaling: Theory and applications*. Springer.

E. Bruni, G. Boleda, M. Baroni, and N. K. Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

S. Clark and S. Pulman. 2007. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55.

B. Coecke, M. Sadrzadeh, and S. Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

J. R. Curran and M. Moens. 2002. Scaling context space. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*.

G. Dinu and M. Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

K. Erk and S. Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*.

J. R. Firth. 1957. *A synopsis of linguistic theory, 1930-1955*.

E. Grave, G. Obozinski, and F. Bach. 2013. Hidden Markov tree models for semantic class induction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*.

E. Grefenstette and M. Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

E. Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*.

Z. S. Harris. 1954. *Distributional structure*. Springer.

T. Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*.

R. Jenatton, N. Le Roux, A. Bordes, and G. Obozinski. 2012. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems 25*.

T. K Landauer and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-volume 2*.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*.

J. Mitchell and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*.

S. Padó and M. Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*.

S. Reddy, I. P. Klapaftis, D. McCarthy, and S. Manandhar. 2011. Dynamic and static prototype vectors for semantic composition. In *IJCNLP*, pages 705–713.

H. Rubenstein and J. B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*.

H. Schutze. 1992. Dimensions of meaning. In *Supercomputing'92. Proceedings*. IEEE.

R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

S. Thater, H. Fürstenau, and M. Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

P. D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*.

T. Van de Cruys, T. Poibeau, and A. Korhonen. 2013. A tensor-based factorization model of semantic compositionality. In *Proceedings of NAACL-HLT*.