TAILORING WORD EMBEDDINGS FOR BILEXICAL PREDICTIONS: AN EXPERIMENTAL COMPARISON

Pranava Swaroop Madhyastha
Universitat Politècnica de Catalunya
Campus Nord UPC, 08034 Barcelona
pranava@cs.upc.edu

Xavier Carreras
Xerox Research Centre Europe
38240 Meylan, France
xavier.carreras@xrce.xerox.com

Ariadna Quattoni
Xerox Research Centre Europe
38240 Meylan, France
ariadna.quattoni@xrce.xerox.com

ABSTRACT

We investigate the problem of inducing word embeddings that are tailored for a particular bilexical relation. Our learning algorithm takes an existing lexical vector space and compresses it such that the resulting word embeddings are good predictors for a target bilexical relation. In experiments we show that task-specific embeddings can benefit both the quality and efficiency in lexical prediction tasks.

1 INTRODUCTION

There has been a large body of work that focuses on learning word representations, either in the form of word clusters (Brown et al., 1992) or vectors (Sahlgren, 2006; Turney & Pantel, 2010; Mikolov et al., 2013; Pennington et al., 2014; Baroni et al., 2014; Bansal et al., 2014) and these have proven useful in many NLP applications (Koo et al., 2008; Turian et al., 2010).

An ideal lexical representation should compress the space of lexical words while retaining the essential properties of words in order to make predictions that correctly generalize across words. The typical approach is to first induce a lexical representation in a task-agnostic setting and then use it in different tasks as features. A different approach is to learn a lexical representation tailored for a certain task. In this work we explore the second approach, and employ the formulation by Madhyastha et al. (2014) to induce task-specific word embeddings. This method departs from a given lexical vector space, and compresses it such that the resulting word embeddings are good predictors for a given lexical relation.

Specifically we learn functions that compute compatibility scores between pairs of lexical items under some linguistic relation. In our work, we refer to these functions as bilexical operators. As an instance of this problem, consider learning a model that predicts the probability that an adjective modifies a noun in a sentence. In this case, we would like the bilexical operator to capture the fact that some adjectives are more compatible with some nouns than others.

Given the complexity of lexical relations, one expects that the properties of words that are relevant for some relation are different for another relation. This might affect the quality of an embedding, both in terms of its predictive power and the compression it obtains. If we employ a task-agnostic low-dimensional embedding, will it retain all important lexical properties for any relation? And, given a fixed relation, can we further compress an existing word representation? In this work we present experiments along these lines that confirm that task-specific embeddings can benefit both the quality and the efficiency of lexicalized predictive models.

2 FORMULATION

Let $\mathcal{V}$ be a vocabulary, and let $x \in \mathcal{V}$ denote a word. We are interested in modeling a target bilexical relation, that is, a relation between pairs of words without context. For example, in a noun-adjective relation we model what nouns can be assigned to what adjectives. We will denote as $\mathcal{Q} \subseteq \mathcal{V}$ the set of query words, or words that appear in the left side of the bilexical relation. And we will use $\mathcal{C} \subseteq \mathcal{V}$ to denote *candidate* words, appearing in the right side of the relation.

In this paper we experiment with the log-linear models by Madhyastha et al. (2014) that given a query word q compute a conditional distribution over candidate words c. The models take the following form:

$$\Pr(c \mid q; W) = \frac{\exp\{\phi(q)^\top W \phi(c)\}}{\sum_{c'} \exp\{\phi(q)^\top W \phi(c')\}} \quad (1)$$

where $\phi : \mathcal{V} \rightarrow \mathbb{R}^n$ is a distributional representation of words, and $W \in \mathbb{R}^{n \times n}$ is a bilinear form.

The learning problem is to obtain ϕ and W from data, and we approach it in a semi-supervised fashion. There exist many approaches to learn ϕ from unlabeled data, and in this paper we experiment with two approaches: (a) a simple distributional approach where we represent words with a bag-of-words of contextual words; and (b) the skip-gram model by Mikolov et al. (2013). To learn W we assume access to labeled data in the form pairs of compatible examples, i.e. $\mathcal{D} = \{(q, c)^1, \ldots, (q, c)^l\}$, where $q \in \mathcal{Q}$ and $c \in \mathcal{C}$. The goal is to be able to predict query-candidate pairs that are unseen during training. Recall that we model relations between words without context. Thus the lexical representation ϕ is essential to generalize to pairs involving unseen words.

With ϕ fixed, we learn W by minimizing the negative log-likelihood of the labeled data using a regularized objective, $L(W) = -\sum_{(q,c) \in D} \log \Pr(c \mid q; W) + \tau \rho(W)$, where $\rho(W)$ is a regularization penalty and τ is a constant controlling the trade-off.

We are interested in regularizers that induce low-rank parameters W, since they lead to task-specific embeddings. Assume that W has rank k, such that $W = UV^\top$ with $U, V \in \mathbb{R}^{n \times k}$. If we consider the product $\phi(q)^\top U V^\top \phi(c)$, we can now interpret $\phi^\top U$ as a k-dimensional embedding of query words, and $\phi(c)^\top V$ as a k-dimensional embedding of candidate words. Thus, if we obtain a low-rank W that is highly predictive, we can interpret U and V as task-specific compressions of the original embedding ϕ tailored for the target bilexical relation, from n to k dimensions.

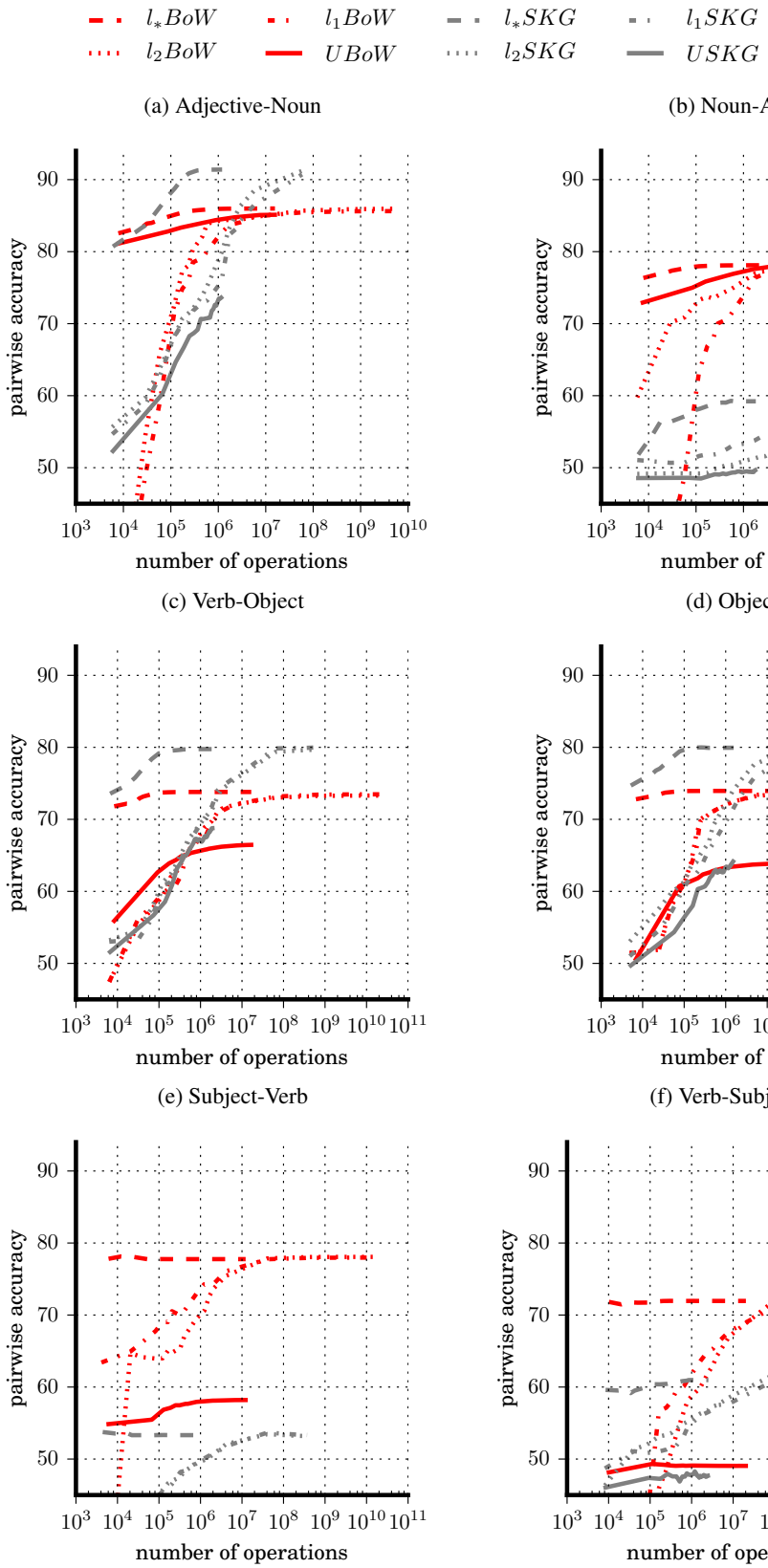
Since minimizing the rank of a matrix is hard, we employ a convex relaxation based on the nuclear norm of the matrix $\ell_\star$ (that is, the ℓ_1 norm of the singular values, see Srebro et al. (2005)). In our experiments we compare the low-rank approach to ℓ_1 and ℓ_2 regularization penalties, which are common in linear prediction tasks. For all settings we use the *forward-backward splitting (FOBOS)* optimization algorithm by Duchi & Singer (2009).

We note that if we set W to be the identity matrix our model scores are inner products between the query-candidate embeddings, a common approach to evaluate semantic similarity in unsupervised distributional approaches. In general, we can compute a low-dimensional projection of ϕ down to k dimensions, using SVD, and perform the inner product in the projected space. We refer to this as the unsupervised approach, since the projected embeddings do not use the labeled dataset specifying the target relation.

3 EXPERIMENTS WITH SYNTACTIC RELATIONS

We conducted a set of experiments to test the performance of the learning algorithm with respect to the initial lexical representation ϕ, for different configurations of the representation and the learner. We experiment with six bilexical syntactic relations using the Penn Treebank corpus (Marcus et al., 1993), following the experimental setting by Madhyastha et al. (2014). For a relation between queries and candidate words, such as noun-adjective, we partition the query words into train, development and test queries, thus test pairs are always unseen pairs.

To report performance, we measure pairwise accuracy with respect to the efficiency of the model in terms of number of active parameters. To measure the efficiency of a model we consider the number of double operations that are needed to compute, given a query word, the scores for all candidates in the vocabulary. See (Madhyastha et al., 2014) for details.



(a) Adjective-Noun

(b) Noun-Adjective

(c) Verb-Object

(d) Object-Verb

(e) Subject-Verb

(f) Verb-Subject

Figure 1: Pairwise accuracy v/s no. of double operations to compute the distribution over candidate words for a query word. Plots are for noun-adjective, verb-object and verb-subject relations, in both directions. The red curves use distributional representation based on bag-of-words (BoW) and the grey curves use the embeddings of the skip-gram model (SKG).

			$\ell_\star$				
Rel	*Type*	*UNS*	*best k*	$k=5$	$k=10$	ℓ_2	ℓ_1
Adj-Noun	BoW	85.12	85.99 (80)	83.99	84.74	85.96	85.63
	SKG	73.61	91.40 (300)	83.70	86.27	91.22	90.72
Obj-Verb	BoW	63.85	78.11 (200)	73.17	73.64	74.08	73.95
	SKG	64.15	79.98 (50)	75.45	78.37	80.30	79.89
Subj-Verb	BoW	58.20	78.13 (2)	71.71	71.73	78.07	77.97
	SKG	49.65	59.28 (90)	53.31	53.32	58.24	58.67
Noun-Adj	BoW	78.09	78.11 (70)	77.48	77.85	78.48	78.85
	SKG	49.65	59.28 (50)	56.42	57.19	58.24	58.67
Verb-Obj	BoW	66.46	73.90 (40)	73.70	73.88	73.30	73.48
	SKG	64.15	79.99 (30)	77.05	78.60	80.29	79.89
Verb-Subj	BoW	49.32	71.97 (30)	71.71	71.23	72.85	71.95
	SKG	32.34	53.75 (2)	53.32	53.32	53.47	53.68

Table 1: Pairwise accuracies for the six relations using the unsupervised, $\ell_\star$, ℓ_2 and ℓ_1 models, using either a distributional bag-of-words representation (BoW) or the skip-gram embeddings (SKG) as initial representation. For $\ell_\star$ we show results for the rank that gives best accuracy (with the optimal rank in parenthesis), as well as for ranks $k = 5$ and 10.

We experiment with two types of initial representations ϕ. The first is a simple high-dimensional distributional representation based on contextual bag-of-words (BoW): each word is represented by the bag of words that appear in contextual windows. In our experiments these were sparse 2,000-dimensional vectors. The second representation are the low-dimensional skip-gram embeddings (SKG) by Mikolov et al. (2013), where we used 300 dimensions. In both cases we induced such representations using the BLIPP corpus (Charniak et al., 2000) and using a context window of size 10 for both. Thus the main difference is that the bag-of-words is an uncompressed representation, while the skip-gram embeddings are a neural-net-style compression of the same contextual windows.

As for the bilexical model, we test it under three regularization schemes, namely ℓ_2, ℓ_1, and $\ell_\star$. For the first two, the efficiency of computing predictions is a function of the non-zero entries in W, while for the latter it is the rank k of W, which defines the dimension of the task-specific embeddings. We also test a baseline unsupervised approach (UNS).

4 Results and Discussion

Figure 1 shows the performance of models for noun-adjective, verb-object and verb-subject relations (in both directions). In line with the results by Madhyastha et al. (2014) we observe that the supervised approach in all cases outperforms the unsupervised case, and that the nuclear norm scheme provides the best performance in terms of accuracy and speed: other regularizers can obtain similar accuracies, but low-rank constraints during learning favor very-low dimensional embeddings that are highly predictive.

In terms of starting with bag-of-words vectors or skip-gram embeddings, in three relations the former is clearly better, while in the other three relations the latter is clearly better. We conclude that task-agnostic embeddings do identify useful relevant properties of words, but at the same time not all necessary properties are retained. In all cases, the nuclear norm regularizer successfully compresses the initial representation, even for the embeddings which are already low-dimensional.

Table 1 presents the best result for each relation, initial representation and regularization scheme. Plus, for the $\ell_\star$ regularizer we present results at three different ranks, namely 5, 10 or the rank that obtains the best result for each relation. These highly compressed embeddings perform nearly as good as the best performing model for each relation.

Table 2 shows a set of query nouns, and two sets of neighbor query nouns, using the embeddings for two different relations to compute the two sets. We can see that, by changing the target relation, the set of close words changes. This suggests that words have a wide range of different behaviors, and different relations might exploit lexical properties that are specific to the relation.

Query	noun-adjective	object-verb
city	province, area, island, township, freeways	residents, towns, marchers, streets, mayor
securities	bonds, mortgage, issuers, debt, loans	bonds, memberships, equities, certificates, syndicate
board	committee directors, commission, nominees, refusal	slate, membership, committee, appointment, stockholder
debt	loan, loans, debts, financing, mortgage	reinvestment, indebtedness, expenditures, outlay, repayment
law	laws, constitution, code, legislation, immigration	laws, ordinance, decree, statutes, state
director	assistant, editor, treasurer, postmaster, chairman	firm, consultant, president, manager, leader

Table 2: Example query words and 5 highest-ranked candidate words for two different bilexical relations: noun-adjective and object-verb.

5 CONCLUSION

We have presented a set of experiments where we compute word embeddings specific to target linguistic relations. We observe that low-rank penalties favor embeddings that are good both in terms of predictive accuracy and efficiency. For example, in certain cases, models using very low-dimensional embeddings perform nearly as good as the best models.

In certain tasks, we have shown that we can refine low-dimensional skip-gram embeddings, making them more compressed while retaining their predictive properties. In other tasks, we have shown that our method can improve over skip-gram models when starting from uncompressed distributional representations. This suggests that skip-gram embeddings do not retain all the necessary information of the original words. This motivates future research that aims at general-purpose embeddings that do retain all necessary properties, and can be further compressed in light of specific lexical relations.

ACKNOWLEDGEMENTS

We thank the reviewers for their helpful comments. This work has been partially funded by the Spanish Government through the SKATER project (TIN2012-38584-C06-01).

REFERENCES

Bansal, Mohit, Gimpel, Kevin, and Livescu, Karen. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL*, 2014.

Baroni, Marco, Dinu, Georgiana, and Kruszewski, Germán. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1023.

Brown, Peter F., deSouza, Peter V., Mercer, Robert L., Pietra, Vincent J. Della, and Lai, Jenifer C. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479, 1992.

Charniak, Eugene, Blaheta, Don, Ge, Niyu, Hall, Keith, and Johnson, Mark. BLLIP 1987–89 WSJ Corpus Release 1, LDC No. LDC2000T43. Linguistic Data Consortium, 2000.

Duchi, John and Singer, Yoram. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

Koo, Terry, Carreras, Xavier, and Collins, Michael. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pp. 595–603, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P08/P08-1068.

Madhyastha, Swaroop Pranava, Carreras, Xavier, and Quattoni, Ariadna. Learning task-specific bilexical embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 161–171. Dublin City University and Association for Computational Linguistics, 2014. URL http://aclweb.org/anthology/C14-1017.

Marcus, Mitchell P., Santorini, Beatrice, and Marcinkiewicz, Mary A. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, 2014. URL http://aclweb.org/anthology/D14-1162.

Sahlgren, Magnus. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, 2006.

Srebro, Nathan, Rennie, Jason D. M., and Jaakola, Tommi S. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pp. 1329–1336. MIT Press, 2005.

Turian, Joseph, Ratinov, Lev-Arie, and Bengio, Yoshua. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-1040.

Turney, Peter D. and Pantel, Patrick. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, January 2010. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=1861751.1861756.