Specializing Word Embeddings for Similarity or Relatedness

Douwe Kiela, Felix Hill and Stephen Clark
Computer Laboratory
University of Cambridge
douwe.kiela|felix.hill|stephen.clark@cl.cam.ac.uk

Abstract

We demonstrate the advantage of specializing semantic word embeddings for either similarity or relatedness. We compare two variants of retrofitting and a joint-learning approach, and find that all three yield specialized semantic spaces that capture human intuitions regarding similarity and relatedness better than unspecialized spaces. We also show that using specialized spaces in NLP tasks and applications leads to clear improvements, for document classification and synonym selection, which rely on either similarity or relatedness but not both.

1 Introduction

Most current models of semantic word representation exploit the *distributional hypothesis*: the idea that words occurring in similar contexts have similar meanings (Harris, 1954; Turney and Pantel, 2010; Clark, 2015). Such representations (or embeddings) can reflect human intuitions about similarity and relatedness (Turney, 2006; Agirre et al., 2009), and have been applied to a wide variety of NLP tasks, including bilingual lexicon induction (Mikolov et al., 2013b), sentiment analysis (Socher et al., 2013) and named entity recognition (Turian et al., 2010; Guo et al., 2014).

Arguably, one of the reasons behind the popularity of word embeddings is that they are "general purpose": they can be used in a variety of tasks without modification. Although this behavior is sometimes desirable, it may in other cases be detrimental to downstream performance. For example, when classifying documents by topic, we are particularly interested in related words rather than similar ones: knowing that *dog* is associated with *cat* is much more informative of the topic than knowing that it is a synonym of *canine*. Conversely, if our embeddings indicate that *table* is closely related to *chair*, that does not mean we should translate *table* into French as *chaise*.

This distinction between "genuine" similarity and associative similarity (i.e., relatedness) is well-known in cognitive science (Tversky, 1977). In NLP, however, semantic spaces are generally evaluated on how well they capture *both* similarity and relatedness, even though, for many word combinations (such as *car* and *petrol*), these two objectives are mutually incompatible (Hill et al., 2014b). In part, this oversight stems from the distributional hypothesis itself: *car* and *petrol* do not have the same, or even very similar, meanings, but these two words may well occur in similar contexts. Corpus-driven approaches based on the distributional hypothesis therefore generally learn embeddings that capture both similarity and relatedness reasonably well, but neither perfectly.

In this work we demonstrate the advantage of specializing semantic spaces for either similarity or relatedness. Specializing for similarity is achieved by learning from both a corpus and a thesaurus, and for relatedness by learning from both a corpus and a collection of psychological association norms. We also compare the recently-introduced technique of graph-based retrofitting (Faruqui et al., 2015) with a skip-gram retrofitting and a skip-gram joint-learning approach. All three methods yield specialized semantic spaces that capture human intuitions regarding similarity and relatedness significantly better than unspecialized spaces, in one case yielding state-of-the-art results for word similarity. More importantly, we show clear improvements in downstream tasks and applications: specialized similarity spaces improve synonym detection, while association spaces work better than both general-purpose and similarity-specialized spaces for document classification.

2 Approach

The underlying assumption of our approach is that, during training, word embeddings can be "nudged" in a particular direction by including information from an additional semantic data

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2044–2048,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

source. For directing embeddings towards genuine similarity, we use the MyThes thesaurus developed by the OpenOffice.org project[1]. It contains synonyms for almost 80,000 words in English. For directing embeddings towards relatedness, we use the University of South Florida (USF) free association norms (Nelson et al., 2004). This dataset contains scores for free association (an experimental measure of cognitive association) of over 10,000 concept words. For raw text data we use a dump of the English Wikipedia plus newswire text (8 billion words in total)[2].

2.1 Evaluations (Intrinsic and Extrinsic)

For instrinsic comparisons with human judgements, we evaluate on SimLex (Hill et al., 2014b) (999 pairwise comparisons), which explicitly measures similarity, and MEN (Bruni et al., 2014) (3000 comparisons), which explicitly measures relatedness. We also consider two downstream tasks and applications. In the TOEFL synonym selection task (Landauer and Dumais, 1997), the objective is to select the correct synonym for a target word from a multiple-choice set of possible answers. For a more extrinsic evaluation, we use a document classification task based on the Reuters Corpus Volume 1 (RCV1) (Lewis et al., 2004). This dataset consists of over 800,000 manually categorized news articles.[3]

2.2 Joint Learning

The standard skip-gram training objective for a sequence of training words $w_1, w_2, ..., w_T$ and a context size c is the log-likelihood criterion:

$$\frac{1}{T}\sum_{t=1}^{T} J_\theta(w_t) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c} \log p(w_{t+j}|w_t)$$

where $p(w_{t+j}|w_t)$ is obtained via the softmax:

$$p(w_{t+j}|w_t) = \frac{\exp^{u_{w_{t+j}}^\top v_{w_t}}}{\sum_{w'} \exp^{u_{w'}^\top v_{w_t}}}$$

where u_w and v_w are the context and target vector representations for word w, respectively, and w' ranges over the full vocabulary (Mikolov et al., 2013a). For our *joint learning* approach, we supplement the skip-gram objective with additional contexts (synonyms or free-associates) from an external data source. In the **sampling** condition, for target word w_t, we modify the objective to include an additional context w^a sampled uniformly from the set of additional contexts A_{w_t}:

$$\frac{1}{T}\sum_{t=1}^{T}\big(J_\theta(w_t) + [w^a \sim \mathcal{U}_{A_{w_t}}]\log p(w^a|w_t)\big)$$

In the **all** condition, all additional contexts for a target word are added at each occurrence:

$$\frac{1}{T}\sum_{t=1}^{T}\left(J_\theta(w_t) + \sum_{w^a\in A_{w_t}} \log p(w^a|w_t)\right)$$

The set of additional contexts A_{w_t} contains the relevant contexts for a word w_t; e.g., for the word *dog*, A_{dog} for the thesaurus is the set of all synonyms of dog in the thesaurus.

2.3 Retrofitting

Faruqui et al. (2015) introduced retrofitting as a post-hoc graph-based learning objective that improves learned word embeddings. We experiment with their method, calling it *graph-based retrofitting*. In addition, we introduce a similar approach that instead uses the same objective function that was used to learn the original skip-gram embeddings. In other words, we first train a standard skip-gram model, and then learn from the additional contexts in a second training stage as if they form a separate corpus:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{w^a\in A_{w_t}} \log p(w^a|w_t)$$

We call this approach *skip-gram retrofitting*. In all cases, our embeddings have 300 dimensions, which has been found to work well (Mikolov et al., 2013a; Baroni et al., 2014)

3 Results for Intrinsic Evaluation

We compare standard skip-gram embeddings with retrofitted and jointly learned specialized embeddings, as well as with “fitted” embeddings that were randomly initialized and learned only from the additional semantic resource. In each case, the

[1] https://www.openoffice.org/lingucomponent/thesaurus.html

[2] The script for obtaining this corpus is available from http://word2vec.googlecode.com/svn/trunk/demo-train-big-model-v1.sh

[3] We exclude articles with multiple topic labels in order to avoid multi-class document classification. The dataset contains a total of 78 topic labels and 33,226 news articles.

Method	SimLex-999	MEN
Skip-gram	0.31	0.68
Fit-Norms	0.08	0.14
Fit-Thesaurus	0.26	0.14
Joint-Norms-Sampled	0.43	**0.72**
Joint-Norms-All	0.42	0.67
Joint-Thesaurus-Sampled	0.38	0.69
Joint-Thesaurus-All	0.44	0.60
GB-Retrofit-Norms	0.32	0.71
GB-Retrofit-Thesaurus	0.38	0.68
SG-Retrofit-Norms	0.35	0.71
SG-Retrofit-Thesaurus	**0.47**	0.69

Table 1: Spearman ρ on a genuine similarity (SimLex-999) and relatedness (MEN) dataset.

training algorithm was run for a single iteration (results from more iterations are presented later).

As shown in Table 1, embeddings that were specialized for similarity using a thesaurus perform better on SimLex-999, and those specialized for relatedness using association data perform better on MEN. Fitting, or learning only from the additional semantic resource without access to raw text, does not perform well. Skip-gram retrofitting with the thesaurus performs best on SimLex-999; joint learning with sampling from the USF norms performs best on MEN, although the two retrofitting approaches are close. There is an interesting difference between the two joint learning approaches: while **sampling** a single free associate as additional context works best for relatedness, presenting **all** additional contexts (synonyms) works best for similarity. In both cases, skip-gram retrofitting matches or outperforms graph-based retrofitting.

More training iterations All the results above were obtained using a single training iteration. When retrofitting, however, it is easy to learn from multiple iterations of the thesaurus or the USF norms. The results are shown in Figure 1, where the dashed lines are the joint learning and standard skip-gram results for comparison with retrofitting scores. As would be expected, too many iterations leads to overfitting on the semantic resource, with performance eventually decreasing after the initial increase. The results show that retrofitting is particularly useful for similarity, as indicated by the large increase in performance on SimLex-999. The highest performance obtained, at 5 iterations, is a Spearman ρ_s correlation of 0.53, which, as far as we know, matches the current state-of-the-art.[4]

For relatedness-specific embeddings, the effect is less clear: joint learning performs comparatively much better. Retrofitting does outperform it, at around 2-10 iterations on the USF norms, but the improvement is marginal. The highest retrofitting score is 0.74; the highest joint learning score is 0.72. Both are highly competitive results on MEN, and outperform e.g. GloVe at 0.71 (Pennington et al., 2014). Joint learning with a thesaurus, however, leads to poor performance on MEN, as expected: the embeddings get dragged away from relatedness and towards similarity.

3.1 Curriculum learning?

The fact that joint learning works better when supplementing raw text input with free associates, but skip-gram retrofitting works better with additional thesaurus information, could be due to *curriculum learning* effects (Bengio et al., 2009). Unlike the USF norms, many of the words from the thesaurus are unusual and have low frequency. This suggests that the thesaurus is more ‘advanced’ (from the perspective of the learning model) than the USF norms as an information source. Its information may be detrimental to model optimization when encountered early in training (in the joint learning condition) because the model has not acquired the basic concepts on which it builds. However, with retrofitting the model first acquires good representations for frequent words from the raw text, after which it can better understand, and learn from, the information in the thesaurus.

4 Downstream Tasks and Applications

4.1 TOEFL Synonym Task

Unsupervised synonym selection has many applications including the generation of thesauri and other lexical resources from raw text (Kageura et al., 2000). In the well-known TOEFL evaluation (Freitag et al., 2005) models are required to identify true synonyms to question words from a selection of possible answers. To test our models on this task, for each question in the dataset, we rank the multiple-choice answers according to the cosine similarity between their word embeddings and that of the target word, and choose the highest-ranked option.

[4] Hill et al. (2014a) obtain a score of 0.52 using neural translation embeddings.

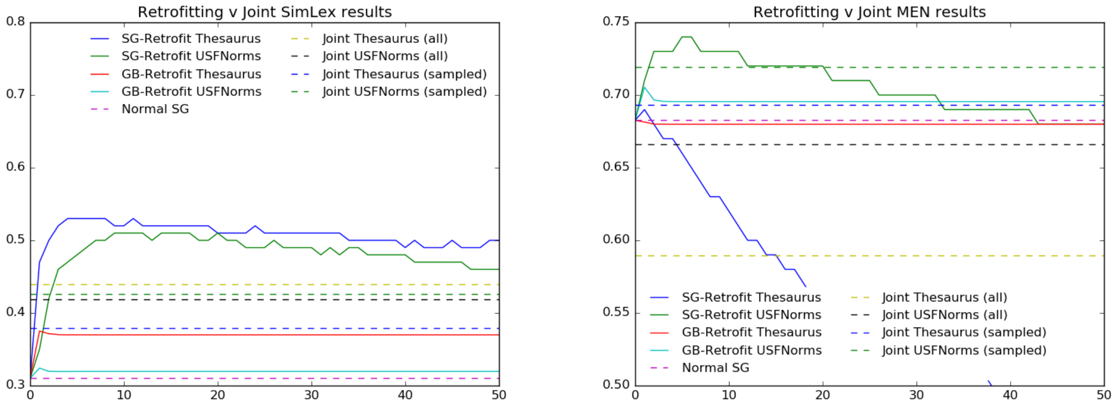

Figure 1: Varying the number of iterations when retrofitting

Method	TOEFL		Doc
Skip-gram	77.50		83.96
Joint-Norms-Sampled	78.75		84.46
Joint-Norms-All	66.25		**84.82**
Joint-Thesaurus-Sampled	81.25		83.90
Joint-Thesaurus-All	80.00		83.56
GB-Retrofit-Norms	80.00		80.58
GB-Retrofit-Thesaurus	83.75		80.24
SG-Retrofit-Norms	80.00		84.56
SG-Retrofit-Thesaurus	**88.75**		84.55

Table 2: TOEFL synonym selection and document classification accuracy (percentage of correctly answered questions/correctly classified documents).

As Table 2 shows, similarity-specialized embeddings perform much better than standard embeddings and relatedness-specialized embeddings. Retrofitting outperforms joint learning, and skip-gram retrofitting matches or outperforms graph-based retrofitting.

4.2 Document Classification

To investigate how well the various semantic spaces perform on document classification, we first construct document-level representations by summing the vector representations for all words in a given document. After setting aside a small development set for tuning the hyperparameters of the supervised algorithm, we train a support vector machine (SVM) classifier with a linear kernel and evaluate document topic classification accuracy using ten-fold cross-validation.

The results are reported in the rightmost column of Table 2. Relatedness-specialized embeddings perform better on document topic classification than similarity embeddings, except with graph-based retrofitting, which in fact performs below the standard skip-gram model. The joint-learning model with all relevant free association norms presented as context for each target word is the best performing model. The differences in the table appear small, but the dataset contains more than 10,000 documents, so every percentage point is worth more than 100 documents. Joint learning while presenting all relevant association norms for each target word performs best on this task.

5 Conclusion

We have demonstrated the advantage of specializing embeddings for the tasks of genuine similarity and relatedness. In doing so, we compared two retrofitting methods and a joint learning approach. Specialized embeddings outperform standard embeddings by a large margin on instrinsic similarity and relatedness evaluations. We showed that the difference in how embeddings are specialized carries to downstream NLP tasks, demonstrating that similarity embeddings are better at the TOEFL synonym selection task and relatedness embeddings at a document topic classification task. Lastly, we varied the number of iterations that we use for retrofitting, showing that performance could be improved even further by going over several iterations of the semantic resource.

Acknowledgments

DK is supported by EPSRC grant EP/I037512/1. FH is supported by St Johns College, Cambridge. SC is supported by ERC Starting Grant DisCoTex

(306920) and EPSRC grant EP/I037512/1. We thank Yoshua Bengio, Kyunghyun Cho and Ivan Vulić for useful discussions and the anonymous reviewers for their helpful comments.

References

Eneko Agirre, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *NAACL*, pages 19–27.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artifical Intelligence Research*, 49:1–47.

Stephen Clark. 2015. Vector Space Models of Lexical Meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics*, chapter 16. Wiley-Blackwell, Oxford.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 25–32.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120.

Zelig Harris. 1954. Distributional Structure. *Word*, 10(23):146—162.

Felix Hill, Kyunghyun Cho, Sébastien Jean, Coline Devin, and Yoshua Bengio. 2014a. Embedding word similarity with neural machine translation. *CoRR*, abs/1412.6448.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014b. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Kyo Kageura, Keita Tsuji, and Akiko N Aizawa. 2000. Automatic thesaurus generation through multiple filtering. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 397–403.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, Arizona, USA.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. In *Proceedings of ICLR*, Scottsdale, Arizona, USA.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, Seattle, WA.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL*, pages 384–394.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, 37(1):141–188, January.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4).