Smoothing Syntax-Based Semantic Spaces: Let The Winner Take It All

Sebastian Padó* **Jan Šnajder†** **Jason Utt*** **Britta D. Zeller***
* Institute for Natural Language Processing, University of Stuttgart
{sebastian.pado,jason.utt,britta.zeller}@ims.uni-stuttgart.de
† Faculty of Electrical Engineering and Computing, University of Zagreb
jan.snajder@fer.hr

Abstract

Syntax-based semantic spaces are more flexible and can potentially better model semantic relatedness than bag-of-words spaces. Their application is however limited by sparsity and restricted coverage. We address these problems by smoothing syntax-based with word-based spaces and investigate when to choose which prediction. We obtain the best results by picking the maximal predicted similarity for each word pair, taking advantage of the tendency of unreliable models to underestimate similarity. We show that smoothing can substantially improve coverage while maintaining prediction quality on two German benchmark tasks.

1 Introduction

Distributional semantics (Turney and Pantel, 2010) assumes that the semantic similarity between words is correlated with usage in the same linguistic contexts. Words can be represented by vectors of their co-occurrence frequencies with context elements.

Two major types of models used today are (a) *bag-of-words (BOW)* models, which use words within a surface window around the target word as contexts, and (b) *syntax-based* models, whose contexts include dependency information. These two types can be found among count models such as those studied here as well as newer predictive models (Mikolov et al., 2013; Levy and Goldberg, 2014).

There is an inherent trade-off between BOW models and syntax-based models: Syntax-based models build on a rich, structured notion of context and can capture fine-grained semantic phenomena such as predicate-argument plausibility (Baroni and Lenci, 2010) and can be considered as allowing more representative semantic similarity predictions. At the same time, syntax-based spaces are more prone to *sparsity problems*: Syntactic co-occurrences are less frequent, and the spaces are very high-dimensional. Vectors for rare words can be so sparse that there is no overlap with any other word, and the words effectively fall out of coverage, resulting in less reliable performance. In contrast, BOW models have almost perfect coverage, but provide a more coarse-grained semantic similarity.

This situation raises the question of how different models of differing levels of granularity can be combined in a globally beneficial manner. There is a research tradition that has developed strategies to unify different input vector spaces into a joint output representation. Andrews et al. (2009) combine feature norms with distributional information. Bruni et al. (2011) experiment with textual and visual distributional features. Fyshe et al. (2013) use word-based and dependency-based features as sources of topical and relational information. All of these studies assume that the information provided by the "input" spaces is of comparable quality, but contains different types of information, and can therefore be combined on equal footing – by dimensionality reduction, feature collation, or even simple addition.

Our work assumes a different point of view, namely that there is an *accuracy-coverage trade-off* among our input spaces, as described above. This resembles the situation in n-gram language modeling where models are typically combined by *smoothing*. We also frame the combination of distributional models a as smoothing problem, combining models not at the level of co-occurrence information, but at the level of *predictions*. To our knowledge, few studies have taken this perspective, with the exception of Utt and Padó (2014) who combine cross-lingual and monolingual syntax-based models, and Padó et al. (2013) who use morphological information for smoothing.

We experiment with two smoothing strategies:

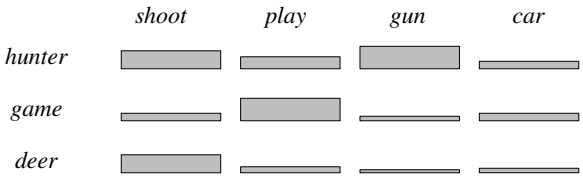

Table 1: Example of a bag-of-words space.

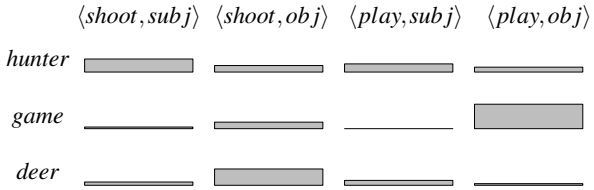

Table 2: Example of a syntax-based space.

Backoff and a *score maximization* strategy, which chooses the highest predicted score. Its intuition is that unreliable distributional models tend to underestimate semantic similarity. Experiments on two German benchmark tasks (semantic similarity prediction and synonym choice) show that score maximization can combine the high precision of syntax-based spaces with the high coverage of BOW spaces.

2 Smoothing Vector Spaces

2.1 Types of Vector Spaces

We concentrate on the two major types: bag-of-words (BOW) and syntax-based models.

BOW models represent target words in terms of context words co-occurring within a surface window. These models are simple, robust, and can be built from any tokenized corpus. They typically have a very high coverage (close to 100%). Different tasks require different context window sizes (Peirsman et al., 2008). Applying dimensionality reduction methods like Singular Value Decomposition (SVD) generally improves space quality.

Syntax-based models are based on word-link-word triples, typically dependency links. This versatile context makes them applicable to languages with free word order and allows them to capture structure-dependent semantic phenomena (Baroni and Lenci, 2010). At the same time, they are much sparser than BOW models, with a lower coverage overall (often 50–70%), which in particular makes the modeling of rare targets problematic. Also, their construction requires a large, well-parsed corpus, which has limited large-scale construction of syntax-based models to few languages (Baroni and Lenci, 2010; Padó and Utt, 2012; Šnajder et al., 2013). Utt and Padó (2014) proposed a cross-lingual method to induce syntax-based models without a parsed corpus, essentially "translating" existing English models. The filter effect created by the use of bilingual lexicon information amplifies the properties of syntax-based models: an even higher quality at the cost of a lower coverage.

2.2 Combining Vector Spaces

As stated above, we assume that there is an *accuracy-coverage* tradeoff between types of vector spaces. Thus, we do not want to unify the individual spaces, but combine their predictions in a sensible way.

Backoff. Backoff and interpolation are two methods that are standardly applied for smoothing in language modeling (Chen and Goodman, 1998). Given our assumptions, Backoff is a straightforward baseline method for combining semantic spaces. It simply defines a linear order on the models and predicts the first model in this order that makes a prediction. This approach was also followed by Utt and Padó (2014).

Score maximization. We propose a second smoothing strategy, *score maximization* or MAX, which chooses the maximum score from the predictions of individual models for each word pair. This strategy is motivated by the hypothesis described in Section 4.

3 Experimental Setup

Tasks. We evaluate on two German lexical-semantic benchmark tasks. The first one is semantic similarity prediction on the Gur350 wordsim dataset (Zesch et al., 2007).[1] It consists of 350 German word pairs with human relatedness ratings on a five-point scale.

The second task is synonym choice: For a target word, its synonym has to be picked from a list of four candidates. We use the German Reader's Digest Word Power dataset (Wallace and Wallace, 2005)[2] with 984 items. It is comparable to the English TOEFL dataset (Landauer and Dumais, 1997), but includes some short phrases as candidates.

[1] Available from: `http://goo.gl/3Dflf1`
[2] Available from: `http://goo.gl/PN42E`

Models. We experiment with three state-of-the-art count models. (1), the BOW space was built from the 800M-token German web corpus SDEWAC (Faaß et al., 2010) using a symmetric context window of size two. A space was extracted with 10k nouns, verbs and adjectives as dimensions, and reduced to 500 dimensions using SVD. (2), the monolingual syntax-based space, "DM", is the German version of Distributional Memory (Baroni and Lenci, 2010), DM.de (Padó and Utt, 2012), induced from a dependency-parsed version of the same corpus. (3), the cross-lingual DM, "tDM", was obtained via translation of the English DM (Utt and Padó, 2014) using the `dict.cc` EN-DE translation lexicon.

We apply both Backoff and score maximization. Model predictions are standardized before smoothing. For Backoff, we assume the linear order (3)>(2)>(1), since (3) has the highest quality, (1) the largest coverage, and (2) assumes an intermediate position. MAX is order-invariant.

Points of Comparison. We consider *random* (for synonym choice) and *frequency* baselines. For word similarity, the frequency baseline predicts the smaller of the two words' frequencies, $\min(f(w_1), f(w_2))$. For synonym choice, it predicts the candidate with the highest frequency. We also compare against current results from the literature, namely UP14 (Utt and Padó, 2014) and PSZ13 (Padó et al., 2013).

Prediction and Evaluation. We compute semantic similarity as cosine similarity. In the case of phrases, we compute the maximum pairwise word similarity. We make a prediction if both words are represented in the model and their vectors have a non-zero cosine. For synonym choice, we make a prediction for an item if we can make a prediction for at least one target–candidate pair.

On both tasks, we compute model coverage, defined as the percentage of items for which a prediction is made. On the similarity task, we measure quality as the Pearson correlation between human rating and model prediction. On the synonym choice task, we compute the accuracy of the covered items with partial credit for ties, following Mohammad et al. (2007). We report performance on all items as well as on the respective subset of covered items. We perform significance testing with bootstrap resampling (Efron and Tibshirani, 1993) on all items.

4 Underestimation Hypothesis

Informally, we believe that noise (e.g., from preprocessing) and sparsity (a perennial issue in distributional semantics) are quite unlikely to increase similarity by chance. To the best of our knowledge, this hypothesis has not been considered yet in the literature:

> **Underestimation hypothesis (UEH).** Unreliable distributional models are more likely to *underestimate* rather than overestimate semantic similarity.

We first develop a geometrical intuition and then corroborate our intuitions with an empirical study.

Geometrical argument . We assume that unreliable distributional models essentially mismeasure co-occurrence frequencies: They do not yield the *ideal vector* $\mathbf{v}$ for a given word, but an *empirical vector* $\hat{\mathbf{v}} = \mathbf{v} + \varepsilon$ that includes a noise vector ε.

We are interested in knowing when cosine similarity decreases due to noise ($cos(\mathbf{v}, \mathbf{w}) > cos(\hat{\mathbf{v}}, \mathbf{w})$). This can be determined by assuming (without loss of generality) that $\mathbf{v}$, $\hat{\mathbf{v}}$, and $\mathbf{w}$ are normalized. This makes them points on the unity hypersphere. Then the cosine decreases if and only if the "empirical" angle $\hat{\alpha}$ between $\hat{\mathbf{v}}$ and $\mathbf{w}$ is larger than the "ideal" angle α between $\mathbf{v}$ and $\mathbf{w}$. As Figure 1 shows, this is the case outside a hypersphere segment of width 2α centered on $\mathbf{w}$. If this segment is maximally wide (180°) if $\alpha = 90°$, it is equally likely that the cosine decreases or increases (in the absence of assumptions on ε). For all smaller angles α, the segment shrinks, and it becomes ever more likely that the cosine decreases, until it necessarily decreases for $\alpha = 0°$ (cf. Fig. 1).

Experimental support for UEH. In order to substantiate the claim of UEH, we designed the following experiment. Ideal vectors are simulated using the entire SDeWaC corpus, giving *'full sims'* for our word pairs. We also construct two halved subspaces by randomly assigning sentences to each half. Word similarities obtained from these two subspaces are termed *'half sims'*. If UEH is true, we would expect the half sims to be, more often than not, lower than the corresponding full sim. A t-test on Gur350 word pairs between full sims and half sims[3] shows a highly significant underes-

[3] As we have two half-sized subspaces, we double the number of wordpairs, pairing each full sim once with half sims

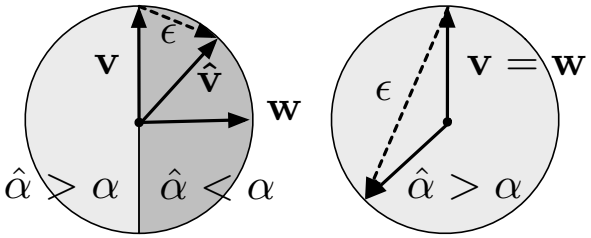

Figure 1: Underestimation hypothesis: ideal and empirical vectors ($\mathbf{v}, \hat{\mathbf{v}}$), point of comparison ($\mathbf{w}$), noise vector ($\varepsilon$). Segments of the hypersphere where angle decreases (dark grey) and increases (light grey). Left: $\alpha = 90°$ (lower $\hat{\alpha}$), Right: $\alpha = 0°$ (higher $\hat{\alpha}$).

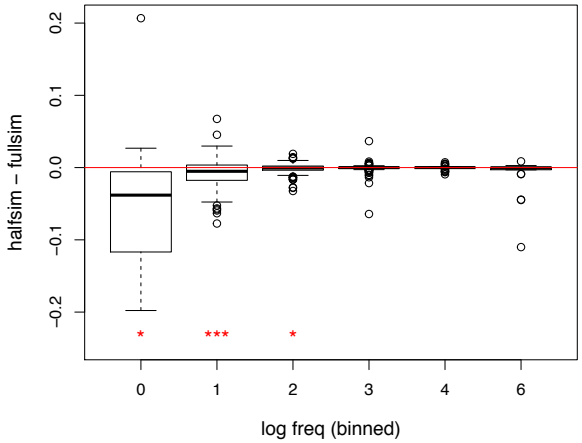

Figure 2: Differences between full and half sims by log frequency of word pairs. (Significance levels are shown for paired t-tests within each bin.)

timation ($t = -4.3647, df = 675, p = 1.473e-05$, mean difference: -0.003277829).

In a second analysis, we test further whether we can further isolate lower frequency word pairs as more reliably showing underestimation. This would correspond to the subnotion with UEH that less evidence for – or more noise in – the representations will intensify the underestimation.

Upon binning word pairs by minimum log frequency, we see that (cf. Figure 2) indeed lower frequency word pairs suffer more from underestimation.

We conclude that, if any of the models predicts a higher similarity, this is a more reliable signal and should be used at the exclusion of others.

from the first subspace, as well as the second. Uncovered items are excluded, in total $df + 1 = 676$ similarity pairs are tested.

	Word similarity			Synonym choice		
Model	r	r_{cov}	cov	acc	acc_{cov}	cov
Random	–	–	–	.25	.25	1
Frequency	.13	.13	1	.31	.31	1
BOW	.34	.34	.97	.52	.53	.95
DM	.38	.43	.60	.48	.53	.84
tDM	.33	.49	.49	.46	**.61**	.58
Smoothed models (sequence tDM>DM>BOW)						
Backoff	.40	.41	**.98**	.56	.57	**.97**
MAX	**.49**	**.50**	**.98**	**.57**	.59	**.97**
Results from the literature						
[UP14]	.42	.47	.69	.55	.59	.89
[PSZ13]	.47	NA	.89	.51	NA	.87

Table 3: Results for baselines and individual models (top), smoothed models (middle) and literature (bottom). Best results per column shown in boldface.

5 Results and Discussion

Table 3 shows the results. All individual models clearly outperform the baselines. Their individual performance matches our accuracy-coverage tradeoff assumptions from above. For example, on the word similarity task, coverage ranges between 97% (BOW) and 49% (tDM). On the covered items, the quality of the tDM predictions outperforms DM, which in turn outperforms BOW (r=.49/.43/.34). The patterns for synonym choice are parallel but less extreme.

The smoothing combination of the three models (tDM>DM>BOW) improves substantially over individual models.[4] In terms of the combination strategy, MAX yields higher results than Backoff.[5] For both tasks, MAX improves highly significantly on all items over the best individual model (word similarity: +.11 r vs. DM; synonym choice: +.05 accuracy vs. BOW; both significant at p<0.01). MAX also outperforms smoothing studies from the literature.

We see different results for the two tasks. On word similarity, smoothing has a larger impact, and the benefit of MAX over Backoff is significant only here (p<0.01). This can be explained by their properties. For word similarity, a regression task, each

[4] In preliminary experiments with the alternative approach of model unification (cf. Section 1), we did not find a comparable benefit for vector concatenation and PCA. This further bolsters our argument from Section 1.

[5] Other combination functions such as arithmetic, geometric and harmonic mean were also tested which however did not provide improvements, in line with UEH.

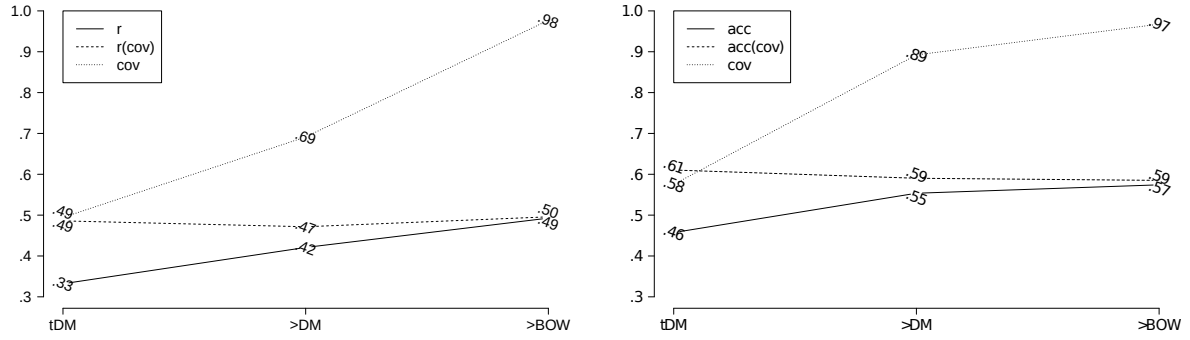

Figure 3: Performance of incremental smoothing (tDM, tDM>DM, tDM>DM>BOW) using score maximization (MAX) for the word similarity (left) and synonym choice (right) tasks

changed prediction influences the evaluation. In synonym choice, a classification task, it only matters which candidate has the highest similarity to the target – the similarities and margins are irrelevant. Consequently, classification is less sensitive to vector changes. This can be observed in practice: Backoff and MAX predictions differ on 155 of 350 word similarity pairs, while the predicted synonym changes only for 52 of 984 targets, i.e., the predictions are almost identical.

Figure 3 shows a more detailed analysis of smoothing. It plots the performance and coverage of MAX for incremental smoothing steps starting from tDM through tDM>DM to tDM>DM>BOW. The plots notably show that the quality on all items increases when adding more models while the quality on the covered items stays almost constant. This shows the robustness of MAX smoothing: The resultant models combine the almost perfect coverage of BOW models with the quality of syntax-based models.

6 Conclusions

This paper investigates the combination of accurate but sparse syntax-based semantic spaces with high-coverage BOW spaces, framing this problem as a smoothing task. We have shown how to reliably smooth by choosing the maximal prediction made by any model. This approach, a "winner-take-all" strategy, exploits the tendency of unreliable distributional models to underestimate semantic similarity making it possible to combine the benefits of different model types, improving both accuracy and coverage across two different semantic tasks and outperforming previous smoothing results. Due to the general nature of the factors giving rise to the underestimation – noise and sparsity in vector representations – we believe that our insights are applicable beyond the models considered in this paper, e.g., to syntax-based continuous vector spaces (Levy and Goldberg, 2014) and document-level models (Landauer and Dumais, 1997).

Acknowledgements

This research is partially funded by the German Research Foundation (SFB 732 at Stuttgart, projects B9 and D10). The second author has been supported by the Croatian Science Foundation under the project UIP-2014-09-7312. We thank the reviewers of this and two other conferences for valuable feedback.

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

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

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32, Edinburgh, UK.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and application of a gold standard for morphological analysis: SMOR in validation. In *Proceedings of LREC*, pages 803–810.

Alona Fyshe, Brian Murphy, Partha Talukdar, and Tom Mitchell. 2013. Documents and dependencies: an exploration of vector space models for semantic composition. In *Proceedings of CoNLL*, pages 84–93, Sofia, Bulgaria.

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):211–240.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308, Baltimore, MD.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, Lake Tahoe, NV.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the 2007 Joint Conference on EMNLP and CoNLL*, pages 571–580, Prague, Czech Republic.

Sebastian Padó and Jason Utt. 2012. A Distributional Memory for German. In *Proceedings of KONVENS 2012 Workshop on Lexical-semantic Resources and Applications*, pages 462–470, Vienna, Austria.

Sebastian Padó, Jan Šnajder, and Britta Zeller. 2013. Derivational smoothing for syntactic distributional semantics. In *Proceedings of ACL*, pages 731–735, Sofia, Bulgaria.

Yves Peirsman, Kris Heylen, and Dirk Geeraerts. 2008. Size matters: Tight and loose context definitions in English word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics – Bridging the Gap Between Semantic Theory and Computational Simulations*, pages 34–41, Hamburg, Germany.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a Distributional Memory for Croatian. In *Proceedings of ACL*, pages 784–789, Sofia, Bulgaria.

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.

Jason Utt and Sebastian Padó. 2014. Crosslingual and multilingual construction of syntax-based vector space models. *Transactions of the Association of Computational Linguistics*, 2(Oct):245–258.

DeWitt Wallace and Lila Acheson Wallace. 2005. *Reader's Digest, das Beste für Deutschland*. Verlag Das Beste, Stuttgart.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Comparing Wikipedia and German Wordnet by evaluating semantic relatedness on multiple datasets. In *Proceedings of NAACL/HLT*, pages 205–208, Rochester, NY.