Applying Multi-Sense Embeddings for German Verbs to Determine Semantic Relatedness and to Detect Non-Literal Language

Maximilian Köper and **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{maximilian.koeper,schulte}@ims.uni-stuttgart.de

Abstract

Up to date, the majority of computational models still determines the semantic relatedness between words (or larger linguistic units) on the type level. In this paper, we compare and extend multi-sense embeddings, in order to model and utilise word senses on the token level. We focus on the challenging class of complex verbs, and evaluate the model variants on various semantic tasks: semantic classification; predicting compositionality; and detecting non-literal language usage. While there is no overall best model, all models significantly outperform a *word2vec* single-sense skip baseline, thus demonstrating the need to distinguish between word senses in a distributional semantic model.

1 Introduction

In recent years, a considerable number of semantic tasks and datasets have been developed, in order to evaluate the semantic quality of computational models. These tasks include general predictions of semantic similarity (e.g., relying on *WordSim-353* (Finkelstein et al., 2001) or *SimLex-999* (Hill et al., 2015)); more specific predictions of semantic relation types (e.g., relying on *BLESS* (Baroni and Lenci, 2011) or the *SemRel* database (Scheible and Schulte im Walde, 2014)); predicting the degree of compositionality for complex nouns and verbs; etc. Computational semantic models predominantly make use of the *distributional hypothesis* in some way or the other, assuming that words with similar distributions have related meanings (Harris, 1954; Firth, 1957). Distributional models thus offer a means to represent meaning vectors of words, and to determine their semantic relatedness (Turney and Pantel, 2010).

Up to date, most distributional semantic models (DSMs) that addressed specific semantic tasks have worked on the type level (e.g., Baroni et al. (2014), Köper et al. (2015), Levy et al. (2015), Pennington et al. (2014)). I.e., each word lemma is represented by a weighted feature vector, where features typically correspond to words that co-occur in particular contexts. When using word embeddings to overcome the problematic sparsity of word vectors, the models rely on neural methods to represent words as low-dimensional vectors.

In contrast, distributional semantic models that break down word type vectors to word sense vectors, have predominantly be applied to Word Sense Disambiguation/Discrimination or (Cross-lingual) Lexical Substitution (McCarthy and Navigli, 2007; Mihalcea et al., 2010; Jurgens and Klapaftis, 2013). As to our knowledge, there is little work on DSMs that distinguishes between word senses and addresses various semantic relatedness tasks. Among the few exceptions are Li and Jurafsky (2015) who evaluated multi-sense embeddings on semantic relation identification (for nouns only) and semantic relatedness between sentences, and Iacobacci et al. (2015) who applied multi-sense embeddings to word and relational similarity.

In this paper, we compare and extend approaches to obtain multi-sense embeddings, in order to model word senses on the token level. We focus on the challenging class of complex verbs, and evaluate the model variants on various semantic tasks: semantic verb classification; the prediction of compositionality; and the detection of non-literal language usage. While there is no overall best model, all models significantly outperform a *word2vec* single-sense skip baseline, thus demonstrating the need to distinguish between word senses in a distributional semantic model.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 535–542,
Valencia, Spain, April 3-7, 2017.

2 Multi-Sense Embeddings

We implemented and applied several variants of state-of-the-art methods for obtaining multi-sense embeddings. In this paper, we restrict the selection to models that perform unsupervised and non-parametric sense learning, i.e., methods that learn potentially different numbers of senses per word, using only a corpus but no sense inventory.

(1) Joint learning of sense representations and application of sense disambiguation From this advanced family of multi-sense embedding induction, we applied the non-parametric multiple-sense skip-grams (**NP-MSSG**), cf. Neelakantan et al. (2014), and skip-grams extended by the Chinese Restaurant Process (**CHINRESTP**), cf. Li and Jurafsky (2015).

(2) Successive learning of single-sense representations and sense disambiguation This class of approaches also relies on skip-grams but learns senses only in a later stage. Pelevina et al. (2016) introduced a non-parametric method that computes a graph relying on cosine-based nearest neighbors, after learning single-sense representations. The graph-clustering algorithm *Chinese Whispers* (Biemann, 2006) identifies senses in the graph, to induce multi-sense embeddings by applying a composition function to word senses. We refer to this approach as **CHINWHISP**.

(3) Single-sense representations for multi-sense corpus annotations In this class of techniques, multi-sense embeddings are also learned in a two-stage procedure: In a first stage, a corpus is automatically sense-annotated by appending a sense index to every word token (e.g., $apple_1$, $apple_2$, etc.). In a second stage, standard techniques are applied to learn single-sense representations for the annotated senses in the corpus. Since the annotations distinguish between senses, the "single-sense" representations effectively represent multi-sense embeddings. For example, Iacobacci et al. (2015) perform the first step by using an off-the-shelf word sense disambiguation tool, and the second step by applying Mikolov's *word2vec* tool (Mikolov et al., 2013b; Mikolov et al., 2013a).

We investigate several variants regarding the automatic corpus sense annotation.

(i) Rather than applying an off-the-shelf WSD tool, we apply the topic-based sense learning method from (Lau et al., 2012), the Hierarchical Dirichlet process (**HDP**) (Teh et al., 2004). The HDP mixture model is a natural non-parametric generalization of the Latent Dirichlet allocation (Blei et al., 2003), where the number of topics can be unbounded and learned directly from the data. We apply HDP by extracting every sentence for each verb type from our corpus. We then train HDP individually for each verb. In the last training iteration we mark each occurrence of a verb type in the corpus with the number of the topic that provided the largest membership value for the respective sentence and that topic.

(ii) As an alternative to the topic model, we apply different clustering algorithms, which not only allows more flexibility in the sense classification technique but also regarding the verb features: we represent each verb token by a vector: We look up the individual vector representations of the verb's context words, and create the verb token vector as the average vector of these context words, ignoring the target verb. This simple kind of phrase/sentence representation has been shown to work well on a variety of tasks (e.g., Milajevs et al. (2014), Hill et al. (2016)). In addition, it allows us to compare different types of context features: (a) all nouns in the sentence (NN), and (b) all words in a symmetrical window of size 10, weighted by the exponential decay function (W10EXP), cf. Iacobacci et al. (2016).

For the actual clustering, we compare non-parametric flat and hierarchical methods. As for HDP, we cluster verb tokens separately, and then mark each verb token with a tag corresponding to a cluster number. The number of clusters containing a specific verb type corresponds to its number of senses. For flat clustering, we use **X-MEANS** (Pelleg and Moore, 2000), which extends the standard hard k-means clustering approach into a non-parametric soft clustering. The algorithm includes a search over the number of clusters k, scores each cluster analysis using the Bayesian Information Criterion (BIC), and chooses the model with k clusters based on the best BIC. For hierarchical clustering, we use *balanced iterative reducing and clustering using hierarchies* **BIRCH** (Zhang et al., 1996), a clustering method that makes use of an internal dendrogram tree structure. Incoming data points are inserted into the tree, and then assigned to the closest sub-trees until they arrive at a leaf node. The entire tree structure changes dynamically over time, while new items are added.

3 Experiments

Corpus & Target Verbs As corpus resource for our target verbs as well as for the experimental setup, we use *DECOW14AX*, a German web corpus containing 12 billion tokens (Schäfer and Bildhauer, 2012; Schäfer, 2015). The corpus sentences were morphologically annotated and parsed using *SMOR* (Faaß et al., 2010), *MarMoT* (Müller et al., 2013) and the MATE dependency parser (Bohnet, 2010). Based on the morphological annotation, we extracted the lemmas of all verb types from the corpus with frequencies $>$100 (regarding base verbs) and $>$200 (regarding complex verbs), and all their sentence contexts. The total selection of German verb types contains 11 869 lemmas, including 6 998 complex verbs.

Experiment Setup The different models have multiple parameters. We set the initial vocabulary to the $200K$ most frequent word types, without removing any of the target verb types. The maximum number of senses per verb type was set to 20. We enabled the multi-sense learning only for our target verbs while all other words obtain only a single sense per model. Regarding the skip-gram architecture, we relied on a symmetrical window of size 10, negative sampling with 15 samples, vector dimensionality of 400 and one corpus iteration. Regarding x-Means and BIRCH, we used a maximum of 5 000 randomly chosen contexts to learn the initial centroids/trees, due to the high-dimensional representations of the sentences. All other individual model-specific parameters were set to the default. Our baseline model is a single-sense skip-gram model as obtained by *word2vec*.

Implementations For HDP, we relied on the python implementation from *gensim*[1]. For x-Means, we used the java implementation *ClodHopper*[2]. For BIRCH we used the java implementation *JBIRCH*[3].

4 Evaluation

We evaluate our models on various semantic tasks: general predictions of semantic similarity, and specific tasks regarding complex German verbs, i.e. semantic classification; prediction of compositionality; detection of non-literal language usage. The goal of the evaluation is to explore whether the distinction of verb senses in our multi-sense embedding models leads to an improvement of model predictions across semantic tasks.

Similarity Traditionally, distributional word representations are predominantly evaluated on their ability to predict the degree of similarity for word pairs in existing benchmarks. The predicted degrees of similarity are compared against human similarity ratings. For our German targets, we use the German versions of *WordSim-353* and *SimLex-999* (Leviant and Reichart, 2015). We predict cosine similarity for multi-sense embeddings by computing a sense-weighted average vector for each word. To assess the predictions, we compare them against the gold standard scores using Spearman's Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988).

The results are presented in Table 1. For this general semantic task, the multi-sense embeddings do not provide significant improvements. The best results are achieved by CHINRESTP for *GerSimLex* and X-MEANS(W10EXP) for *GerWS353*, but these results are close to the baselines.

Model	GerWS353	GerSimLex
NP-MSSGR	.62	.42
ChinRestP	.64	**.46**
ChinWhisp	.64	.36
HDP	.63	.45
x-Means(NN)	.64	.43
x-Means(w10Exp)	**.65**	.44
BIRCH(NN)	.63	.44
BIRCH(w10Exp)	.64	.45
Baseline	**.65**	.45

Table 1: Results for the word similarity datasets.

Compositionality Addressing the compositionality of complex words is a crucial ingredient for lexicography and NLP applications, to know whether the expression should be treated as a whole, or through its constituents, and what the expression means. In this evaluation, we predict the degree of compositionality of German complex verbs, i.e., the degree of relatedness between a complex verb and its corresponding base verb (such as *abnehmen–nehmen* 'take over–take', and *anfangen–fangen* 'begin–catch'). The predictions are evaluated against an existing dataset of human ratings on compositionality (Bott et al., 2016), containing a total of 400 German particle verbs

[1] https://radimrehurek.com/gensim/models/hdpmodel.html

[2] https://github.com/rscarberry-wa/clodhopper

[3] https://github.com/perdisci/jbirch

across 11 particle types. The results are presented in Table 2. CHINWHISP performs significantly better than the baseline, while most other models are performing equally to or even inferior to the baseline.

Model	Prediction
NP-MSSGR	.20
ChinRestP	.30
ChinWhisp	**.32**
HDP	.19
x-Means(NN)	.19
x-Means(w10Exp)	.26
BIRCH(NN)	.28
BIRCH(w10Exp)	.26
Baseline	.26

Table 2: Results for predicting compositionality.

Semantic Verb Classification Semantic verb classifications are of great interest to NLP, specifically regarding the pervasive problem of data sparseness in the processing of natural language. Such classifications have been used in applications such as *word sense disambiguation* (Dorr and Jones, 1996; Kohomban and Lee, 2005; McCarthy et al., 2007), *parsing* (Carroll et al., 1998; Carroll and Fang, 2004), *machine translation* (Prescher et al., 2000; Koehn and Hoang, 2007; Weller et al., 2014), and *information extraction* (Surdeanu et al., 2003; Venturi et al., 2009).

We target the semantic classification of German complex verbs by applying hard clustering to multi-sense embeddings, rather than using soft clustering. Focusing on particle verbs across three particles (*ab, an, auf*), we aim to obtain cluster analyses that resemble existing manual sense classifications based on formal semantic definitions (Kliche, 2011; Lechler and Roßdeutscher, 2009; Springorum, 2011). All datasets represent fuzzy gold standards. The *ab* classification contains 205 particle verbs in 9 classes; the *an* classification contains 188 particle verbs in 8 classes; the *auf* classification contains 234 particle verbs in 11 classes. *All* refers to the concatenation of all tasks.

Using multi-sense embeddings in a hard clustering (rather than single-sense embeddings in a soft clustering) avoids the usage of a cluster membership threshold, which most soft clustering algorithms require. In contrast, the clustering algorithm outputs a membership degree for each element and each cluster, i.e., a fuzzy membership. We rely on k-Means for clustering our multi-sense embeddings, and compare against a fuzzy c-Means baseline with single-sense embeddings. (using every possible threshold within a range of $[0.01, 0.99]$ to determine the memberships, and reporting the one providing the highest score). As evaluation measure we relied on *B-Cubed* (Bagga and Baldwin, 1998) and report f-score between the soft extension of precision and recall.

Table 3 presents the results. Overall, CHINRESTP works best, and CHINWHISP and the BIRCH variants work similarly well. NP-MSSGR is worst. A manual inspection revealed that NP-MSSGR assigns many verbs to multiple clusters, resulting in too large and fuzzy clusters.

Model	*ab*	*an*	*auf*	all
NP-MSSGR	.12	.18	.15	.05
ChinRestP	.24	**.31**	.27	**.13**
ChinWhisp	**.26**	.30	**.28**	.11
HDP	.24	.28	.25	.10
x-Means(NN)	.17	.25	.18	.09
x-Means(w10Exp)	.17	.24	.20	.09
BIRCH(NN)	**.26**	.30	.26	.12
BIRCH(w10Exp)	**.26**	.32	.25	.12
Baseline	.25	.26	.19	.11

Table 3: Results for semantic classification.

Detecting Non-Literal Meaning We explore the prediction of literal vs. non-literal language usage of German complex verbs, relying on an existing dataset containing 159 particle verbs within 6 436 sentences (Köper and Schulte im Walde, 2016). Each sentence is annotated on literal vs. non-literal language usage, comprising 4 174 literal and 2 262 non-literal uses across the 159 particle verbs. Köper and Schulte im Walde (2016) relied on the Multinomial Naive Bayes (MNB) classifier by McCallum and Nigam (1998). We applied the same experimental setup using ten-fold cross validation. Further we re-implemented their system as a baseline, using bag-of-words unigram context features, and added sense information based on the embeddings. For a given sentence, we compare which sense vector fits best to the specific context. This is done by computing a cosine similarity score between a verb sense vector $verb_i$ and the vectors of all context words in the sentence. We then add a verb-sense specific token based on the most similar sense embedding to the unigram list. The underlying assumption is that a specific sense is used either in literal or in non-literal usage. When feeding the training data to the classifier, it should thus automatically assign a high probability for features that predominantly occur for the respective classes.

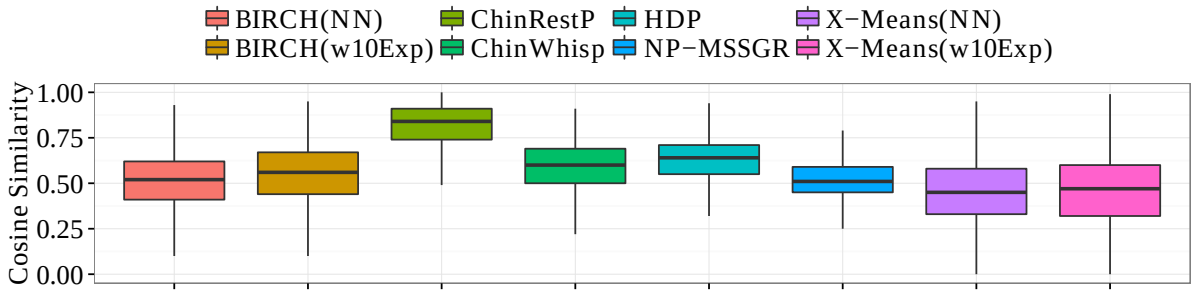

Figure 1: Cosine similarity between all sense pairs within a specific embedding model: many senes are highly similar to each other.

A major difference between our setup and the one by Köper and Schulte im Walde (2016) is the information about the verb itself. In our experiments, the classifier has knowledge about the verb in a sentence, while in their setup the verb has been removed, to avoid learning a verb-specific majority baseline (since some verbs have only literal/non-literal sentences). For this reason, our baseline (i.e., one sense per verb) is already higher than their reported baseline. The remaining parts of our experimental setting are however done as by Köper and Schulte im Walde (2016). To evaluate the classifiers, we calculate the precision, recall and f-score values regarding the non-literal class.

Table 4 shows the results. All multi-sense embedding models clearly outperform the single-sense baseline model. The overall best models are the clustering models X-MEANS and BIRCH.

Model	P	R	F1
NP-MSSGR	90.1	80.3	84.9
ChinRestP	89.0	79.7	84.1
ChinWhisp	90.1	81.2	85.4
HDP	90.8	80.1	85.1
x-Means(NN)	**93.2**	**83.7**	**88.2**
x-Means(w10Exp)	91.9	81.4	86.3
BIRCH(NN)	91.4	81.6	86.2
BIRCH(w10Exp)	91.1	82.7	86.7
Baseline (K&SiW)	91.1	66.0	76.5

Table 4: Results for non-literal language.

5 Discussion & Conclusions

Overall, our experiments demonstrated that the variants of multi-sense embeddings we applied across semantic tasks are successful in comparison to single-sense baselines. In all the tasks we presented, some, most or even all of the multi-sense embeddings outperformed the single-sense baselines, thus demonstrating the need to distinguish between word senses in a distributional semantic model.

The best multi-sense embeddings varied across the semantic tasks. I.e., there was no type of multi-sense embedding that performed superior to all other multi-sense embedding types. Even CHINWHISP, which was among the most successful embeddings across many tasks, exhibited a weakness on one task (i.e., compositionality). We also looked into the inter-sense similarity within the embedding models. Figure 1 presents box-plots on the cosine similarity between all sense pairs within a specific embedding model. The plot shows that overall, the identified senses in the models are quite similar to each other. The strongest inter-sense similarity can be found for CHINRESTP.

Looking into the embeddings across multi-sense approaches, we found that –even though the embeddings were trained on the same data– the average number of senses differs strongly across the embedding models: NP-MSSGR, CHINRESTP and CHINWHISP have an average number of less than 2 senses per word, while the X-MEANS and BIRCH models have an average number between 3.2 and 7.6 senses. Most senses are obtained by HDP (15.4), but many senses received little weight.

This diversity of success across embedding types and semantic tasks demonstrates that an evaluation of semantic models on a general task such as semantic similarity is not sufficient.

Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732 (Maximilian Köper) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde).

References

Amit Bagga and Breck Baldwin. 1998. Entity-based Cross-document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 79–85, Montréal, Canada.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed Distributional Semantic Evaluation. In *Proceedings of the EMNLP Workshop on Geometrical Models for Natural Language Semantics*, pages 1–10, Edinburgh, UK.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A Systematic Comparison of Context-counting and Context-predicting Semantic Vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, MD.

Chris Biemann. 2006. Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems. In *Proceedings of the 1st Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 73–80, Stroudsburg, PA, USA.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Stefan Bott, Nana Khvtisavrishvili, Max Kisselew, and Sabine Schulte im Walde. 2016. G$_h$ost-PV: A Representative Gold Standard of German Particle Verbs. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*, pages 125–133, Osaka, Japan.

John Carroll and Alex C. Fang. 2004. The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 107–114, Sanya City, China.

John Carroll, Guido Minnen, and Ted Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, pages 118–126, Montréal, Canada.

Bonnie J. Dorr and Doug Jones. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 322–327, Copenhagen, Denmark.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 803–810, Valletta, Malta.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414, Hong Kong, Hong Kong.

John R. Firth. 1957. *Papers in Linguistics 1934-51*. Oxford University Press, London, UK.

Zellig Harris. 1954. Distributional Structure. *Word*, 10(23):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating Semantic Models with (genuine) Similarity Estimation. *Computational Linguistics, Volume 41*, pages 665–695.

Felix Hill, KyungHyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 95–105, Beijing, China.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 897–907, Berlin, Germany.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 290–299, Atlanta, Georgia, USA.

Fritz Kliche. 2011. Semantic Variants of German Particle Verbs with *"ab"*. *Leuvense Bijdragen*, 97:3–27.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Upali S. Kohomban and Wee Sun Lee. 2005. Learning Semantic Classes for Word Sense Disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41, Ann Arbor, MI.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual Reliability and "Semantic" Structure of Continuous Word Spaces. In *Proceedings of the 11th Conference on Computational Semantics*, pages 40–45, London, UK.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing Literal and Non-Literal Usage of German Particle Verbs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California, USA.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word Sense Induction for Novel Sense Detection. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France.

Andrea Lechler and Antje Roßdeutscher. 2009. German Particle Verbs with *auf*. Reconstructing their Composition in a DRT-based Framework. *Linguistische Berichte*, 220:439–478.

Ira Leviant and Roi Reichart. 2015. Judgment Language Matters: Multilingual Vector Space Models for Judgment Language Aware Lexical Semantics. *Preprint pubslished on arXiv*, abs/1508.00106.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Jiwei Li and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal.

Andrew McCallum and Kamal Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of the AAAI Workshop on Learning for Text Categorization*, pages 41–48, Budapest, Hungary.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53, Prague, Czech Republic.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics*, 33(4):553–590.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, GA.

Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 708–719, Doha, Qatar.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, WA, USA.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1059–1069, Doha, Qatar.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183, Berlin, Germany, August.

Dan Pelleg and Andrew Moore. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the 17th International Conference on Machine Learning*, pages 727–734, San Francisco.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, pages 1532–1543.

Detlef Prescher, Stefan Riezler, and Mats Rooth. 2000. Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 649–655, Saarbrücken, Germany.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28 – 34.

Silke Scheible and Sabine Schulte im Walde. 2014. A Database of Paradigmatic Semantic Relation Pairs for German Nouns, Verbs and Adjectives. In *Proceedings of the COLING Workshop Lexical and Grammatical Resources for Language Processing*, pages 111–119, Dublin, Ireland.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA.

Sylvia Springorum. 2011. DRT-based Analysis of the German Verb Particle *"an"*. *Leuvense Bijdragen*, 97:80–105.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association, Volume 101*.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Giulia Venturi, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. 2009. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In Alexander Gelbukh, editor, *Linguistics and Intelligent Text Processing*, pages 137–148. Springer, Heidelberg.

Marion Weller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Using Noun Class Information to model Selectional Preferences for Translating Prepositions in SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 275–287, Vancouver, Canada.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 103–114.