Quantifying Context Overlap for Training Word Embeddings

Yimeng Zhuang, Jinghui Xie, Yinhe Zheng and Xuan Zhu
Samsung Research Institute China - Beijing (SRC-B)
{ym.zhuang, jh.xie, yh.zheng, xuan.zhu}@samsung.com

Abstract

Most models for learning word embeddings are trained based on the context information of words, more precisely first order co-occurrence relations. In this paper, a metric is designed to estimate second order co-occurrence relations based on context overlap. The estimated values are further used as the augmented data to enhance the learning of word embeddings by joint training with existing neural word embedding models. Experimental results show that better word vectors can be obtained for word similarity tasks and some downstream NLP tasks by the enhanced approach.

1 Introduction

In the last decade, the distributed word representation (a.k.a word embedding) has attracted tremendous attention in the field of natural language processing (NLP). Instead of large vectors, such as the one-hot representation, the distributed word representation embeds semantic and syntactic characteristics of words into a low-dimensional space, which makes it popular in NLP applications.

The main idea of most word embedding models follows the distributional hypothesis (Harris, 1954), i.e., the embedding of each word may be inferred using its context. An important model family for distributional word representation learning is built based on the global matrix factorization approach (Deerwester et al., 1990; Lee and Seung, 2001; Srebro et al., 2005; Mnih and Hinton, 2007; Li et al., 2015; Wang and Cohen, 2016), in which a dimensionality reduction over a sparse matrix is performed to capture the statistical information about a corpus in low-dimensional vectors. Another model family is neural word embeddings (Levy and Goldberg, 2014b), some attempts include the famous Neural Probabilistic Language Model (Bengio et al., 2003), SGNS and CBOW (Mikolov et al., 2013a,b), GloVe (Pennington et al., 2014) and their variants (Shazeer et al., 2016; Kenter et al., 2016; Ling et al., 2017; Patel et al., 2017).

Most of these models capture the context information of each word using the co-occurrence matrix. However, the co-occurrence matrix only represents relatively local information, i.e., it describes context associations based on word pairs' co-occurrence counts without considering global context perspective. Besides, the co-occurrence matrix is only an estimation of a corpus, which is only a sample of a language. A mass of related word pairs may not be observed in the corpus, and the latent relations between unobserved word pairs may not be modeled well due to the missing knowledge.

Few attempts are carried out to indirectly deal with unobserved co-occurrence for dense neural word embeddings. SGNS (Mikolov et al., 2013a,b) indirectly addresses this problem through negative sampling. Swivel (Shazeer et al., 2016) improves GloVe by using a "soft hinge" loss to prevent from over-estimating zero co-occurrences. However, the latent relations between unobserved word pairs are not explicitly represented. There are also some works around semantic composition and distributional inference (Mitchell and Lapata, 2008; Erk and Padó, 2008, 2010; Reisinger and Mooney, 2010; Thater et al., 2011; Kartsaklis et al., 2013; Kober et al., 2016) that are explored to address the sparseness problem, but they are not designed for training neural word embeddings.

In this paper, we explore an approach that utilizes context overlap information to dig up more effective co-occurrence relations and propose extensions for GloVe and Swivel to validate the positive impact of introducing context overlap.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 587–593
Brussels, Belgium, October 31 - November 4, 2018.

2 Quantify Context Overlap

In this work, we explore quantifying context overlap based on the observation that to a certain extent the overlap of Point-wise Mutual Information (PMI) (Church and Hanks, 1990) reflects context overlap.

As shown in Figure 1, two separate words may exhibit a particular aspect of interest or be semantically related when the overlap area between their PMI is relatively large.

The calculation of complete PMI-weighted context overlap may be time-consuming when the number of words is large. To make the time complexity affordable, only the context words that have strong lexical association with a target word i are considered:

$$S_i = \{k \in V | PMI(i,k) > h_{PMI}\} \qquad (1)$$

in which V is the vocabulary, h_{PMI} is a threshold which acts as a magnitude to shift PMI, and S_i denotes the set that consists of the context words that have enough large PMI values with the target word i. It is expected that most context information associated with the word i can be captured by its PMI values over S_i. Then, we measure the degree of context overlap (CO) between two target words i, j as a function of their PMI values over the intersection of S_i and S_j, i.e.,

$$CO(i,j) = \sum_{k \in S_i \cap S_j} \min(f(PMI(i,k)), f(PMI(j,k))) \qquad (2)$$

where f is a monotonic mapping function to rectify the data characteristics for certain objective function in word embedding training.

Compared to identity function $f(x) = x$, we find exponential function $f(x) = exp(x)$ works much better in our experiments. For the quantized context overlap, the exponential mapping function results in a similar data distribution as the co-occurrence counts, i.e., few word pairs have extremely large values while most word pairs' values are distributed in a relatively small range.

3 Extend to Existing Models

We consider the original co-occurrence matrix as a description of first order co-occurrence relations, while the quantized context overlap as a description of second order co-occurrence relations (Schütze, 1998), i.e., co-co-occurrences, which is represented by "non-logarithmic PMI-weighted context overlap" in this work. The context overlap between two words can be inferred even when they never co-occur in the corpus. According to our statistics, more than 84% word pairs in the second order co-occurrence matrix are not included in the first order co-occurrence matrix. We expect introducing second order co-occurrence relations may enhance the quality of the word embedding that is originally trained on first order co-occurrence relations. GloVe (Pennington et al., 2014) and Swivel (Shazeer et al., 2016) are extended by joint training with context overlap information in this paper.

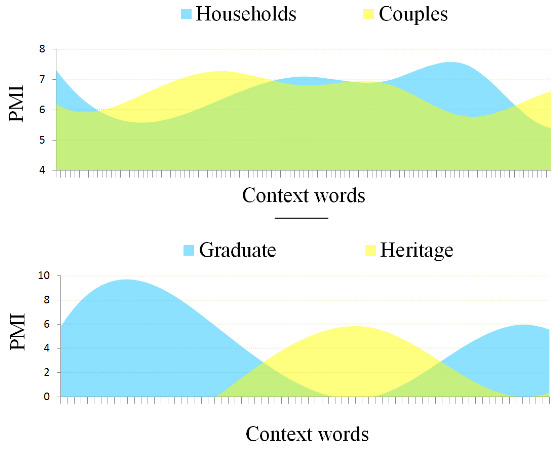

Figure 1: PMI of different words. The x-axis represents a series of context words in a subset of the whole vocabulary, the y-axis denotes PMI values between the target word and context words. (a) The upper part of this figure illustrates the large overlap between semantics related words. (b) The lower part, on the contrary, is an example of relatively unrelated word pair, in which the overlap is relatively small.

GloVe The logarithmic co-occurrence matrix is factorized in GloVe with bias terms, and a weighted least squares loss function is optimized:

$$\mathcal{J}_{GloVe} = \sum_{i,j} \lambda_{ij}(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \quad (3)$$

where X_{ij} denotes the word-context co-occurrence count between a target word i and a context word j. The model parameters to be learned include $w_i \in \mathbb{R}^d$, $\tilde{w}_j \in \mathbb{R}^d$, b_i and $\tilde{b}_j$, which correspond to target word vector, context word vector, bias terms associated with the target word and the context word, respectively. λ_{ij} is a weight whose value equals to $(\min(X_{ij}, x_{max})/x_{max})^\alpha$.

To extend GloVe, two tasks are trained in parallel during the training process: One is the main task that follows the original GloVe training process as above; Another one is an auxiliary task that tunes word embeddings using context overlap. The parameters of word embeddings are shared in both tasks.

Following GloVe-style loss function, in the auxiliary task, the dot products of word vectors are pushed to estimate logarithmic second order co-occurrence.

$$\mathcal{J}^{(2)}_{GloVe} = \sum_{i,j} \lambda^{(2)}_{ij} (Aw_i^T w_j + b_i^{(2)} + b_j^{(2)} - \log X_{ij}^{(2)})^2 \tag{4}$$

where the superscripts $^{(2)}$ are used to differentiate with the terms in the original GloVe. $X_{ij}^{(2)} = CO(i,j)$ represents context overlap, a word independent learnable scale A is adopted to relieve the potential inconformity between first order and second order co-occurrences. The weight $\lambda_{ij}^{(2)}$ is similar to the original λ_{ij}, but using a different hyperparameter $x_{max}^{(2)}$.

The multi-task (Ruder, 2017) loss function is the weighted sum of the two tasks, i.e., $\mathcal{J} = \mathcal{J}_{GloVe} + \beta \cdot \mathcal{J}^{(2)}_{GloVe}$, where the weight β is a hyperparameter.

Swivel As pointed out by (Levy et al., 2015) , if the bias terms in GloVe are fixed to the logarithmic count of the corresponding word, the dot products of target word vectors and context word vectors are almost equivalent to the approximation of logarithmic PMI matrix with a shift of $\log \sum_{i,j} X_{ij}$. Submatrix-wise Vector Embedding Learner (Swivel) directly reconstructs the PMI matrix by dot product between target vectors and context vectors and deals with unobserved co-occurrences using a "soft hinge" loss function. (Shazeer et al., 2016) details its loss functions and training process. In our extended version, we add a supplementary loss function to handle second order co-occurrences. When the second order co-occurrence $X_{ij}^{(2)}$ is more than zero, the PMI of context overlap is approximated.

$$\frac{1}{2}\lambda_{ij}^{(2)} (Aw_i^T w_j + B - PMI^{(2)}(i,j))^2 \tag{5}$$

in which A, B are word independent learnable scale parameters, and $PMI^{(2)}(i,j)$ is the Pointwise Mutual Information computed on the second order co-occurrence matrix $[X_{ij}^{(2)}]$.

4 Experiments

4.1 Setup

Corpus The training dataset contains 6 billion tokens collected from diversified corpora, including the News Crawl corpus (Chelba et al., 2013), the April 2010 Wikipedia dump (Shaoul, 2010; Lee and Chen, 2017), and a year-2012 subset of the Reddit comment datasets [1].

Preprocessing Following (Lee and Chen, 2017), the Stanford tokenizer is used to process the training corpus, which are split into sentences with characters converted to lower cases. Punctuations are removed.

Parameter Configuration The vocabularies are limited to the 200K most frequent words. Following (Pennington et al., 2014), a decreasing weighting function is adopted to construct the co-occurrence matrix. We use symmetric context window of five words to the left and five words to the right.

For GloVe, recommended parameters in (Pennington et al., 2014) are used. Specifically, we set $\alpha = \frac{3}{4}$, $x_{max} = 100$, initial learning rate as 0.05, 100 iterations. For Swivel, recommended parameters in (Shazeer et al., 2016) are used. The weighting function is $0.1 + 0.25x_{ij}^{0.5}$, each shard is sampled about 100 times. But we set the block size as 4000 so that the vocabulary size can be divided exactly.

For the auxiliary tasks, we tune the hyperparameters on the small News Crawl corpus. And we find that in an appropriate range, the threshold h_{PMI} is not sensitive to the performance. In this paper, h_{PMI}, $x_{max}^{(2)}$ and β are set to $\log 100$, 10000 and 0.2 respectively. Since there is no difference between target vectors and context vectors (except random initialization), in order to keep symmetry, we not only approximate context overlap between target vectors, but also approximate context overlap between context vectors simultaneously. Final vectors are the sum of w and $\tilde{w}$ in both GloVe and Swivel.

4.2 Intrinsic Evaluation

Table 1 shows the evaluation results of word similarity tasks and word analogy tasks. Word similarity is measured as the Spearman's rank correlation ρ between human-judged similarity and cosine distance of word vectors. In word analogy

[1] Available at `https://files.pushshift.io/reddit/comments/`

Method	WS353	SL999	SCWS	RW	MEN	MT771	Analogy	
							Sem	Syn
GloVe	66.8	35.0	59.3	44.1	74.7	69.9	76.0	75.3
GloVe + CO	69.7	38.0	63.8	45.1	77.6	71.3	78.6	75.0
SGNS	71.1	40.7	**67.1**	52.8	78.1	70.4	67.2	77.3
Swivel	73.1	39.9	66.4	53.4	79.1	71.7	78.6	78.0
Swivel + CO	**74.0**	**41.2**	66.3	**53.6**	**79.8**	**72.5**	**79.4**	**78.1**

Table 1: Word similarity and analogy results ($\rho \times 100$ and analogy accuracy). We denote context overlap enhanced method with "+ CO". 300-dimensional embeddings are used. The datasets used include WS353 (Finkelstein et al., 2001), SL999 (Hill et al., 2016), SCWS (Huang et al., 2012), RW (Luong et al., 2013), MEN (Bruni et al., 2014), MT771 (Halawi et al., 2012), and Mikolov's analogy dataset (Mikolov et al., 2013a).

task, the questions are answered over the whole vocabulary through 3CosMul (Levy and Goldberg, 2014a). In addition to GloVe and Swivel, the evaluations of SGNS are also reported for reference. We train SGNS with the word2vec tool, using symmetric context window of five words to the left and five words to the right, and 5 negative samples.

As can be seen from the table, the context overlap information enhanced word embeddings perform better in most word similarity tasks and get higher analogy accuracy in semantic aspect at the cost of syntactic score. The improved semantics performance, to a certain extent, reflects second order co-occurrence relations are more semantic.

4.3 Text Classification

Text classification tasks are conducted on five shared benchmark datasets from (Kim, 2014) including binary classification tasks CR (Hu and Liu, 2004), MR (Pang and Lee, 2005), Subj (Pang and Lee, 2004) and multiple classification tasks TREC (Li and Roth, 2002), SST1 (Socher et al., 2013). Texts are preprocessed following the description of Section 4.1. We train Convolutional Neural Networks (CNN) on top of our static pretrained word vectors following (Kim, 2014). To avoid the high-risk of single-run estimate being false (Melis et al., 2017; Reimers and Gurevych, 2017), average classification accuracies of 20 runs are reported as the final scores. The results are shown in Table 2. As can be seen from the results that the enhanced word embeddings outperform the baselines.

Method	CR	MR	SST1	Subj	TREC
GloVe	80.9	76.5	46.9	90.9	89.7
+ CO	81.7†	76.4	47.6†	91.4†	90.2†
Swivel	81.7	76.7	47.9	91.4	90.4
+ CO	82.4†	76.7	48.3†	91.7†	90.5
CBOW	80.6	75.3	46.5	89.8	89.6
SGNS	81.6	77.0	48.0	91.2	90.6

Table 2: Text classification results (Acc.%). Pretrained word vectors with 300 dimensions are reported here. Enhanced runs statistically significantly (t-test, p-value < 0.05) different from the GloVe/Swivel baseline runs are marked with a $\dagger$. The results of CBOW and SGNS are also given for reference.

5 Model Analysis

As it is known to all, word frequency plays an important role in the computation of word embeddings (Gittens et al., 2017). Inspired from the graph in (Shazeer et al., 2016), relations between word analogy accuracy and the log mean frequency of the words in analogy questions and answers are plotted on Figure 2. The word embeddings trained by GloVe with or without context overlap information are used here.

An obvious semantic performance improvement is observed in the range of low frequency. Our observation of second order co-occurrences may explain this fact. We randomly sample 1 million word pairs, and rank these word pairs in descending order by their quantized context overlap. In all the word pairs, average word frequency is 13934.4. However, it is only 1676.1 in the top 0.1% word pairs, it is 3984.8 in the top 1%, and it is 7904.9 in the top 10%. This may be caused by PMI's bias towards infrequent words, but it illustrates infrequent words carry more information in second order co-occurrence relations.

6 Conclusion

In this paper, we propose an empirical metric to enhance the word embeddings through estimating second order co-occurrence relations using con-

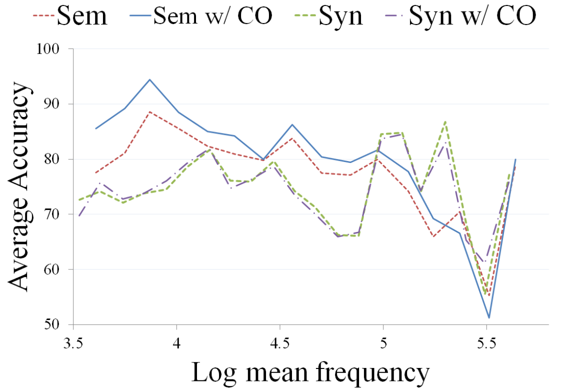

Figure 2: Relations between word analogy accuracy and the log mean frequency.

text overlap. Instead of only local statistical information, context overlap leverages global association distribution to measure word pairs correlation.

The proposed method is easy to extend to existing models, such as GloVe and Swivel, by an auxiliary objective function. The improvement in experimental results helps to validate the positive impact of introducing quantized context overlap.

We have considered the feasibility of enriching SGNS and CBOW with information from context-overlap. However, because of their training mode, we can't remake them in a straightforward way following their "original spirit". When training SGNS and CBOW, the program scans the training text. The target and context words are chosen using a slide window and negative sampling is used. In this process, no co-occurrence matrix is explicitly computed, and we fail to extend it in a united form as we extend GloVe and Swivel. The extensions for GloVe and Swivel can also be used for reference for extending other word embedding approaches that are trained on co-occurrence matrix. The exploration for second order co-occurrence can be traced back to 1990s. We think it is helpful to revive the classical method in a modern, embedding driven way. How to integrate second order co-occurrence information for approaches like SGNS, CBOW should be an interesting future work.

As future works, we suggest further investigating the characteristics of context overlap in diversified ways.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906. Association for Computational Linguistics.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the acl 2010 conference short papers*, pages 92–97. Association for Computational Linguistics.

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. ACM.

Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. 2017. Skip-gram-zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 69–76.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.

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

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 114–123.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Thomas Kober, Julie Weeds, Jeremy Reffin, and David Weir. 2016. Improving sparse word representations with distributional inference for semantic composition. *arXiv preprint arXiv:1608.06794*.

Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Guang-He Lee and Yun-Nung Chen. 2017. Muse: Modularizing unsupervised sense embeddings. *arXiv preprint arXiv:1704.04601*.

Omer Levy and Yoav Goldberg. 2014a. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

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.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Yitan Li, Linli Xu, Fei Tian, Liang Jiang, Xiaowei Zhong, and Enhong Chen. 2015. Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *IJCAI*, pages 3650–3656.

Yuan Ling, Yuan An, Mengwen Liu, Sadid A Hasan, Yetian Fan, and Xiaohua Hu. 2017. Integrating extra knowledge into word embedding models for biomedical nlp tasks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 968–975. IEEE.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648. ACM.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics.

Kevin Patel, Divya Patel, Mansi Golakiya, Pushpak Bhattacharyya, and Nilesh Birari. 2017. Adapting pre-trained word embeddings for use in medical coding. *BioNLP 2017*, pages 302–306.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Cyrus Shaoul. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta*.

Noam Shazeer, Ryan Doherty, Colin Evans, and Chris Waterson. 2016. Swivel: Improving embeddings by noticing what's missing. *arXiv preprint arXiv:1602.02215*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. 2005. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143.

William Yang Wang and William W Cohen. 2016. Learning first-order logic embeddings via matrix factorization. In *IJCAI*, pages 2132–2138.