Integrating Semantic Knowledge into Lexical Embeddings Based on Information Content Measurement

Hsin-Yang Wang
Institute of Information Science
Academia Sinica
Nankang, Taipei, Taiwan
wang@iis.sinica.edu.tw

Wei-Yun Ma
Institute of Information Science
Academia Sinica
Nankang, Taipei, Taiwan
ma@iis.sinica.edu.tw

Abstract

Distributional word representations are widely used in NLP tasks. These representations are based on an assumption that words with a similar context tend to have a similar meaning. To improve the quality of the context-based embeddings, many researches have explored how to make full use of existing lexical resources. In this paper, we argue that while we incorporate the prior knowledge with context-based embeddings, words with different occurrences should be treated differently. Therefore, we propose to rely on the measurement of information content to control the degree of applying prior knowledge into context-based embeddings - different words would have different learning rates when adjusting their embeddings. In the result, we demonstrate that our embeddings get significant improvements on two different tasks: Word Similarity and Analogical Reasoning.

1 Introduction

Distributed word representation maps each word into a real-valued vector. The produced vector has implied the abstract meaning of the word for their syntactic (Collobert and Weston, 2008; Luong et al., 2013; Mnih and Hinton, 2007; Turian et al., 2010) and semantic (Huang et al., 2012; Socher et al., 2013b) information. These vectors have been used as features in a variety of applications, such as information retrieval (Salton and McGill, 1984), document classification (Sebastiani, 2002), question answering (Tellex et al., 2003), name entity recognition (Turian et al., 2010), and syntactic parsing (Socher et al., 2013a).

In past few years, several unsupervised methods for word embeddings (Collobert et al., 2011; Dhillon et al., 2012; Lebret and Collobert, 2014; Li and Zhang, 2015; Mikolov et al., 2013a; Pennington et al., 2014) have been proposed and have had great results in various evaluations. Through exploiting local context of target words, these algorithms learn word embeddings by maximizing the contextual distribution of a large corpus.

Knowledge bases provide rich semantic relatedness between words, which are more likely to capture the desired semantics on certain NLP tasks. To improve the quality of context-based embeddings, some researchers attempted to incorporate knowledge base, such as WordNet (Miller, 1995) and Paraphrase Database (Ganitkevitch et al., 2013) into the learning process. Recent work has shown that aggregating the knowledge base information into context-based embeddings can significantly improve the embeddings (Bian et al., 2014; Chang et al., 2013; Faruqui et al., 2015; Xu et al., 2014; Yih et al., 2012; Yu and Dredze, 2014).

One implicit but critical reason of the success on using knowledge bases, based on our insight, is that knowledge bases can complement the embedding quality of those words which lack enough statistics of word occurrences, such as enough occurrences or diversity of their context. These words may suffer the difficulty obtaining meaningful information from the given corpus. Following this idea, we argue that while incorporating prior knowledge into context-based embeddings, words with different statistics of word occurrences should be treated differently. With this idea, we propose to rely on the measurement of information content to control the degree of applying prior knowledge into context-based embeddings.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 509–515,
Valencia, Spain, April 3-7, 2017.

2 Learning Embeddings

In this section, we will first review word2vec, a popular context-based embedding approach, and then introduce Relation Constrained Model (RCM) to incorporate prior knowledge. Finally we propose our approach to utilize the both two models, making words with different statistics of word occurrences be treated differently while incorporating prior knowledge.

2.1 Context-based Embedding

Context-based embedding has two main model families: *global matrix factorization methods*, such as latent semantic analysis (LSA) (Bullinaria and Levy, 2007; Lebret and Collobert, 2014; Pennington et al., 2014; Rohde et al., 2006) and *local context window methods* (Bengio, 2013; Collobert and Weston, 2008; Mikolov et al., 2013a). Both training models learn the embedding by using the statistical information of the word context from a large corpus. In this paper, we adopt continuous bag-of-word (CBOW) in word2vec (Mikolov et al., 2013a) as our context-based embedding model. CBOW is an unsupervised learning algorithm using a neural language models, given a target word w_t and its c neighboring words, the model is aimed at maximizing the log-likelihood of each word given its context.

The objective function is shown as following:

$$J = \frac{1}{T}\sum_{t=1}^{T} \log p\left(w_t | w_{t-c}^{t+c}\right) \tag{1}$$

In CBOW, $p\left(w_t | w_{t-c}^{t+c}\right)$ defined as:

$$\frac{\exp({e'_{w_t}}^{\top} \cdot \sum_{-c\le j\le c, j\neq 0} e_{w_{t+j}})}{\sum_w \exp({e'_w}^{\top} \cdot \sum_{-c\le j\le c, j\neq 0} e_{w_{t+j}})} \tag{2}$$

where e_w and e'_w represent the input and output embeddings respectively.

CBOW use stochastic gradient descent to learn embeddings, the update of e'_w and e_{w_j} are:

$$e'_w - \alpha(\sigma(f(w)) - \mathbb{I}_{[w=w_t]}) \cdot \sum_{j=t-c}^{t+c} e_{w_j} \tag{3}$$

$$e_{w_j} - \alpha \sum_w (\sigma(f(w)) - \mathbb{I}_{[w=w_t]}) \cdot e'_w \tag{4}$$

where

$$\sigma(x) = \exp\{x\}/(1+\exp\{x\}) \tag{5}$$

$I_{[x]}$ is 1 when x is true, $f(w) = {e'_w}^{\top} \sum_{j=t-c}^{t+c} e_{w_j}$, α is learning rate.

2.2 Relation Constrained Model(RCM)

RCM (Yu and Dredze, 2014) designed a simple but effective method to incorporate prior knowledge into context-based embeddings. Given a set of relation pairs (w, w_i) in a given knowledge base, by maximizing the log probability of w and w_i, the model aims to increase the similarity between w and w_i. To simplify the formula, we can define $\mathbb{R}$ as a set of relations between w and w_i. $\mathbb{R}_w$ is the subset of $\mathbb{R}$ which involve word w.

The objective function is shown as following:

$$J = \frac{1}{N}\sum_{i=1}^{N} \sum_{w\in\mathbb{R}_{w_i}} \log p\left(w|w_i\right) \tag{6}$$

where

$$p(w|w_i) = \exp({e'_w}^{\top} e_{w_i}) / \sum_{\bar{w}} \exp({e'_{\bar{w}}}^{\top} e_{w_i}) \tag{7}$$

The objective function of RCM is similar to the CBOW but without the context. RCM only revise output embeddings e'_w and e'_{w_i} when it trains with CBOW jointly.

RCM use stochastic gradient descent to learn embeddings, the update of e'_w and e'_{w_i} are:

$$e'_w - \alpha(\sigma(f'(w)) - \mathbb{I}_{[w\in\mathbb{R}_{w_i}]}) \cdot e'_{w_i} \tag{8}$$

$$e'_{w_i} - \alpha \sum_w (\sigma(f'(w)) - \mathbb{I}_{[w\in\mathbb{R}_{w_i}]}) \cdot e'_w \tag{9}$$

where

$$\sigma(x) = \exp\{x\}/(1+\exp\{x\}) \tag{10}$$

$I_{[x]}$ is 1 when x is true, $f'(w) = {e'_w}^{\top} e'_{w_i}$, α is the learning rate.

2.3 Information Content Measurement

No matter which kind of context-based embedding approach, statistics of word occurrences play a primary role. Under this statement, the embedding quality of those words which lack enough statistics of word occurrences, such as enough occurrences or diversity of their context, may suffer the difficulty obtaining meaningful information from the given corpus. We argue that while incorporating prior knowledge into context-based embeddings, words with different statistics of word occurrences should be treated differently. With this idea, we investigate several score functions S_{IC} to adjust the learning rate, aiming to make words with less

statistical information be adjusted more via prior knowledge, and words with richer statistical information be adjusted less.

The update formula of e'_w and e'_{w_i} are:

$$e'_w - (S_{IC}(w, w_i) * \alpha)(\sigma(f'(w)) - \mathbb{I}_{[w \in \mathbb{R}_{w_i}]}) \cdot e'_{w_i} \quad (11)$$

$$e'_{w_i} - (S_{IC}(w_i, w) * \alpha) \sum_w (\sigma(f'(w)) - \mathbb{I}_{[w \in \mathbb{R}_{w_i}]}) \cdot e'_w \quad (12)$$

In this paper, we propose three kinds of score functions to control the adjustment: **Threshold**, **Function(Freq.)**, and **Function(Ent.)**.

a. Threshold: The first one is a binary indicator based on a threshold of word frequency. We can distinguish the word relations into two groups.

$$S_{IC}(w, w_i) = \begin{cases} 1, & if\ f_w < f_{thres.}\ and\ f_{w_i} \geq f_{thres.} \\ 0, & otherwise \end{cases} \quad (13)$$

This strategy will only revise low frequency word in a word relation pair, when one word of the relation word pair has low frequency and the other has high frequency.

b. Function (Freq.): In contrast to the previous strategy, we make the score function smoother, we use a relative value between two words frequencies and a hyperbolic tangent function to determine the score.

$$S_{IC}(w, w_i) = \tanh\left(\frac{f_{w_i}}{f_w}\right) \quad (14)$$

This strategy still can revise relatively lower frequency word in a word relation pair, when one word of the relation word pair has relatively lower frequency and the other has relatively high frequency. This scoring function is based on our assumption that if a word has relatively higher occurrence, its embedding quality is better, so it does not need to be adjusted much.

c. Function (Ent.): In addition to the word's frequency, in fact, we believe that the contextual diversity plays a critical role of affecting the quality of word embedding. Therefore, we propose a score function based on the conditional entropy (information content) from the information theory.

We define the score function as the follows:

$$S_{IC}(w, w_i) = \tanh\left(\frac{H(C|w_i)}{H(C|w)}\right) \quad (15)$$

$$H(C|w) = \sum_j p(c_j, w) \log \frac{p(w)}{p(c_j, w)} \quad (16)$$

where C is a set of all context words of w, and c_j is the jth context word of w.

In here, the occurrence probability of w (denoted as $p(w)$) and the occurrence probability of w with its context c_j (denoted as $p(c_j, w)$) are defined as:

$$p(w) \equiv \sum_{c_j \in Context(w)} p(c_j, w) \quad (17)$$

$$p(c_j, w) \equiv \frac{\#(c_j, w)}{\sum_w \sum_{c_k \in Context(w)} \#(c_k, w)} \quad (18)$$

The output value of this entropy function conditions on two main points. First, as we defined in Equ. 16, if there's a high frequency word w, the output value will be high. Second, for a word w with many different contextual words, the output value will be higher. This score function is based on our assumption that if a word has context with higher diversity, its embedding quality is supposed to be better and does not need to be adjusted much.

3 Experiments

We conduct two experiments to evaluate our approach: Word Similarity and Analogical Reasoning. These two experiments directly test the quality of information embedded in the word vector. We integrate semantic information from knowledge bases using the four strategies: Baseline(Joint), Threshold, Function(Freq.), and Function(Ent.). We compare our proposed methods under the setting of using both prior knowledge and context to adjust the embeddings.

3.1 Experiment Setup

3.1.1 Training Data

We use New York Times (NYT) 1994-97 subset from Gigaword v5.0 (Parker et al., 2011) as the training corpus for CBOW, which is the same setting as (Yu and Dredze, 2014). After preprocessing of tokenization, the final training corpus contains 555.4 million tokens. We use two knowledge bases: Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) and WordNet (Miller, 1995). For PPDB, we use the XXL package,

Resource	Method	MEN-3k	RW	WS353	WS353r	WS353s	Average
	CBOW	63.6	33.9	57.7	46.7	68.5	54.1
PPDB	Baseline (Joint)	66.3	36.8	59.6	48.7	70.4	56.4
	Threshold	66.0	35.5	60.2	50.2	71.0	56.6
	Function (Freq.)	**68.7**	37.4	**61.1**	**51.8**	**71.3**	**58.1**
	Function (Ent.)	68.6	**37.8**	60.5	50.3	71.1	57.7
WordNet	Baseline (Joint)	66.4	**35.7**	59.6	**49.9**	69.7	56.3
	Threshold	66.3	35.5	**59.8**	49.4	**71.2**	**56.4**
	Function (Freq.)	66.3	35.4	58.9	49.5	68.9	55.8
	Function (Ent.)	**66.6**	35.2	58.2	48.3	68.1	55.3

Table 1: Spearman rank correlation on word similarity task. All embeddings are 300 dimensions. The best result for each dataset is highlighted in bold.

which shows the best result in (Yu and Dredze, 2014). It contains 587,439 synonym word pairs. For WordNet, we extract relation pairs from synonym. It contains 132,046 word pairs.

3.1.2 Parameter Setting

We set all our embedding size to 300, which is a suitable embedding size mentioned in (Melamud et al., 2016). The training iteration for RCM is 100. Learning rate for CBOW is 0.025. We experiment on an array of learning rates for the Baseline(Joint) and the best one is 0.0001. While the learning rate for Threshold remains 0.0001, we attempt various learning rates for Function(Freq.) and Function(Ent.) and the best one is 0.001, which is larger than 0.0001. This setting can be actually explained by that the output values of the two functions are between 0 to 1, which is used to decrease the learning rate. In other words, the learning rate of the two functions needs to be set a larger value than the baseline in order to be decreased by the two functions.

The Window Size is 5. Negative Sample is 15. We experiment on the threshold values of 10, 50 and 100. In our experiments, 50 gets the best result. We first learn the embeddings using CBOW with a random initialization, and take this pre-trained embeddings to initialize a joint model, where CBOW and RCM are jointly trained, and their learning rates are adjusted by using our proposed functions. Following (Yu and Dredze, 2014), we use asynchronous stochastic gradient ascent in training, where the threads to the CBOW and RCM are set to be a balance of 12:1 and the shared embeddings are updated by each thread based on training data within the thread. We let the CBOW threads to control convergence; training stops when CBOW threads finish processing the data. The joint model without using our proposed functions is taken as the baseline system, denoted by Baseline(Joint)

3.2 Word Similarity Task

The aim of word similarity task is to check whether a given word would have the similarity score which closely corresponds to human judges. These datasets contain relatedness scores for pairs of words; the cosine similarity of the embedding for two words should have high correlation. We use five datasets to evaluate: **MEN-3k** (Bruni et al., 2014), **RW** (Luong et al., 2013), **WordSim-353** (Finkelstein et al., 2002), also the partitioned dataset from WordSim-353, separated into the dataset into two different relations, **WS353-Similarity** and **WS353-Relatedness** (Agirre et al., 2009; Zesch et al., 2008).

Table 1 shows that comparing to the baseline, all of our proposed three methods get significant improvement. The results support our argument that incorporating prior knowledge into context-based embeddings can complement the embedding quality of those words which lack enough statistics of word occurrences.

Resource	Method	Google	MSR	Avg.
	CBOW	43.0	52.0	47.5
PPDB	Baseline (Joint)	46.8	54.9	50.9
	Threshold	46.2	54.2	50.2
	Function (Freq.)	**46.8**	**55.6**	**51.2**
	Function (Ent.)	46.8	55.0	50.9
WordNet	Baseline (Joint)	45.9	53.9	49.9
	Threshold	46.3	53.9	50.1
	Function (Freq.)	45.8	53.7	49.8
	Function (Ent.)	**46.6**	**53.9**	**50.2**

Table 2: Accuracy on analogical reasoning task. All embeddings are 300 dimensions. The best result for each dataset is highlighted in bold.

3.3 Analogical Reasoning Task

Analogical reasoning task was popularized by (Mikolov et al., 2013b). The dataset is composed

Resource	Method	MEN-3k	RW	WS353	WS353r	WS353s	Average
	CBOW	14.4	9.1	27.7	16.8	37.3	21.1
PPDB	Baseline (Joint)	21.4	9.7	33.6	22.5	41.6	25.8
	Threshold	**22.7**	9.7	34.1	22.2	**42.5**	26.2
	Function (Freq.)	22.4	**9.8**	33.9	22.7	41.3	26.0
	Function (Ent.)	22.2	9.7	**34.6**	**23.7**	41.9	**26.4**
WordNet	Baseline (Joint)	21.4	9.7	33.2	22.5	40.5	25.5
	Threshold	22.1	10.0	34.4	**23.4**	41.5	26.3
	Function (Freq.)	**22.2**	**10.1**	**34.6**	23.3	**42.5**	**26.5**
	Function (Ent.)	22.2	9.8	33.2	21.9	41.4	25.7

Table 3: Spearman rank correlation on word similarity task. All embeddings are 300 dimensions. The corpus is the same as Table 1, but the size is 1/100. The best result for each dataset is highlighted in bold.

of analogous word pairs. It contains pairs of tuples of word relations that follow a common syntactic relation. The goal of this task is to find a term c for a given term d so that c:d best resembles a sample relationship a:b. We use the vector offset method (Levy and Goldberg, 2014; Mikolov et al., 2013b), computing $e_d = e_a - e_b + e_c$ and returning the vector which has the highest cosine similarity to e_d. We use two datasets, **Googles analogy dataset** (Mikolov et al., 2013b), which contains 19,544 questions, about half of the questions are syntactic analogies and another half of a more semantic nature, and **MSR analogy dataset** (Mikolov et al., 2013b), which contains 8,000 syntactic analogy questions.

Table 2 shows the similar result as Word Similarity and demonstrates our proposed methods are stable and can be applied to different tasks.

3.4 Corpus Size

We also apply our models on the corpus with a smaller size. The same corpus is used but its size is 1/100. All parameters are the same except that the threads to the CBOW and RCM are set to be a balance of 2:1, and only the learning rates of positive samples are adjusted by our functions. The results are shown in Table 3 and Table 4, which shows our proposed models also improve the CBOW and outperform the baseline. In our experiments, we find out that for a smaller corpus, adjusting the learning rates of both positive samples and negative samples can not gain as much improvement as only using positive samples. Our conjecture is that since the quality of the embeddings trained from a smaller corpus might not be as high as the ones trained from a larger corpus, and the number of negative samples is much more than the positive sample (15:1 in our setting) each time, negative sample with the learning rate adjustment are more likely to mislead the training for a smaller corpus.

Resource	Method	Google	MSR	Avg.
	CBOW	3.5	7.5	5.5
PPDB	Baseline (Joint)	4.7	8.9	6.8
	Threshold	4.7	**9.5**	**7.1**
	Function (Freq.)	**4.8**	9.2	7.0
	Function (Ent.)	4.7	9.1	6.9
WordNet	Baseline (Joint)	4.5	8.8	6.7
	Threshold	4.6	8.9	6.8
	Function (Freq.)	**4.8**	**9.3**	**7.1**
	Function (Ent.)	4.8	9.2	7.0

Table 4: Accuracy on analogical reasoning task. All embeddings are 300 dimensions. The corpus is the same as Table 2, but the size is 1/100. The best result for each dataset is highlighted in bold.

4 Conclusion

In this paper, we argue that while applying prior knowledge into context-based embeddings, statistics of word occurrences should be considered, which based on the assumption that a embedding with more contextual information is supposed to have higher quality, and thus should be treated in a different way while incorporating with knowledge bases. We propose three models and demonstrate our embeddings got improved on two different tasks: Word Similarity and Analogical Reasoning. The implementation is based on RCM package and we have released the code for academic use. [1] In the future, under this framework, we plan to further investigate other possible score functions of learning rate based on information theory or dynamic consideration of training process for the incorporation of context and knowledge base information.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive suggestions to improve the quality of the paper.

[1] https://github.com/hywangntut/KBE

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 *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics - Human Language Technologies*, pages 19–27, Boulder, Colorado.

Yoshua Bengio. 2013. Deep learning of representations: Looking forward. In *Proceedings of the Statistical Language and Speech Processing*, pages 1–37, Tarragona, Spain.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases*, pages 132–148, Nancy, France.

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

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Kai-Wei Chang, Wen-tau Yih, and Christopher Meek. 2013. Multi-relational latent semantic analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1602–1612, Seattle, USA.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the International Conference on Machine Learning*, pages 160–167, Helsinki, Finland.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Paramveer S. Dhillon, Jordan Rodu, Dean P. Foster, and Lyle H. Ungar. 2012. Two step CCA: a new spectral method for estimating vector models of words. In *Proceedings of the International Conference on Machine Learning*, pages 1551–1558, Edinburgh, Scotland.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics - Human Language Technologies*, pages 1606–1615, Denver, Colorado.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20:116–131.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: the paraphrase database. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics - Human Language Technologies*, pages 758–764, Atlanta, USA.

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 Association for Computational Linguistics*, pages 873–882, Jeju, Korea.

Rémi Lebret and Ronan Collobert. 2014. Word embeddings through hellinger PCA. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–490, Gothenburg, Sweden.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the Association for Computational Linguistics*, pages 302–308, Baltimore, USA.

Ping Li and Cun-Hui Zhang. 2015. Compressed sensing with very sparse gaussian random projections. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 617–625, San Diego, USA.

Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics - Human Language Technologies*, pages 1030–1040, San Diego, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 3111–3119, Long Beach, USA.

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 of Computational Linguistics - Human Language Technologies*, pages 746–751, Atlanta, USA.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Andriy Mnih and Geoffrey E. Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the International Conference of Machine Learning*, pages 641–648, Corvallis, Oregon.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition.

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

Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut. 2006. An improved model of semantic similarity based on lexical co-occurence. *Communications of the ACM*, 8:627–633.

Gerard Salton and Michael McGill. 1984. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013a. Parsing with compositional vector grammars. In *Proceedings of the Association for Computational Linguistics*, pages 455–465, Sofia, Bulgaria.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, USA.

Stefanie Tellex, Boris Katz, Jimmy J. Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 41–47, Toronto, Canada.

Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: a general framework for incorporating knowledge into word representations. In *Proceedings of the International Conference on Conference on Information and Knowledge Management*, pages 1219–1228, Shanghai, China.

Wen-tau Yih, Geoffrey Zweig, and John C. Platt. 2012. Polarity inducing latent semantic analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1212–1222, Jeju, Korea.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the Association for Computational Linguistics*, pages 545–550, Baltimore, USA.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. In *Proceedings of the Conference on Artificial Intelligence*, pages 861–866, Chicago, Illinois.