Joint Learning of Sense and Word Embeddings

Mohammed Alsuhaibani Danushka Bollegala
Department of Computer Science, University of Liverpool.
{m.a.alsuhaibani, danushka.bollegala}@liverpool.ac.uk

Abstract

Methods for learning lower-dimensional representations (embeddings) of words using unlabelled data have received a renewed interested due to their myriad success in various Natural Language Processing (NLP) tasks. However, despite their success, a common deficiency associated with most word embedding learning methods is that they learn a single representation for a word, ignoring the different senses of that word (polysemy). To address the polysemy problem, we propose a method that jointly learns sense-aware word embeddings using both unlabelled and sense-tagged text corpora. In particular, our proposed method can learn both word and sense embeddings by efficiently exploiting both types of resources. Our quantitative and qualitative experimental results using unlabelled text corpus with (a) manually annotated word senses, and (b) pseudo annotated senses demonstrate that the proposed method can correctly learn the multiple senses of an ambiguous word. Moreover, the word embeddings learnt by our proposed method outperform several previously proposed competitive word embedding learning methods on word similarity benchmark datasets.

Keywords: Sense Embeddings, Word embeddings, Labelled Data, Unlabelled Data

1. Introduction

The ability to accurately represent the meanings of words is a fundamental requirement for many natural language processing (NLP) tasks. By using accurate word representations, it is possible to improve the performance of downstream NLP applications such as name entity recognition (NER) (Turian et al., 2010), word similarity measurement (Huang et al., 2012), sentiment analysis (Dhillon et al., 2015), word analogy detection (Bollegala et al., 2014), syntactic parsing (Socher et al., 2013) and dependency parsing (Bansal et al., 2014). Moreover, compositional approaches can be used to compute phrase-, sentence- or document-level embeddings from word embeddings (Baroni et al., 2014). Consequently, various methods have been proposed recently that *embed* words in lower-dimensional dense vector spaces, for example, using word co-occurrence information such as skip-gram with negative sampling (SGNS), continuous bag-of-words model (CBOW) (Mikolov et al., 2013) and Global Vectors (GloVe) (Pennington et al., 2014), to name a few.

A common limitation associated with existing prediction-based word embedding learning methods is that they represent each word by a *single* vector, ignoring the possibly multiple senses of a word. For example, consider the ambiguous word *bank* that could mean either a *financial institution* or a *river-bank*. The two senses of *bank* are significantly different, and embedding both senses to the same point is inadequate.

Several solutions have been proposed in the literature to overcome this limitation and learn *sense embeddings*, which capture the sense related information of words. For example, Reisinger and Mooney (2010) proposed a method for learning sense-specific high dimensional distributional vector representations of words, which was later extended by Huang et al. (2012) using global and local context to learn multiple sense embeddings for an ambiguous word. Neelakantan et al. (2014) proposed a multi sense skip-gram (MSSG), an online cluster-based sense-specific word representations learning method, by extending SGNG. Unlike SGNG, which updates the gradient of the word vector according to the context, MSSG predicts the nearest sense first, and then updates the gradient of the sense vector.

Aforementioned methods apply a form of word sense discrimination by clustering a word contexts, before learning sense-specific word embeddings based on the induced clusters to learn a fixed number of sense embeddings for each word. In contrast, a nonparametric version of MSSG (NP-MSSG) (Neelakantan et al., 2014) estimates the number of senses per word and learn the corresponding sense embeddings. On the other hand, Iacobacci et al. (2015) used a Word Sense Disambiguation (WSD) tool to sense annotate a large text corpus and then used an existing prediction-based word embeddings learning method to learn sense and word embeddings with the help of sense information obtained from the BabelNet (Navigli and Ponzetto, 2010) sense inventory. Similarly, Camacho-Collados et al. (2015) used the knowledge in two different lexical resources: WordNet (Miller, 1995) and Wikipedia. They use the contextual information of a particular concept from Wikipedia and WordNet synsets prior to learning two separate vector representations for each concept.

Above-mentioned methods for learning word and sense embeddings require either (a) sense inventories (dictionaries defining the different senses of a word), and (b) word sense taggers that can be applied on unlabelled corpora to generate sense-labelled training data, or (c) manually sense-annotated corpora. Unfortunately, such resources are either under developed or not available for most resource poor languages. On the other hand, methods that learn only word embeddings such as SGNS, CBOW, GloVe etc. can operate on unlabelled corpora. It remains unclear whether unlabelled data can help the process of learning sense embeddings, thereby reducing the manual effort required for creating sense tagged corpora for learning sense embeddings. Revisiting our previous example, only few instances of the word *bank* might be annotated in the labelled data with its sense as a financial institute, however, there might be many other words such as *cash*, *ATM*, *transaction* etc. that co-

occur with bank that could contribute information about this particular sense towards the embedding of *bank*. Importantly, such word-level co-occurrences can be obtained purely using unlabelled texts, which are comparatively easier to obtain than sense-labelled texts.

In this paper, we propose a method that uses a large collection of unlabelled texts and a comparatively smaller collection of sense-labelled sentences to learn both word and sense embeddings simultaneously. Our proposed method randomly initialises each word w_i and each of its senses s_{ij} with unique embedding vectors, and update those vectors such that the rank loss between words and senses that co-occur in unlabelled or labelled contexts is minimised over the entire vocabulary of words. In particular, we do not require sense lexicons or dictionary definitions (glosses) of words/senses in this process. Moreover, the proposed method works in an *online* fashion, where we require only a single pass over the data considering one sentence at a time. This is particularly attractive when learning from large collections of unlabelled texts, such as the ukWaC corpus (Baroni et al., 2009) used in our experiments.

We conduct two sets of experiments to evaluate the word/sense embeddings learnt by the proposed method. First, (in § 3.1.) we create a *pseudo sense-labelled* corpus by replacing two words by a unique identifier to create an artificially sense tagged corpus. This approach enables us to generate arbitrarily large sense-labelled data considering different frequency levels of the ambiguous words. Our experimental results on this dataset show that the proposed method can indeed learn word embeddings that are sensitive to the different senses appearing in the dataset. Second, (in § 3.2.) we use the learnt word embeddings to compute the semantic similarity between two words for word-pairs that have been rated by humans. This experiment reveals that by incorporating unlabelled data, we can indeed learn better word embeddings that are sensitive to the word senses compared to what we would get if we had used only labelled data, which is encouraging given the abundance of unlabelled text corpora. Moreover, the experiment shows that by considering the senses in the learning process we can not only learn better sense embeddings, but it also improves the accuracy of the word embeddings as well.

2. Learning Sense Aware Word Embeddings

We propose a method to jointly embed words and their senses in the same lower-dimensional dense vector space. To explain our method, let us consider the lemma of the *target word* $l_i \in \mathcal{V}$ for which we are interested in learning a word embedding $\boldsymbol{l}_i \in \mathbb{R}^d$ in some d-dimensional real space. Here, $\mathcal{V}$ is the vocabulary of words and we use bold fonts to denote word/sense embedding vectors. Given an unlabelled (i.e. not sense-tagged) corpus $\mathcal{U}$, let us denote the set of contexts in which l_i occurs by $\mathcal{K}_i$. Here, for example, a context can be a window of fixed/dynamic length, a sentence or a document. Next, let us consider the lemma of a *context word* l_n that co-occurs with l_n, denoted by $l_n \in \mathcal{K}_i$. Inspired by the negative sampling method used in SGNS, we would like to learn the embeddings of l_i and l_n close to each other than a word $l_m (\notin \mathcal{K}_i)$ that does not co-occur with l_i. We sample $l_n \sim P_u$ from the unigram distribution P_u such that words that are frequent in the corpus (therefore likely to occur in a given sentence) but do not co-occur with l_i as the *negative* examples. We define the hinge loss J_{ww} for predicting l_n over l_m in all contexts $\mathcal{K}(l_i)$ over the entire vocabulary by

$$J_{ww} = \sum_{l_i \in \mathcal{V}} \sum_{l_n \in \mathcal{K}_i} \sum_{\substack{l_m \sim P_u \\ l_m \notin \mathcal{K}_i}} \max\left(-\boldsymbol{l_i}^\top \boldsymbol{l_n} + \boldsymbol{l_i}^\top \boldsymbol{l_m} + 1, 0\right) \quad (1)$$

J_{ww} can be computed using unlabelled data and does not involve sense embeddings.

We require that the word embeddings must be able to predict not only the co-occurrences of a context word in contexts where a target word occurs, but also must be able to predict the senses associated with the target and contexts words. To model such word vs. sense co-occurrences, given a sense-tagged corpus $\mathcal{L}$, we compute the hinge loss J_{ws} associated with predicting the correct sense s_{nt} of the context word l_n and a randomly sampled sense s_{mg} from the distribution of senses in unigrams P_s that does not occur with l_i as follows:

$$J_{ws} = \sum_{l_i \in \mathcal{V}} \sum_{s_{nt} \in \mathcal{K}_i} \sum_{\substack{s_{mg} \sim P_s \\ s_{mg} \notin \mathcal{K}_i}} \max\left(-\boldsymbol{l_i}^\top \boldsymbol{s}_{nt} + \boldsymbol{l_i}^\top \boldsymbol{s}_{mg} + 1, 0\right) \quad (2)$$

Here, P_s is computed by counting the occurrences of senses in $\mathcal{L}$.

Likewise, we can compute the hinge loss J_{sw} for predicting a context word using the sense s_{ij} of the target word over a randomly sampled word $l_m \sim P_u$ as follows:

$$J_{sw} = \sum_{l_i \in \mathcal{V}} \sum_{l_n \in \mathcal{K}_i} \sum_{\substack{l_m \sim P_u \\ l_m \notin \mathcal{K}_i}} \max\left(-\boldsymbol{s}_{ij}^\top \boldsymbol{l_n} + \boldsymbol{s}_{ij}^\top \boldsymbol{l}_m + 1, 0\right) \quad (3)$$

Finally, we require that sense embeddings must be able to predict the correct sense s_{nt} of a context word given the sense s_{ij} of the target word. This requirement is captured by the hinge loss given by (4), where the inner-product between $\boldsymbol{s}_{ij}$ and $\boldsymbol{s}_{nt}$ must be greater than with $\boldsymbol{s}_{mg}$, a randomly sampled sense $s_{mg} \sim P_s$, as given by (4).

$$J_{ss} = \sum_{l_i \in \mathcal{V}} \sum_{s_{nt} \in \mathcal{K}_i} \sum_{\substack{s_{mg} \sim P_s \\ s_{mg} \notin \mathcal{K}_i}} \max\left(-\boldsymbol{s}_{ij}^\top \boldsymbol{s}_{nt} + \boldsymbol{s}_{ij}^\top \boldsymbol{s}_{mg} + 1, 0\right) \quad (4)$$

We combine the four losses given above into a single linearly-weighted objective given by (5), for some $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$ coefficients.

$$J = J_{ww} + \lambda_1 J_{ws} + \lambda_2 J_{sw} + \lambda_3 J_{ss} \quad (5)$$

We find the word embeddings $\boldsymbol{l}_i$, $\boldsymbol{l}_m$, $\boldsymbol{l}_n$ and sense embeddings $\boldsymbol{s}_{ij}$, $\boldsymbol{s}_{nt}$, $\boldsymbol{s}_{mg}$, $\boldsymbol{s}_{ft}$ such that J is minimised. For this purpose, we compute the partial derivatives of J w.r.t. word and sense embeddings and use stochastic gradient descent (SGD) with initial learning rate set to 0.01.

3. Experiments and Results

We conduct two sets of experiments to evaluate the embeddings learnt by the proposed method. First, in § 3.1., we qualitatively evaluate the ability of the proposed method to

Words	Unique Identifier (ambigious word)	Nearest Neighbours unlabelled corpus	Nearest Neighbours joint (labelled+unlabelled) corpora sense#1	sense#2
career (0.8) africa (0.2)	careeryafrica	south, australia, development,education, professional, developing, china,west, seeking, experience, job, academic	careers, professional, profession, graduate, academic, employment, training, development, job, successful, skills, pursue	india, europe, asia, south, kenya, australia, china, african, southern, countries, pacific, brazil
stock (0.7) dance (0.3)	dancystock	market, music, shares, exchange, company, rolling, markets, art, mix, stocks, theatre, dancing	stocks, market, markets, price, exchange,purchase, prices, investment, company, shares, trading, products	dancing, music, musical, jazz, theatre, art, singing, ballet, drama, artists, opera, song
sea (0.6) chapter (0.4)	chapterysea	river, ocean, introduction, atlantic, island, coastal, section, shore, coast, above, waters,north	ocean, river, coast, mountains, bay, atlantic, shore, beach, coastal, island, sand, water	introduction, section, chapters, book, summary, article, describes,act, review, notes, paragraph,report
dog (0.5) chairman (0.5)	dogychairman	executive, cat, chief, president, bob, david, director,john, horse, cats, brown,fox	cat, puppy, pet, horse, cats,dogs, rat, girl, breed, sheep, horses, boy	executive, chief, committee, director, treasurer, secretary, john, vice, officer, turner, superintendent, deputy

Table 1: Nearest Neighbours of the learnt sense and word embeddings.

discover known senses in a pseudo-labelled dataset. Second, in § 3.2., we compare the word embeddings learnt by the proposed method against prior work on multiple word similarity benchmarks.

3.1. Qualitative Analysis

To verify that the proposed method can learn sense embeddings for the the different senses of an ambiguous word as expected, we conduct the following experiment. We create a pseudo sense-tagged corpus by replacing all occurrences of two words by an artificial word in a corpus and tag the mentions of original words as different senses of the artificial word. Due to space limitations, few examples are shown in Table 1, where we select words with different frequencies (ratio of frequencies indicated within brackets in the first column). For example, we replace *dog* and *chairman* with the artificial ambiguous word *dogychairman* with two senses corresponding dog and chairman. Using ukWaC as the unlabelled corpus, we produced a pseudo-labelled corpus following this procedure. This approach enables us to create arbitrarily large sense-tagged corpora with known senses (and frequencies), which is useful for verifying that the proposed method is working as expected.

We run the proposed method independently on the (a) unlabelled corpus, and (b) the combination of unlabelled and pseudo-labelled corpora to compute word (in the case of both (a) and (b)) and sense ((b) only) embeddings. The nearest neighbouring words (computed using the cosine similarity between the learnt 300 dimensional embeddings) for setting (a) (third column) and for setting (b) (fourth and fifth columns) are shown in Table 1. From Table 1 we see that the nearest neighbours of the word embeddings learnt using only the unlabelled corpus are a mixture of the multiple senses of the ambiguous artificial word. On the other hand, the sense embeddings learnt by the proposed method using both unlabelled and labelled data enable us to produce coherent neighbourhoods, capturing a single sense of the artificial ambiguous word.

To further illustrate the ability of the proposed method for learning the sense and word embeddings, we use t-SNE (Maaten and Hinton, 2008) to project the word embeddings to two-dimensional space as shown in Figure 1. Nearest neighbours of *dogychairman* and its two senses are highlighted. We see that the proposed method successfully learns the different senses of the ambiguous word in the embedding space. For example, the *dog* sense of *doggychaieman* has neighbours such as *dogs, cats* and *pet*, whereas the *chairman* sense has *executive, president* and *director* as the neighbours.

3.2. Word Similarity

To empirically compare the proposed method against prior work, we use ukWaC as the unlabelled corpus and SemCor (Miller et al., 1993) as the sense-tagged corpus, and learn word and sense embeddings using the proposed method. We set the context window to 10 tokens to the right and left of a word in the sentence. We used 5 negative samples for both words $\boldsymbol{l}_m$ and senses $\boldsymbol{s}_{mg}$ with 0.75 as a uniform sampling rate. The proposed model converged to a solution with 20 training epochs. We used the Rubenstein-

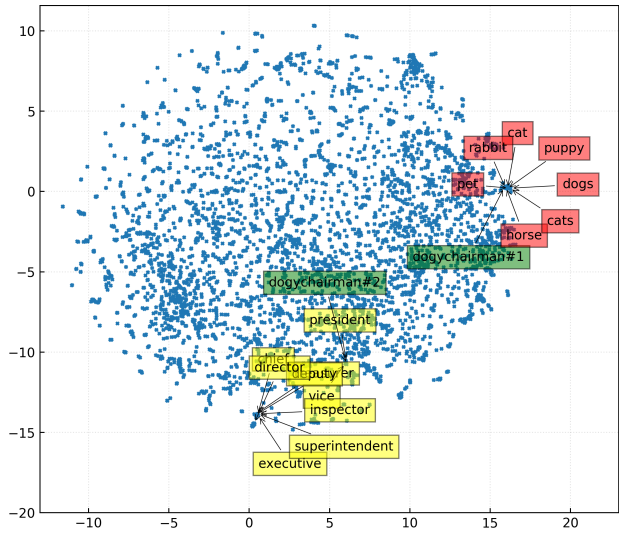

Figure 1: t-SNE projection of word/sense embeddings. Green labels show the two sense embeddings for *doggy-chairman*, whereas yellow and red labels show the nearest neighbours for the two senses. Best viewed in colour.

Model	WS	MC	RW	SCWS	MEN	SimLex
CBOW	0.587	0.569	0.251	0.523	0.654	0.291
SGNS	0.633	**0.746**	0.259	0.582	0.677	0.356
GloVe	0.465	0.664	0.265	0.483	0.701	0.327
MSSG	0.658	0.738	0.152	0.632	0.676	0.341
NP-MSSG	0.653	0.715	0.153	**0.639**	0.674	0.355
Proposed	**0.668**	0.702	**0.282**	0.606	**0.734**	**0.372**

Table 2: Performance of the proposed method in comparison with prior work evaluated on word similarity benchmark datasets.

Goodenough (**RG**, 65 word-pairs) (Rubenstein and Goodenough, 1965) dataset as a validation dataset to tune the hyperparameters λ_1, λ_2 and λ_3 defined in (5). In particular, we vary the values of the coefficients λ_1, λ_2 and λ_3 and learn the sense and word embeddings using the proposed method afore measuring the Spearman correlation on **RG** dataset. Next, λ_1, λ_2 and λ_3 values are selected based on the highest reported correlation score.[1]

Next, we measure the cosine similarity between two words in human similarity benchmarks using their embeddings, and measure Spearman correlation coefficient between human similarity ratings and computed cosine similarities. A higher correlation with human similarity ratings implies that the word embeddings learnt by the proposed method accurately capture the semantics of the words.

We use several benchmark datasets in our evaluations: WordSim353 (**WS**, 353 word-pairs) (Finkelstein et al., 2002), Miller-Charles (**MC**, 30 word-pairs) (Miller and Charles, 1998), rare words dataset (**RW**, 2034 word-pairs) (Luong et al., 2013), Stanford's contextual word similarities (**SCWS**, 2023 word-pairs) (Huang et al., 2012), **MEN** test collection (3000 word-pairs) (Bruni et al., 2012) and the SimLex-999 (**SimLex**, 999 word-pairs) (Hill et al., 2016).

[1]Setting $\lambda_1 = \lambda_2 = \lambda_3 = 10$ performed consistently well in our experiments.

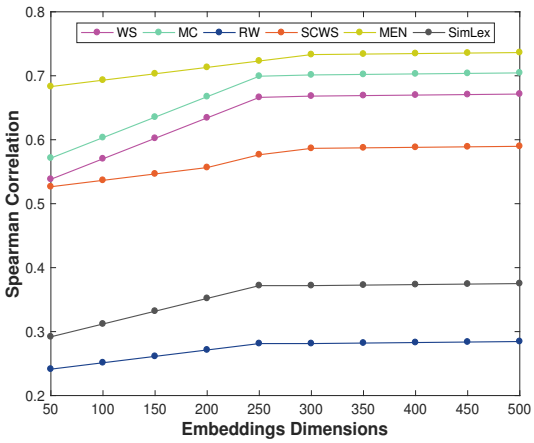

Figure 2: Accuracy vs Dimensionality of the word embeddings evaluated on the **WS**, **MC**, **RW**, **SCWS**, **MEN** and **SimLex** datasets.

In Table 2, we compare several word embedding learning methods such as sense-insensitive embeddings **CBOW**, **SGNG** (Mikolov et al., 2013) and **GloVe** (Pennington et al., 2014), and sense-sensitive embeddings **MSSG** and **NP-MSSG** (Neelakantan et al., 2014) for learning sense embeddings. We limit the comparison to the state-of-the-art methods for which source codes are publicly available such that we can train all methods on the same datasets and same dimensionality (i.e 300) for a fair comparison.

From Table 2, we see that the proposed method reports the best perfomance in most benchmark datasets, except for the smallest dataset **MC** and the contextual dataset **SCWS**. Table 2 shows that using a sense-tagged corpus is not only beneficial for learning sense embeddings, but also helps in learning better word embeddings. For example, the proposed method report the highest score among all other models in two of the largest word similarity datasets **MEN** and **SimLex**. **NP-MSSG** reports the best performance in **SCWS** where sentential information is available, which shows an advantage of cluster-based models of capturing the senses. However, the proposed method significantly outperforms (Fisher transformation at $p < 0.05$) **NP-MSSG** and **MSSG** in **RW**, **MEN** and **SimLex**.

Figure 2 shows the effect of the dimensionality of the embeddings learnt by the proposed method. Overall, in all benchmarks, the proposed method is able to learn accurate word embeddings with as small as 50 dimensions. Moreover, the performance gradually increase with the dimensionality reaching a peak around 300 dimensions.

4. Conclusion

We proposed a method for jointly learning word and sense embeddings using both an unlabelled corpus and a sense-tagged corpus. Our experiments on multiple similarity benchmarks show that the proposed method learns accurate word embeddings by modelling senses.

5. Bibliographical References

Bansal, M., Gimpel, K., and Livescu, K. (2014). Tailoring continouos word representations for dependency parsing. In *Proc. of ACL*.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014). Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9:5–110.

Bollegala, D., Maehara, T., Yoshida, Y., and ichi Kawarabayashi, K. (2014). Learning word representations from relational graphs. In *Proc. of AAAI*, pages 2146 – 2152.

Bruni, E., Boleda, G., Baroni, M., and Tran, N. K. (2012). Distributional semantics in technicolor. In *Proc. of ACL*, pages 136–145.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015). Nasari: a novel approach to a semantically-aware representation of items. In *HLT-NAACL*, pages 567–577.

Dhillon, P. S., Foster, D. P., and Ungar, L. H. (2015). Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16:3035–3078.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., z. Solan, Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20:116–131.

Hill, F., Reichart, R., and Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*, pages 873 – 882.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proc. of ACL-IJCNLP*, pages 95–105.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *CoNLL*.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Mikolov, T., Chen, K., and Dean, J. (2013). Efficient estimation of word representation in vector space. In *Proc. of International Conference on Learning Representations*.

Miller, G. and Charles, W. (1998). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proc. of the Workshop on Human Language Technology*, pages 303–308.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proc. of ACL*, pages 216–225.

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proc. of EMNLP*, pages 1059–1069, October.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: global vectors for word representation. In *Proc. of EMNLP*, pages 1532 – 1543.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Proc. of NAACL*, pages 109–117.

Rubenstein, H. and Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8:627–633.

Socher, R., Bauer, J., Manning, C. D., and Andrew Y., N. (2013). Parsing with compositional vector grammars. In *Proc. of ACL*, pages 455–465.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384 – 394.