Enriching Word Sense Embeddings with Translational Context

Mehdi Ghanimifard
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
mmehdi.g@gmail.com

Richard Johansson
Språkbanken, Department of Swedish
University of Gothenburg
richard.johansson@gu.se

Abstract

Vector-space models derived from corpora are an effective way to learn a representation of word meaning directly from data, and these models have many uses in practical applications. A number of unsupervised approaches have been proposed to automatically learn representations of word *senses* directly from corpora, but since these methods use no information but the words themselves, they sometimes miss distinctions that could be possible to make if more information were available.

In this paper, we present a general framework that we call *context enrichment* that incorporates external information during the training of multi-sense vector-space models. Our approach is agnostic as to which external signal is used to enrich the context, but in this work we consider the use of *translations* as the source of enrichment. We evaluated the models trained using the translation-enriched context using several similarity benchmarks and a word analogy test set. In all our evaluations, the enriched model outperformed the purely word-based baseline soundly.

1 Introduction

Word meaning representations derived from corpora have recently seen much attention in natural language processing (NLP), most importantly because they can be used very effectively to abstract over the word level in lexicalized NLP systems (Miller et al., 2004; Koo et al., 2008; Turian et al., 2010; Bansal et al., 2014; Guo et al., 2014; Sienčnik, 2015). These representations are derived from corpus statistics, building on the *distributional hypothesis* that the meaning of a word is reflected in statistical distributions of the contexts in which it appears (Harris, 1954). This intuition can be implemented in a number of ways in practice; in this work, we focus on models that represent word meaning as a point in a metric space (Widdows, 2005; Sahlgren, 2006; Turney and Pantel, 2010; Clark, 2015). In particular, one member of this family that has been particularly influential recently is the *skip-gram* learning algorithm (Mikolov et al., 2013a), which is derived from the log-bilinear language model by Mnih and Hinton (2007). The main reasons for its popularity are its computational efficiency (Mikolov et al., 2013b), its high performance in several evaluations, and the availability of an implementation in the form of the easily usable `word2vec` package.

In most cases distributional word representations disregard the fact that many words have more than one possible interpretation, or *word sense*, and in lexicographical descriptions of a language we will typically list the senses of a word in different sub-entries (Cruse, 1986). For instance, the English word *bass* can refer to a fish, a musical instrument, the low part of a musical range, etc. It is imaginable that we could use standard techniques to learn a vector-space semantic representation from a sense-annotated corpus, but this is infeasible in practice since fairly large corpora are needed to induce data-driven representations of a high quality, while corpora with hand-annotated sense identifiers are small and scarce. Instead, there have been several attempts to use *unsupervised* methods that create vectors representing the senses of ambiguous words, most of them based on some variant of the idea that was first proposed by Schütze (1998): that the different senses of a word can be discovered by applying a clustering algorithm to the set of contexts where it has appeared. Variations on this idea have turned up in a number of recent papers (Huang et al., 2012; Moen et al., 2013; Neelakantan et al., 2014; Kågebäck et al., 2015). However, unsupervised

Proceedings of Recent Advances in Natural Language Processing, pages 208–215,
Hissar, Bulgaria, Sep 7–9 2015.

models for discovering word senses are solipsistic in the sense that they are not *grounded* in the external world in the way that a language user is. This leads to the problem that they sometimes tend to discover different *discourses* or domains, rather than true word senses (Tahmasebi, 2013). Because of this lack of external signals, it seems natural to try to introduce additional sources of information into the learning process.

In this paper, we enrich the multi-sense skip-gram model (Neelakantan et al., 2014) by introducing external signals, which are implemented as additional context features during training. In particular, we use a *parallel corpus*, where the foreign-language words work as a source of external information that helps the algorithm form more distinct clusters. For instance, the fish sense of *bass* can be clearly distinguished from the musical senses if we have access to a Swedish translation: the fish is called *havsabborre*, while most of the musical senses would be translated as *bas*. Our approach can be seen as a form of *distant supervision* (Mintz et al., 2009), in contrast to the fully unsupervised approaches mentioned above.

We evaluated the context-enriched model on a collection of word similarity benchmarks and analogy tests, including the *contextual word similarity set* used in previous work on learning representations of different senses (Huang et al., 2012), and we saw large improvements when comparing to a baseline without access to the enriched context.

2 Background: the Skip-gram Model and its Multi-sense Extension

In the *skip-gram* model (Mikolov et al., 2013a), a target word w and a context feature c are represented using vectors from two different vector spaces, denoted $v_t(w)$ and $v_c(c)$ respectively. Intuitively, we would like the training algorithm to fit the vectors so that $v_c(c) \cdot v_t(w)$ is a high number if we are likely to see c near w, and a low number otherwise.

In the original formulation of the model, these two vectors are combined into probability of the occurrence of a context feature c near a target word w using the following equation:

$$\log P(c|w) = v_c(c) \cdot v_t(w) - \log Z(c)$$

where $Z(c)$ is a normalization factor so that the probabilities sum to 1. In principle, the model could be fit to a training corpus by maximizing the likelihood of all the contexts in the corpus, but due to the normalization factors $Z(c)$ – which are computed by summing over the whole vocabulary – this is computationally inefficient, leading to a number of approximations. Mikolov et al. (2013a) used a hierarchical decomposition, but after a simplification of the the idea of *noise-contrastive estimation* (Mnih and Kavukcuoglu, 2013), the most recent `word2vec` implementation estimates the word vectors using an approach called *skip-gram with negative sampling* (SGNS) (Mikolov et al., 2013b). This model treats word–context pairs actually occurring in a corpus as positive training examples, and synthetic pairs that were generated randomly as negative examples, and then fits a logistic model that discriminates between positive and negative examples:

$$P(\text{true pair}|c, w) = \frac{1}{1 + e^{-v_c(c) \cdot v_t(w)}}$$

During training of the SGNS model, when we consider a true pair (w, c), we generate N synthetic pairs (w, c') with the same word but with the c' randomly selected from the context vocabulary. While SGNS is not guaranteed to converge to the same solution as the original skip-gram model, it is more efficient and has achieved comparable results in evaluations.

The *multi-sense skip-gram* model (MSSG) by Neelakantan et al. (2014) generalizes SGNS by taking multiple senses into account. This algorithm uses context vectors as in the original skip-gram model, but it replaces the target word vector $v_t(w)$ for a word w with K different *sense vectors* $v_s(w, k)$.[1] In addition, it uses K vectors $\mu(w, k)$ that represent the centers of the clusters of contexts. The learning algorithm works in a fashion similar to SGNS, but extends it by introducing an additional sense discrimination step. When the algorithm encounters a word w, it first represents the full context window by building a sum $\bar{v}_c$ of the context vectors of the words appearing in the window. It then selects sense k whose context cluster $\mu(w, k)$ maximizes the dot product with $\bar{v}_c$. Finally, it carries out a gradient update (similar to that in SGNS) of the sense vector $v_s(w, k)$ and the context vectors $v_c(c)$, and adds $\bar{v}_c$ to the context cluster $\mu(w, k)$.

[1] Neelakantan et al. (2014) also described a *nonparametric* variant where the number of senses was determined automatically. We did not use that model since the distributed code did not include that part.

3 Context Enrichment

One of the fundamental criticisms against distributional word learning claims that the disembodiment from physical world will cause problems due to the fact that many concepts are actually grounded in perception and a sample text from a language alone does not carry all information about the concept behind the word (Andrews et al., 2009).[2] The perceptual information which has been claimed to improve these models are usually multi-modal data, for instance images as visual context of word usage in a language. In this work, we will instead enrich the training context with another type of supplementary text – the translation of the English text into Swedish – in order to improve the final word sense discrimination model.

In our method, we use a parallel corpus such as Europarl (Koehn, 2005), which provides sentence-by-sentence translations. Then by aligning words in each sentence we will add corresponding list of words in enhancing language into the list of words in skip-gram context window. Figure 1 illustrates why we expect this to be useful for forming better word sense clusters. In the figure, the first occurrence of the word *plant*, meaning an industrial or power plant, is translated by the Swedish word *anläggning*; the second example means a botanical plant and is translated as *planta.* This shows clearly that the external context in the form of a translation can be useful for discriminating between senses: an industrial plant would never occur in Swedish as *planta*, or vice versa.

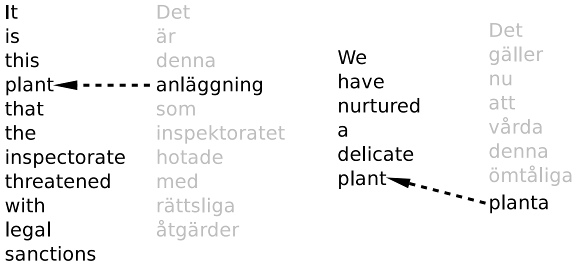

Figure 1: Examples of two occurrences in Europarl of the English word *plant* and their respective translations into Swedish.

3.1 Preprocessed Corpus

In order to facilitate and simplify the training process, we isolated the word alignment process from the rest of the training. In this isolated process in addition to the word alignment process which takes two parallel corpora and suggests one-to-many word alignments per sentence [3], we produce an enriched corpus by annotating the source corpus with words from the target corpus.

In order to get better results from word alignments, we applied a part-of-speech tagger on the Swedish and English words before running the aligner. Then we by taking the union of two word alignments with *fast_align* (Dyer et al., 2013) in both forward and reverse setups, we produced one-to-many mappings. We then read sentences from both corpora in parallel with their word mappings and generated the annotated corpus, which we refer to as the *enriched* or augmented corpus. The enriched corpus simply is the annotated source corpus which each word has its list of aligned words from target corpus.

During the training process, the Enriched Multi-Sense Skip-Gram Model will parse the annotated tokens, and add the enriched context to the skip-gram contexts as we describe in next section.

3.2 Enriched Multi-Sense Skip-Gram Model

The Enriched Multi-Sense Skip-Gram Model (EMSSG) extends the previous work by Neelakantan et al. (2014) by adding an extra step that incorporates external information into the context representation. In this procedure, sense vectors will be trained only for words in the source language; however, for any token occurring as context – including the translations – we produce a context vector. The enriched corpus is made of words and their enriched context (w, C). From each word from the source corpus $w_t \in W$ the corresponding enrichment is a subset of tokens from a parallel corpus $C_t \subset W'$:

$$W = \{w_t\}_{t \in 1,\ldots,T}, W' = \{w'_t\}_{t \in 1,\ldots,T'}$$

Basically, each token (w_t, C_t) is a result of word alignment which we produce in the preprocessing phase:

$$C_t = \{w'_{a_t(1)} \ldots w'_{a_t(m_t)}\}$$

[2] One can also relate this problem to the "symbol grounding problem", by saying that the result of a distributional learning algorithm will be just meaningless symbolic relations between words. But the symbol grounding problem is a problem for specific application of these models in cognitive modeling, which is also mentioned by Harnad (1990).

[3] In more complicated translation alignments, such as phrase-to-phrase mappings, we still can take the one-to-many implementation of these alignments in our one directional process.

In the training process, the enrichment context C_t will be added to the skip-gram context words $C_{sg} = \{w_{t-R_t}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+R_t}\}$ to create a combined context: $C = C_t \cup C_{sg}$. As in the original MSSG, the vector representation of the combined context will then be used to predict the right sense for the observed context. We first build a representation of the full context by summing all the individual context vectors:

$$\bar{v}_c = \sum_{w \in C} v_c(w)$$

This vector is then compared to all the context cluster centroids in order to predict the sense:

$$s_t = \operatorname*{argmax}_{k=1,2,\ldots,K} sim(\mu(w_t, k), \bar{v}_c)$$

Algorithm 1 shows the pseudocode of how we use the enriched context representation to improve the sense prediction and their corresponding clusters. The enriched context is only used during training as a form of distant supervision: at test time, only the skip-gram contexts are used when predicting the sense.

Algorithm 1 Training Algorithm of EMSSG model

```
input $(w_t, C_t)_{t \in \{1,2,\ldots,T\}}$, $d$, $K$, $N$.
for $t = 1, 2, \ldots, T$
for $k \in \{1, \ldots, K\}$
  initialize $\mu(w_t, k) = 0$
  randomly initialize $v_s(w_t, k)$, $v_c(w_t)$
for $t = 1, 2, \ldots, T'$
  randomly initialize $v_c(w'_t)$
for $t = 1, 2, \ldots, T$
    $R_t \sim \{1, \ldots, N\}$
    $C_{sg} \leftarrow \{w_{t-R_t}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+R_t}\}$
    $C \leftarrow C_t \cup C_{sg}$
    $\bar{v}_c \leftarrow \sum_{w \in C} v_c(w)$
    $s_t \leftarrow \mathrm{argmax}_{k=1,2,\ldots,K}\, sim(\mu(w_t, k), \bar{v}_c)$
    update cluster center:
      $\mu(w_t, s_t)$ with new context $C$
    for $c$ words in $C$
      gradient update: $v_s(w_t, s_t)$ with $v_c(c)$
      gradient update: $v_c(c)$ with $v_s(w_t, s_t)$
    $C' \leftarrow Noisy\_Samples(C)$
    for $c$ words in $C'$.
      negative gradient update:
        $v_s(w_t, s_t)$ with $v_c(c)$
return $v_s(w, k)$, $v_c(w)$, $v_c(w')$ ,$\mu(w, k)$
    for $w \in W, w' \in W', k \in 1, ..., K$
```

4 Experiments

To evaluate the enrichment model, we trained a baseline MSSG model without enrichment from English Europarl. Then by enriching the English Europarl with Swedish parallel corpus, as described in previous section, we trained the enriched model with the same setup.

In these models the dimension size is $d = 300$ and window size is $N = 5$, and number of senses is $k = 2$. To enable faster training we chose to train sense vectors only for top 1000 most frequent words, excluding stop words.

4.1 Word similarity tests

We evaluate our models with 3 different word similarity tests:

- the SimLex999 similarity test (Hill et al., 2014)
- the WordSim353 tests in both similarity and relatedness (Ponzetto and Strube, 2011)
- the Stanford Contextual Word Similarity test (Huang et al., 2012)

The evaluation procedures for word sense models in all of these test sets are identical:

- Disambiguate word senses for each pair of words.
- Quantify the similarity of pairs with the cosine similarity measure between two sense vectors.
- Calculate the correlation between gold standard and the estimated similarity.

In order to disambiguate the sense for a word, we need its context to find the most likely sense vector for that word. The sense disambiguation separate these tests in two groups: those with word contexts and those without word contexts.

4.1.1 Non-contextual tests

Both SimLex999 and WordSim353 are designed for evaluating word vector representations. Although the lack of context to describe the actual usage of word makes them unsuitable for word sense evaluation, they have been used to evaluate sense-aware vector-space models (Reisinger and Mooney, 2010; Neelakantan et al., 2014), so we include a comparison for completeness. However, despite the absence of context, human judges estimate their similarity based on their own understanding of senses of those words. Similar to *passive sense selection* in humans[4], we consider each word as context for the other word to select the best sense. With a twist, instead of using context vectors to predict the sense of the other one, we basically choose the most similar vectors pairs as desired vectors. This is equivalent to what Reisinger and Mooney (2010) term the *MaxSim* score.

To understand why we use this procedure, consider two very different words: in this case, we expect that all of their senses should be very different. Considering two words that the evaluators considered to be similar, it is likely that this does not apply to *all* of the senses, but only a specific pair. This motivates why we take the highest similarity of senses, and we think that this procedure is more meaningful than the *AvgSim* score used by (Reisinger and Mooney, 2010).

The English-Swedish Europarl's vocabulary covers 758 of word pairs in SimLex999 and 163 pairs in WordSim353 similarity test and 218 pairs WordSim353 relatedness test.

Table 1 shows the results of the evaluations on the three non-contextual benchmarks. As is customary in this type of evaluation, the similarity scores output by the model are compared to the gold standard using the Spearman correlation coefficient. In all three tests, the model with access to an enriched context representation clearly outperforms the baseline MSSG model.

Model	SL999	WS353-sim	WS353-rel
MSSG	0.29	0.44	0.35
EMSSG	0.36	0.52	0.39

Table 1: Spearman correlation values of the two systems when evaluated on the three non-contextual similarity test sets.

4.1.2 Contextual test

The Stanford Contextual Word Similarity test (Huang et al., 2012) consists of pairs of words and a sentence as an example for their usage. The sense disambiguation with the provided sample will be done by making a context vector as we have in MSSG models: the evaluation using this procedure is equivalent to the *localSim* procedure used by Neelakantan et al. (2014).

The English-Swedish Europarl's vocabulary covers 1498 samples of this dataset. In Table 2, we present the results (again, Spearman correlations) of the evaluation with this set. Again, the enriched model outperforms the baseline.

Model	Correlation
MSSG	0.45
EMSSG	0.53

Table 2: Evaluation on the Stanford contextual word similarity test set.

4.2 Word analogy test

The word analogy data set provided by Google (Mikolov et al., 2013c) is also another test for vector representations of words. The judgement on the word relation are based on their semantic or syntactic identity. For instance, an example of a semantic analogy is *Paris*:*France* = *Stockholm*:*Sweden*, while *sleeping*:*sleep* = *breaking*:*break* is an example of a syntactic analogy.

The test is about guessing the correct word vector by only having the three other word vectors. For instance, if the missing vector is $v_{gold} = v(``queen")$, the nearest neighbour word vector to the vector $v_{analogy} = v(``king") - v(``man") + v(``woman")$ should be v_{gold}. Similar to non-contextual word similarity tests, this test also needs a novel sense disambiguation method.

To find those word-senses that intended to be in each analogy test, we can suppose that correct senses in these tests should lead to only one correct answer. It means that the nearest neighbour to analogy vector $v_{analogy}$ should have a significant similarity comparing to other close neighbours of this vector. We can define a score to find the best analogy vector based on maximized margin from other neighbours. With k number of senses per word in the model, there are k^3 possible $v_{analogy}$.

For each possible $v_{analogy}$ and its top 10 closest sense vectors $V = \{v_1, ..., v_{10}\}$, we define the score of $v_{analogy}$ based on similarity of the nearest neighbour and its margin with other neighbours:

- δ_i is the *similarity margin* between $v_i \in V$

[4] Cruse (1986) used this term "*passive selection*" in contrast with "*productive selection*" as psycholinguistic matter, to describe sense selection among pre-established senses. Whenever we use this type of corpus driven word sense models, we only have passive selection because we only have pre-established senses. By using this term here, we want to emphasize that even in absence of context we can take most related senses as most obvious choice of sense

and the nearest neighbour v_1:

$$\delta_i = sim(v_1, v_{analogy}) - sim(v_i, v_{analogy})$$

- The score of $v_{analogy}$:

$$score = \frac{\sum_{i=1}^{10} \delta_i^2}{\delta_{10}^2} \times sim(v_1, v_{analogy})$$

Higher score in this formula indicates that v_1, the most similar vector to $v_{analogy}$, has a significant similarity to $v_{analogy}$ compering to other possible neighbour vectors. By taking the best $v_{analogy}$ from all possible $v_{analogy}$, we automatically pick 3 sense vectors for analogy test.

Table 3 shows the results of the evaluation on the Google analogy test set (Mikolov et al., 2013c). For the third time, the translation-enriched model outperforms the MSSG baseline in all tests.

Model	Total	Syntactic	Semantic
MSSG	0.13	0.04	0.17
EMSSG	0.25	0.09	0.32

Table 3: Evaluation on the Google analogy test set.

5 Related Work

The idea of integrating different modalities into corpus-based vector representations has generated much interest recently (Lazaridou et al., 2014; Socher et al., 2014). The work in this area that is most similar to ours is that by Hill and Korhonen (2014) and : they extend the context representation of the skip-gram model with features representing the external information like we do, although they do not take word senses into account.

Parallel corpora have been used in a number of research projects in order to derive *crosslingual* word representations; this is different from our goal, which is to use them to help the monolingual model form better sense clusters. Klementiev et al. (2012) presented a neural multi-task learning model that used bilingual cooccurrence data as a way to connect the models in two languages, and Utt and Padó (2014) described a syntactically informed context-counting method. Faruqui and Dyer (2014) presented a method that combine two monolingual vector spaces into a multilingual one by Canonical Correlation Analysis. In addition to vector-space models, bilingual and multilingual corpora have been used to derive a number of non-geometric corpus-based representations, such as Brown clusters (Täckström et al., 2012) and topic models (Vulić et al., 2015).

Finally, the use of word translations as a way to distantly supervise word sense disambiguation and discrimination systems is an idea that goes far back (Dagan et al., 1991; Dyvik, 2004) and has reappeared many times. This intuition was behind a number of SemEval cross-lingual word sense disambiguation and lexical substitution tasks (Lefever and Hoste, 2010; Mihalcea et al., 2010).

6 Conclusions

We have presented a general technique called *context enrichment* that allows us to use external information to multi-prototype vector-space models of word meaning. The intention of this approach is that the external signal helps the model form more coherent and well-separated clusters during the training process, and it is not necessary during testing. The approach that we have evaluated is a straightforward extension of the multi-sense skip-gram model by Neelakantan et al. (2014), but we imagine that other models (for instance Huang el al., 2012) could be extended in a similar fashion. The model can integrate any kind of language-external signal as long as it can be represented as a contextual feature taken from a finite vocabulary. In this work, we enriched the context using word translations taken from the Europarl corpus (Koehn, 2005).

We evaluated the multi-sense vector models trained with translation-enriched contexts using a number of different benchmarks: word similarity tests, a contextual similarity test, and a word analogy test. In every experiment we tried, the enriched model outperformed the non-enriched baseline.

It seems straightforward to extend our work to a setting where other types of features are used, and we would like to explore this area further. In particular, we would like to integrate multimodal input (Hill and Korhonen, 2014), for instance with information extracted from images. This could lead to several interesting experiments where the effect of different modalities on word sense discovery could be investigated.

Acknowledgements

We would like to thanks the reviewers for their constructive comments. The second author was funded by the Swedish Research Council under grant 2013–4944, *Distributional methods to represent the meaning of frames and constructions*.

References

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

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815, Baltimore, United States.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics, second edition*. Wiley-Blackwell.

D. A. Cruse. 1986. *Lexical semantics*. Cambridge University Press.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 130–137, Berkeley, United States.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *HLT-NAACL*, pages 644–648. Citeseer.

Helge Dyvik. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. *Language and computers*, 49(1):311–326.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120, Doha, Qatar.

Stevan Harnad. 1990. Symbol Grounding Problem: Turing-Scale Solution Needed. 42:335–346.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23).

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 255–265, Doha, Qatar.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Association for Computational Linguistics 2012 Conference (ACL 2012)*, Jeju Island, Korea.

Mikael Kågebäck, Fredrik Johansson, Richard Johansson, and Devdatt Dubhashi. 2015. Neural context embeddings for automatic discovery of word senses. In *Proceedings of the Workshop on Vector Space Modeling for NLP*, Denver, United States. To appear.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603, Columbus, USA.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, United States.

Els Lefever and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden.

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.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations, Workshop Track*, Scottsdale, USA.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, USA.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, SA.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore.

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.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 26*, pages 2265–2273.

Hans Moen, Erwin Marsi, and Björn Gambäck. 2013. Towards dynamic word sense discrimination with random indexing. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 83–90, Sofia, Bulgaria.

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 (EMNLP)*, pages 1059–1069, Doha, Qatar.

Simone Paolo Ponzetto and Michael Strube. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9-10):1737–1756.

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.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

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

Scharolta Katharina Sienčnik. 2015. Adapting *word2vec* to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 239–243, Vilnius, Lithuania.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada.

Nina N. Tahmasebi. 2013. *Models and Algorithms for Automatic Detection of Language Evolution*. Ph.D. thesis, Gottfried Wilhelm Leibniz Universität Hannover.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

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

Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147.

Dominic Widdows. 2005. *Geometry and Meaning*. CSLI Publications.