Learning Topic-Sensitive Word Representations

Fadaee, M.; Bisazza, A.; Monz, C.

DOI
10.18653/v1/P17-2070

Publication date
2017

Document Version
Final published version

Published in
The 55th Annual Meeting of the Association for Computational Linguistics

License
CC BY

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Learning Topic-Sensitive Word Representations

Marzieh Fadaee Arianna Bisazza Christof Monz
Informatics Institute, University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands
{m.fadaee,a.bisazza,c.monz}@uva.nl

Abstract
Distributed word representations are widely used for modeling words in NLP tasks. Most of the existing models generate one representation per word and do not consider different meanings of a word. We present two approaches to learn multiple topic-sensitive representations per word by using Hierarchical Dirichlet Process. We observe that by modeling topics and integrating topic distributions for each document we obtain representations that are able to distinguish between different meanings of a given word. Our models yield statistically significant improvements for the lexical substitution task indicating that commonly used single word representations, even when combined with contextual information, are insufficient for this task.

1 Introduction
Word representations in the form of dense vectors, or word embeddings, capture semantic and syntactic information (Mikolov et al., 2013a; Pennington et al., 2014) and are widely used in many NLP tasks (Zou et al., 2013; Levy and Goldberg, 2014; Tang et al., 2014; Gharbieh et al., 2016). Most of the existing models generate one representation per word and do not distinguish between different meanings of a word. However, many tasks can benefit from using multiple representations per word to capture polysemy (Reisinger and Mooney, 2010). There have been several attempts to build repositories for word senses (Miller, 1995;Navigli and Ponzetto, 2010), but this is laborious and limited to few languages. Moreover, defining a universal set of word senses is challenging as polysemous words can exist at many levels of granularity (Kilgarriff, 1997; Navigli, 2012).

In this paper, we introduce a model that uses a nonparametric Bayesian model, Hierarchical Dirichlet Process (HDP), to learn multiple topic-sensitive representations per word. Yao and van Durme (2011) show that HDP is effective in learning topics yielding state-of-the-art performance for sense induction. However, they assume that topics and senses are interchangeable, and train individual models per word making it difficult to scale to large data. Our approach enables us to use HDP to model senses effectively using large unannotated training data. We investigate to what extent distributions over word senses can be approximated by distributions over topics without assuming these concepts to be identical. The contributions of this paper are: (i) We propose three unsupervised, language-independent approaches to approximate senses with topics and learn multiple topic-sensitive embeddings per word. (ii) We show that in the Lexical Substitution ranking task our models outperform two competitive baselines.

2 Topic-Sensitive Representations
In this section we describe the proposed models. To learn topics from a corpus we use HDP (Teh et al., 2006; Lau et al., 2014). The main advantage of this model compared to non-hierarchical methods like the Chinese Restaurant Process (CRP) is that each document in the corpus is modeled using a mixture model with topics shared between all documents (Teh et al., 2005; Brody and Lapata, 2009). HDP yields two sets of distributions that we use in our methods: distributions over topics for words in the vocabulary, and distributions over topics for documents in the corpus.

Similarly to Neelakantan et al. (2014), we use neighboring words to detect the meaning of the context, however, we also use the two HDP dis-
Figure 1: Illustration of our topic-sensitive representation models: (a) hard-topic labeled representations (HTLE), (b) hard topic-labeled representations plus generic word representation (HTLEadd), (c) soft topic-labeled representations (STLE).

tributions. By doing so, we take advantage of the topic of the document beyond the scope of the neighboring words, which is helpful when the immediate context of the target word is not sufficiently informative. We modify the Skipgram model (Mikolov et al., 2013a) to obtain multiple topic-sensitive representations per word type using topic distributions. In addition, we do not cluster context windows and train for different senses of the words individually. This reduces the sparsity problem and provides a better representation estimation for rare words. We assume that meanings of words can be determined by their contextual information and use the distribution over topics to differentiate between occurrences of a word in different contexts, i.e., documents with different topics. We propose three different approaches (see Figure 1): two methods with hard topic labeling of words and one with soft labeling.

### 2.1 Hard Topic-Labeled Representations

The trained HDP model can be used to hard-label a new corpus with one topic per word type through sampling. Our first model variant (Figure 1(a)) relies on hard labeling by simply considering each word-topic pair as a separate vocabulary entry. To reduce sparsity on the context side and share the word-level information between similar contexts, we use topic-sensitive representations for target words (input to the network) and standard, i.e., unlabeled, word representations for context words (output). Note that this results in different input and output vocabularies. The training objective is then to maximize the log-likelihood of context words $w_{i+j}$ given the target word-topic pair $w^*_i$:

$$
L_{HardT-SG} = \frac{1}{I} \sum_{i=1}^{I} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{i+j} | w^*_i)
$$

where $I$ is the number of words in the training corpus, $c$ is the context size and $\tau$ is the topic assigned to $w_i$ by HDP sampling. The embedding of a word in context $h(w_i)$ is obtained by simply extracting the row of the input lookup table ($\mathbf{r}$) corresponding to the HDP-labeled word-topic pair:

$$h_{HTLE}(w_i) = \mathbf{r}(w^*_i) \tag{1}$$

A possible shortcoming of the HTLE model is that the representations are trained separately and information is not shared between different topic-sensitive representations of the same word. To address this issue, we introduce a model variant that learns multiple topic-sensitive word representations and generic word representations simultaneously (Figure 1(b)). In this variant (HTLEadd), the target word embedding is obtained by adding the word-topic pair representation ($\mathbf{r}'$) to the generic representation of the corresponding word ($\mathbf{r}_0$):

$$h_{HTLEadd}(w_i) = \mathbf{r}'(w^*_i) + \mathbf{r}_0(w_i) \tag{2}$$

### 2.2 Soft Topic-Labeled Representations

The model variants above rely on the hard labeling resulting from HDP sampling. As a soft alternative to this, we can directly include the topic distributions estimated by HDP for each document (Figure 1(c)). Specifically, for each update, we use the topic distribution to compute a weighted sum over the word-topic representations ($\mathbf{r}''$):

$$h_{STLE}(w_i) = \sum_{k=1}^{T} p(\tau_k | d_i) \mathbf{r}''(w^*_k) \tag{3}$$

where $T$ is the total number of topics, $d_i$ the document containing $w_i$, and $p(\tau_k | d_i)$ the probability assigned to topic $\tau_k$ by HDP in document $d_i$. The training objective for this model is:

$$L_{SoftT-SG} = \frac{1}{I} \sum_{i=1}^{I} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{i+j} | w_i, \tau)$$

442
where \( \tau \) is the topic of document \( d_i \) learned by HDP. The STLE model has the advantage of directly applying the distribution over topics in the Skipgram model. In addition, for each instance, we update all topic representations of a given word with non-zero probabilities, which has the potential to reduce the sparsity problem.

### 2.3 Embeddings for Polyseymous Words

The representations obtained from our models are expected to capture the meaning of a word in different topics. We now investigate whether these representations can distinguish between different word senses. Table 1 provides examples of nearest neighbors in different representation spaces using cosine similarity. \texttt{w2v} and GloVe are pre-trained embeddings from (Mikolov et al., 2013a) and (Pennington et al., 2014) respectively. \texttt{SGE} is the Skipgram baseline and \texttt{TSE} is our topic-sensitive Skipgram (cf. Equation (1)), both trained on Wikipedia. \( \tau_k \) stands for HDP-inferred topic \( k \).

<table>
<thead>
<tr>
<th>word</th>
<th>bat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-trained \texttt{w2v}</td>
<td>bats, batting, Pinch_hitter_Brayan_Pena, batsman, batted, Hawaiian_hoary, Batting</td>
</tr>
<tr>
<td>Pre-trained GloVe</td>
<td>bats, batting, Bat, catcher, fielder, hitter, outfield, hitting, batted, catchers, balls</td>
</tr>
<tr>
<td>SGE on Wikipedia</td>
<td>uroderma, magnirostrum, sorenseni, miniopiterus, promops, luctus, microcyotertis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>word</th>
<th>jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-trained \texttt{w2v}</td>
<td>jaguars, Macho_B, panther, lynx, rhino, lizard, tapir, tiger, leopard, Florida_panther</td>
</tr>
<tr>
<td>Pre-trained GloVe</td>
<td>jaguars, panther, mercedes, Jaguar, porsche, volvo, ford, audi, mustang, bmw, buick</td>
</tr>
<tr>
<td>SGE on Wikipedia</td>
<td>electromotive, viper, roadster, saleen, siata, chevrolet, camaro, dodge, nissan, escort</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>word</th>
<th>appeal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-trained \texttt{w2v}</td>
<td>appeals, appealing, appealed, Appeal, rehearing, apeal, Appealing, ruling, Appeals</td>
</tr>
<tr>
<td>Pre-trained GloVe</td>
<td>appeals, appealed, appealing, Appeal, court, decision, conviction, plea, sought</td>
</tr>
<tr>
<td>SGE on Wikipedia</td>
<td>court, case, appeals, appealed, decision, proceedings, disapproves, ruling</td>
</tr>
</tbody>
</table>

Table 1: Nearest neighbors of three examples in different representation spaces using cosine similarity. \texttt{w2v} and GloVe are pre-trained embeddings from (Mikolov et al., 2013a) and (Pennington et al., 2014) respectively. SGE is the Skipgram baseline and TSE is our topic-sensitive Skipgram (cf. Equation (1)), both trained on Wikipedia. \( \tau_k \) stands for HDP-inferred topic \( k \).

3 Evaluation

In this section we present the setup for our experiments and empirically evaluate our approach on the context-aware word similarity and lexical substitution tasks.

3.1 Experimental setup

All word representations are learned on the English Wikipedia corpus containing 4.8M documents (1B tokens). The topics are learned on a 100K-document subset of this corpus using the HDP implementation of Teh et al. (2006). Once the topics have been learned, we run HDP on the whole corpus to obtain the word-topic labeling (see Section 2.1) and the document-level topic distributions (Section 2.2). We train each model vari-
Table 2: Spearman’s rank correlation performance for the Word Similarity task on SCWS.

We compare our models to several baselines: Skipgram (SGE) and the best-performing multi-sense embeddings model per word type (MSSG) (Neelakantan et al., 2014). All model variants are trained on the same training data with the same settings, following suggestions by Mikolov et al. (2013a) and Levy et al. (2015). For MSSG we use the best performing similarity measure (avgSimC) as proposed by Neelakantan et al. (2014).

3.2 Context-Aware Word Similarity Task

Despite its shortcomings (Faruqui et al., 2016), word similarity remains the most frequently used method of evaluation in the literature. There are multiple test sets available but in almost all of them word pairs are considered out of context. To the best of our knowledge, the only word similarity data set providing word context is SCWS (Huang et al., 2012). To evaluate our models on SCWS, we run HDP on the data treating each word’s context as a separate document. We compute the similarity of each word pair as follows:

$$\text{Sim}(w_1, w_2) = \cos(h(w_1), h(w_2))$$

where \(h(w_i)\) refers to any of the topic-sensitive representations defined in Section 2. Note that \(w_1\) and \(w_2\) can refer to the same word.

Table 2 provides the Spearman’s correlation scores for different models against the human ranking. We see that with dimensions 100 and 300, two of our models obtain improvements over the baseline. The MSSG model of Neelakantan et al. (2014) performs only slightly better than our HLTE model by requiring considerably more parameters (600 vs. 100 embedding size).

3.3 Lexical Substitution Task

This task requires one to identify the best replacements for a word in a sentential context. The presence of many polysemous target words makes this task more suitable for evaluating sense embedding. Following Melamud et al. (2015) we pool substitutions from different instances and rank them by the number of annotators that selected them for a given context. We use two evaluation sets: LS-SE07 (McCarthy and Navigli, 2007), and LS-CIC (Kremer et al., 2014).

Unlike previous work (Szarvas et al., 2013; Kremer et al., 2014; Melamud et al., 2015) we do not use any syntactic information, motivated by the fact that high-quality parsers are not available for most languages. The evaluation is performed by computing the Generalized Average Precision (GAP) score (Kishida, 2005). We run HDP on the evaluation set and compute the similarity between target word \(w_t\) and each substitution \(w_s\) using two different inference methods in line with how we incorporate topics during training:

**Sampled (Smp):**

$$\text{Sim}_{\text{smp}}(w_s, w_t) = \cos(h(w_s), h(w_t)) + \frac{\sum \cos(h(w_s^\tau), o(w_c))}{C}$$

**Expected (Exp):**

$$\text{Sim}_{\text{exp}}(w_s, w_t) = \frac{\sum \tau \cdot p(\tau) \cdot p(\tau')}\cos(h(w_s^\tau), h(w_t^\tau')) + \frac{\sum \tau \cdot \tau' \cdot p(\tau) \cdot p(\tau') \cdot \cos(h(w_s^\tau), o(w_c))}{C}$$

where \(h(w_s^\tau)\) and \(h(w_t^\tau')\) are the representations for substitution word \(s\) with topic \(\tau\) and target word \(t\) with topic \(\tau'\) respectively (cf. Section 2), \(w_c\) are context words of \(w_t\) taken from a sliding window of the same size as the embeddings, \(o(w_c)\) is the context (i.e., output) representation of \(w_c\), and \(C\) is the total number of context words. Note
<table>
<thead>
<tr>
<th>Model</th>
<th>Dim = 100</th>
<th>Dim = 150</th>
<th>Dim = 300</th>
<th>Dim = 600</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n. v. adj. adv.</td>
<td>All</td>
<td>n. v. adj. adv.</td>
<td>All</td>
</tr>
<tr>
<td>SGE + C</td>
<td>37.2</td>
<td>31.6</td>
<td>36.6</td>
<td>37.2</td>
</tr>
<tr>
<td>HTLE</td>
<td>42.4</td>
<td>33.9</td>
<td>40.3</td>
<td>42.4</td>
</tr>
</tbody>
</table>

Table 4: GAP scores on the candidate ranking task on LS-SE07 for different part-of-speech categories.

that these two methods are consistent with how we train HTLE and STLE.

The sampled method, similar to HTLE, uses the HDP model to assign topics to word occurrences during testing. The expected method, similar to STLE, uses the HDP model to learn the probability distribution of topics of the context sentence and uses the entire distribution to compute the similarity. For the Skipgram baseline we compute the similarity \( \text{Sim}_{\text{SGE+C}}(w_s, w_t) \) as follows:

\[
\cos(h(w_s), h(w_t)) + \frac{1}{C} \sum_{q \in Q} \cos(h(w_s), o(w_q))
\]

which uses the similarity between the substitution word and all words in the context, as well as the similarity of target and substitution words.

Table 3 shows the GAP scores of our models and baselines.\(^1\) One can see that all models using multiple embeddings per word perform better than SGE. Our proposed models outperform both SGE and MSSG in both evaluation sets, with more pronounced improvements for LS-CIC. We further observe that our expected method is more robust and performs better for all embedding sizes.

Table 4 shows the GAP scores broken down by the main word classes: noun, verb, adjective, and adverb. With 100 dimensions our best model (HTLE) yields improvements across all POS tags, with the largest improvements for adverbs and smallest improvements for adjectives. When increasing the dimension size of embeddings the improvements hold up for all POS tags apart from adverbs. It can be inferred that larger dimension sizes capture semantic similarities for adverbs and context words better than other parts-of-speech categories. Additionally, we observe for both evaluation sets that the improvements are preserved when increasing the embedding size.

4 Related Work

While the most commonly used approaches learn one embedding per word type (Mikolov et al., 2013a; Pennington et al., 2014), recent studies have focused on learning multiple embeddings per word due to the ambiguous nature of language (Qiu et al., 2016). Huang et al. (2012) cluster word contexts and use the average embedding of each cluster as word sense embeddings, which yields improvements on a word similarity task. Nee-lakantan et al. (2014) propose two approaches, both based on clustering word contexts: In the first, they fix the number of senses manually, and in the second, they use an ad-hoc greedy procedure that allocates a new representation to a word if existing representations explain the context below a certain threshold. Li and Jurafsky (2015) used a CRP model to distinguish between senses of words and train vectors for senses, where the number of senses is not fixed. They use two heuristic approaches for assigning senses in a context: ‘greedy’ which assigns the locally optimum sense label to each word, and ‘expectation’ which computes the expected value for a word in a given context with probabilities for each possible sense.

5 Conclusions

We have introduced an approach to learn topic-sensitive word representations that exploits the document-level context of words and does not require annotated data or linguistic resources. Our evaluation on the lexical substitution task suggests that topic distributions capture word senses to some extent. Moreover, we obtain statistically significant improvements in the lexical substitution task while not using any syntactic information. The best results are achieved by our hard topic-labeled model which learns topic-sensitive representations by assigning topics to target words.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 639.021.646, and a Google Faculty Research Award. We thank the anonymous reviewers for their helpful comments.
References


Arvind Neelakantan, Jeevan Shankar, Alexandre Pas-sos, and Andrew McCallum. 2014. Efficient non-


