Topic Modeling for Word Sense Induction

Johannes Knopp, Johanna Völker*, and Simone Paolo Ponzetto

Data & Web Science Research Group
University of Mannheim, Germany
{johannes,johanna,simone}@informatik.uni-mannheim.de

Abstract. In this paper, we present a novel approach to Word Sense Induction which is based on topic modeling. Key to our methodology is the use of word-topic distributions as a means to estimate sense distributions. We provide these distributions as input to a clustering algorithm in order to automatically distinguish between the senses of semantically ambiguous words. The results of our evaluation experiments indicate that the performance of our approach is comparable to state-of-the-art methods whose sense distinctions are not as easily interpretable.

Keywords: word sense induction, topic models, lexical semantics.

1 Introduction

Computational approaches to the identification of meanings of words in context, a task commonly referred to as Word Sense Disambiguation (WSD) [12], typically rely on a fixed sense inventory such as WordNet [5]. But while WordNet provides a high-quality semantic lexicon in which fine-grained senses are connected by a rich network of meaningful semantic relations, it is questionable whether or not it provides enough coverage to be successfully leveraged for high-end, real-world applications, e.g., Web search, or whether these need to rely, instead, on sense distinction automatically mined from large text collections [15,4].

An alternative to WSD approaches is offered by methods which aim at automatically discovering senses from word (co-)occurrence in texts, i.e., performing so-called Word Sense Induction (WSI). WSI is viewed as a clustering task where the goal is to assign different occurrences of the same sense of a word to the same cluster and, by converse, to discover different senses of the same word in an unsupervised fashion by assigning their occurrences in text to different clusters. To this end, a variety of clustering methods can be used [12].

All clustering methods, however, crucially depend on the representation of contexts as their input. A standard approach is to view texts simply as vectors of words: a vector space model [17], in turn, can be complemented by dimensionality reduction techniques like, for instance, Latent Semantic Analysis (LSA) [3,18].

* Financed by a Margarete-von-Wrangell scholarship of the European Social Fund (ESF) and the Ministry of Science, Research and the Arts Baden-Württemberg.

I. Gurevych, C. Biemann, and T. Zesch (Eds.): GSCL 2013, LNAI 8105, pp. 97–103, 2013.

An alternative method proposed by Brody and Lapata [2], instead, consists of a generative model. In their approach, occurrences of ambiguous words in context are viewed as samples from a multinomial distribution over senses. These, in turn, are generated by sampling a sense from the multinomial distribution and then choosing a word from the sense-context distribution.

All in all, both vector space and generative models achieve competitive performance by exploiting the distributional hypothesis [8], i.e., the assumption that words that occur in similar contexts will have similar meanings. But while distributional methods have been successfully applied to the majority, if not all, of Natural Language Processing (NLP) tasks, they are still difficult to interpret for humans. LSA, for instance, can detect that a word may appear near completely different kind of words, but it does not, and cannot, encode explicitly in its representation that it has multiple senses.[1] In this paper we propose to overcome this problem by exploring the application of a state-of-the-art generative model, namely probabilistic topic models [16] to the task of Word Sense Induction. To this end, we propose to use Topic Models (TMs) as a way to estimate the distribution of word senses in text, and use topic-word distributions as a way to derive a semantic representation of ambiguous words in context that are later clustered to identify their senses. This is related to the approach presented in [10] where each topic is considered to represent a sense, while in this work we use all topics to represent a word's sense.

TMs, in fact, have been successfully used for a variety of NLP tasks, crucially including WSD [1] – thus providing a sound choice for a robust model – and, thanks to their ability to encode different senses of a polysemous word as a distribution over different topics, we expect them to provide a model which is easier to interpret for humans.

2 Method

The intuition behind TMs is that each topic has a predominant theme and ranks words in the vocabulary accordingly. A word's probability in a topic represents its importance with respect to the topic's theme and consequently reflects how dominant the respective theme is for the word. Our assumption in this work is that the meaning of a word consists of a distribution over topics. This assumption is analogue to the distributional hypothesis: words with a similar distribution over topics have similar meanings.

2.1 Representing Word Semantics with Topic Models

A topic model induced from a corpus provides topics $t_1, \ldots, t_i$ and each topic consists of a probability distribution over all words $w_1, \ldots, w_j$ in the vocabulary. Therefore, each word is associated with a probability for a topic, namely

[1] Cf., e.g., [7]: "the representation of words as points in an undifferentiated euclidean space makes it difficult for LSA to solve the disambiguation problem".

$p(w_{ji}) = p(w_i|t_j)$.[a] We will represent each word by means of a *topic signature* where the topics define the signature's features, and the word's probabilities in these topics constitute the feature values: $tsig(w_j) = \langle p(w_{j1}), p(w_{j2}), \ldots, p(w_{ji}) \rangle$

The whole document can be represented by aggregating the topic signatures of all the words in the document resulting in a single topic signature that describes the topical focus of the document. This representation of documents can be used to compute euclidean distances between documents which can be input to an established unsupervised clustering algorithm. This fact is utilized to identify word senses.

2.2 Identifying Word Senses

We use the arithmetic mean to aggregate the topic signatures of the words in any given document *doc*.

$$aggr(doc) = \frac{tsig(c_1) + \ldots + tsig(c_m)}{m} \quad \text{where} \quad c_1, \ldots, c_m \in doc \tag{1}$$

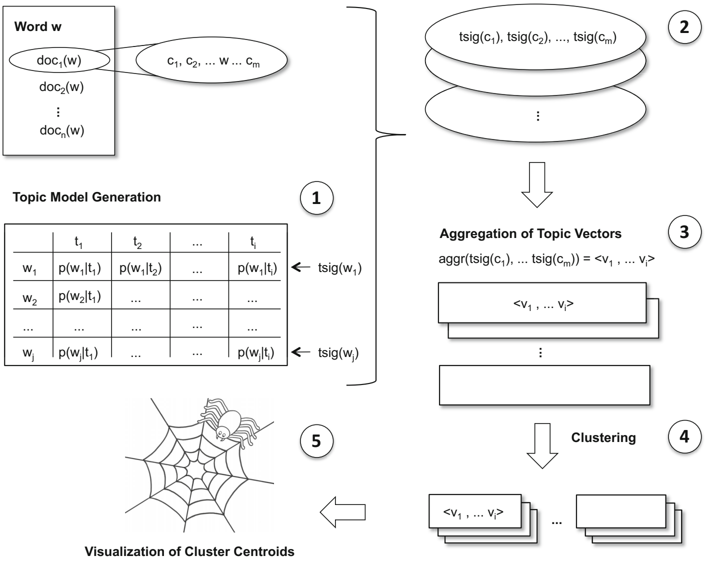
We write c when we refer to words in a document (tokens) while w is used for words in the vocabulary (types). Clustering the aggregated document vectors forms groups of documents that are close to each other and thus share similar topic distributions for the contained words. The cluster centroids constitute what we were looking for, a topic based representation of a sense, the *sense blueprints*. Because topics can be interpreted by humans by listing the top words of a topic, the sense blueprints can be interpreted as well. We just look at the top words in each topic along with its amplitude for a single cluster to get an intuition what theme is important in the cluster. This can also be helpful for comparing clusters and identify the topical dimensions where they differ from each other. The complete workflow is depicted in Figure 1.

3 Experiments

For our experiments we use data provided by the WSI task of SemEval 2010 [11]. For each of the 100 ambiguous words (50 nouns and 50 verbs, each having an entry in WordNet) the training data set provides text fragments of 2-3 sentences that were downloaded semi-automatically from the web. Each fragment is a short document that represents one distinct meaning of a target word.

The documents were preprocessed and only the lemmatized nouns were kept to be used as contextual features. We leave the inclusion of verbs for future work, because the average number of senses is higher for verbs than for nouns and thus would introduce more noise [13]. One topic model was built for each

[a] Usually the notation includes the information that t_j is sampled from a distribution itself, but as we do not rely on this sampling step in our approach we keep the notation simple.


Word w
doc1(w)
c1, c2, ... w ... cm
doc2(w)
⋮
docn(w)
Topic Model Generation
1
tsig(w1)
tsig(wj)
2
tsig(c1), tsig(c2), ..., tsig(cm)
⋮
Aggregation of Topic Vectors
3
aggr(tsig(c1), ... tsig(cm)) = <v1 , ... vi>
<v1 , ... vi>
⋮
Clustering
4
<v1 , ... vi>
...
5
Visualization of Cluster Centroids

	t_1	t_2	...	t_i
w_1	$p(w_1\|t_1)$	$p(w_1\|t_2)$	...	$p(w_1\|t_i)$
w_2	$p(w_2\|t_1)$	...	...	...
...	...	...	...	...
w_j	$p(w_j\|t_1)$	...	...	$p(w_j\|t_i)$

Fig. 1. The complete workflow

ambiguous word on its preprocessed documents using David Blei's TM implementation *ldac*[2]. We set the number of topics to 3, 4, . . . , 10 in order to test how the number of topics influences the results. The value for alpha was 0.3.

TMs are used as described in Section 2.1 to represent the documents in the test set as topic signatures. They are clustered using K-means clustering where the number of clusters was determined by the number of WordNet senses of a word.[3]

A spider diagram visualization of two of the cluster results for the word "promotion" can be found in Figure 2. Each dimension corresponds to one topic and is labeled with the respective most probable words. Each cluster centroid is a vector that spans a plane in the diagram, the number of documents per cluster is specified next to the cluster name in the legend. The word "promotion" is dominant in every topic because it appears in every document. With a higher number of topics more thematic details emerge. For example with the number of topics set to 3, there is one topic clearly dealing with advertising (lower right corner) while with 9 topics there are several topics dealing with aspects of advertising.

[2] Available at `http://www.cs.princeton.edu/~blei/lda-c/index.html`

[3] In order to evaluate if the presented approach is worth investigating further we leave the tuning of the number of cluster results for future work. By relying on external resources we avoid introducing more possible variation on the quality of the results by estimating the right number of clusters automatically.

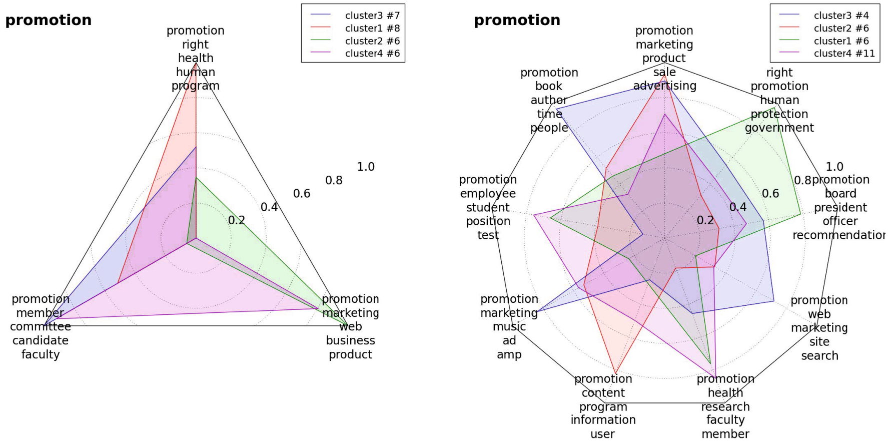


Fig. 2. Clustering result for the word "promotion" with 4 Clusters. The setup on the left side uses 3 topics, the one on the right 9 topics. The cluster labels in the diagrams do not correspond to each other.

4 Results

Following the standard evaluation of the Semeval WSI task, we used paired F-Score and V-measure [14] to evaluate our results. Paired F-Score is the harmonic mean of precision and recall. V-measure is the harmonic mean of the homogeneity and completeness scores of a clustering result. The results of our system are presented in Table 1.

The system's F-Score is not able to outperform the random baseline and does not improve with the number of topics used. Inspecting the detailed results shows that the reason for the significantly lower F-Score lies in low recall values. On average our system has 19% recall in comparison to 99% for MFS while the systems precision is 49% on average in comparison to 39% for MFS. This indicates that the choice of the number of clusters for each word – which was the number of WordNet senses in our experiment – is higher than than the actual number of senses in the gold standard. In fact, the gold standard data generally uses fewer senses than listed by WordNet. Interestingly the F-Score for nouns is similar to the score of verbs, while most other systems reported better results for verbs. The main reason in our opinion is that the training set for nouns was significantly bigger than for verbs, which resulted in more accurate topics. In general working with a corpus of short documents like the Semeval training data makes it harder to identify meaningful topics.

The outcome is much better for V-measure where the results indicate that the system learned useful information because it easily outperforms the random baseline. In comparison to the original results the system would achieve rank four out of 26 systems, with the best 3 systems reporting V-measures of 16.2% (*Hermit*) and 15.7% (both *UoY* and *KSU KDD*). Still, the results are not in the range of recent work like [10].

Table 1. V-measure (VM) and F-Score (FS) results for different topic settings along with the most frequent sense (MFS) and random baseline. The best clustering results are highlighted in **bold**.

Number of topics	**VM (%) (All)**	**VM (%) (Nouns)**	**VM (%) (Verbs)**	**FS (%) (All)**	**FS (%) (Nouns)**	**FS (%) (Verbs)**
3	12.3	14.6	9	**27**	**26.7**	**27.5**
4	12.6	15.3	8.7	25.6	25.7	25.4
5	12.9	15.2	9.6	25.4	25.9	26.2
6	12.4	14.4	9.4	24.8	24.1	25.7
7	12.5	14.6	9.4	24.8	24.2	25.7
8	13.2	15.9	9.2	25.4	25.2	25.7
9	13	15.5	9.4	24.9	24.3	25.9
10	**14**	**16.7**	**10.1**	25.9	25.5	26.3
MFS	0	0	0	63.5	57	72.2
Random	4.4	4.2	4.6	31.9	30.4	34.1

5 Conclusion and Outlook

In this paper we explored the embedding of information from a generative model in a vector space, in order to create interpretable clustering results. We presented an approach to WSI that uses probabilistic Topic Modeling to create a semantic representation for documents that allows clustering to find word senses. The results do not outperform other approaches to the Semeval 2010 WSI task, but the general idea might be helpful for tasks where interpretable results are desirable like near synonym detection [9] or exploratory data analysis.

There are many directions which we plan to explore in the very near future. Instead of the training set, a big corpus like Wikipedia could be used for creating the topic models. In general we expect the clustering performance to improve when bigger training data sets are available for the topic model creation. In this work, the complete generative model was not incorporated: The infered topic distribution for single documents could be used to add weights to the topic distribution of the words. In order to have a completely unsupervised WSI approach a clustering method that does not need to know the number of clusters beforehand needs to be developed. Additionally hierarchical topic models [6] could be used to find a more fine grained semantic representation.

References

1. Boyd-Graber, J., Blei, D., Zhu, X.: A topic model for word sense disambiguation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL), pp. 1024–1033 (2007)
2. Brody, S., Lapata, M.: Bayesian word sense induction. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 103–111 (2009)

3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
4. Di Marco, A., Navigli, R.: Clustering and diversifying web search results with graph-based word sense induction. Computational Linguistics 39(4) (2013)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998)
6. Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested chinese restaurant process. Advances in Neural Information Processing Systems 16, 106–114 (2004)
7. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological Review 114(2), 211 (2007)
8. Harris, Z.S.: Distributional structure. Word (1954)
9. Hirst, G.: Near-synonymy and the structure of lexical knowledge. In: AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity, pp. 51–56 (1995)
10. Lau, J.H., Cook, P., McCarthy, D., Newman, D., Baldwin, T.: Word sense induction for novel sense detection. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 591–601. Association for Computational Linguistics, Avignon (2012)
11. Manandhar, S., Klapaftis, I., Dligach, D., Pradhan, S.: Semeval-2010 task 14: Word sense induction & disambiguation. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 63–68. Association for Computational Linguistics, Uppsala (July 2010)
12. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41(2), 10 (2009)
13. Ng, H.T.: Getting serious about word sense disambiguation. In: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics, pp. 1–7 (1997)
14. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), vol. 410, p. 420 (2007)
15. Schuetze, H., Pedersen, J.O.: A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing and Management 33(3), 307–318 (1997)
16. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., Mcnamara, D., Dennis, S., Kintsch, W. (eds.) Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum (2007)
17. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Artificial Intelligence 37(1), 141–188 (2010)
18. Van de Cruys, T., Apidianaki, M., et al.: Latent semantic word sense induction and disambiguation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1476–1485 (2011)