Knowledge-Based Word Sense Disambiguation with Distributional Semantic Expansion

Hossein Rouhizadeh
Shahid Behehshti University
h.rouhizadeh@mail.sbu.ac.ir

Mehrnoush Shamsfard
Shahid Behehshti University
m-shams@sbu.ac.ir

Masoud Rouhizadeh
Johns Hopkins University
mrouhizadeh@gmail.com

Abstract

In this paper, we propose a new knowledge-based method for Word Sense Disambiguation (WSD). Using a pre-trained LDA model, we first retrieve the topics of document and assign each ambiguous word to one of the topics. For each possible sense s of a given word w in the document, we first compute the similarity between the WordNet gloss of s and the words of the assigned topic of w and linearly combine it with the rank value of s, computed by sense frequency information from SemCor. Finally, the sense with the highest final score will be chosen as the most probable one. Given the fact that LDA topics mostly consist of nouns, we rely on this method only for noun sense disambiguation. For other parts-of-speech, we utilize WordNet first sense, which is a hard to beat approach.

1 Introduction

Word Sense Disambiguation (WSD) is an open problem in Natural Language Processing (NLP), with the goal of identifying the most relevant meaning of ambiguous words in the context. WSD methods can be classified into two major classes: supervised and knowledge-based.

In this paper, we present a new knowledge-based method for WSD that also uses latent Dirichlet allocation (LDA) (Blei et al., 2003) for semantic expansion. Although KB approaches work on the basis of lexical resources such as WordNet, our method is still partially dependent on sense annotated corpora to extract sense-frequency (available in WordNet) which have shown useful in KB disambiguation pipelines, it can be a step forward to a completely KB method.

Although other KB systems only use context words to disambiguate words meaning, we find that it could be helpful to use words of document topics (retrieved from a pre-trained LDA model) which have semantic similarity with context words but may never occure in the context.

As observed in Gale et al. (1992) all repetitions of each word in a document have tendency to represent one sense. However, the WSD systems which use a window of words around the target word, due to the change of the context, have to run the disambiguation algorithm for all repetitions of a word. Using the whole document as context, our proposed method assigns all the repetitions of a word in a document to one of the document topics which results in proposing one meaning for all of them.

2 Method

Our WSD system takes a document as input and identifies the most relevant meaning of polysemous content words. For each ambiguous word w in the document, the system retrieves all possible senses from WordNet. Next, for each sense s it computes two scores and consider the linear combination of them as the final score of s as shown in Equation 1, and selects the sense with the highest overall score:

$$Final\ Score(s) = \alpha * Score_1(s) + (1 - \alpha) * Score_2(s) \quad (1)$$

When no topic is assigned to the word (for instance in compound words) only $score_2$ is used to disambiguate it. In the following, we provide details on the computing $score_1$ and $score_2$ of each sense.

For computing $score_1$ we use document topics (retrieved from a pre-trained LDA model). A key assumption in our method is that words in any given LDA topic are semantically inter-related. Using Gensim we first trained an LDA model on WikiPedia dump[1] (Rehurek and Sojka, 2010), and then we utilized existing functions of Gensim to retrieve topics of a document and assign each ambiguous word w to

[1] www.dumps.wikimedia.org

System	Senseval-2	Senseval-3	Semeval-07	Semeval-13	Semeval-15	All
Basile14	63.0	63.7	**56.7**	66.2	64.6	63.7
Moro14	67.0	63.5	51.6	66.4	70.3	65.5
Agirre 18	68.8	66.1	53.0	**68.8**	70.3	**67.3**
Chaplot 18	**69.0**	**66.9**	55.6	65.3	69.6	66.9
WSD-DSE	67.3	66.1	54.1	68.5	**71.7**	67.1

Table 1: comparison of our F-Scores with different knowledge-based WSD systems on the (Raganato et al., 2017) datasets. Our proposed method denoted by WSD-DSE. Best results are in bold.

one of these topics. In addition, we obtain the word vectors from word2vec word embeddings (Mikolov et al., 2013). Next, for each sense s of word w, we identify $score_1$ as the cosine similarity between the WordNet sense vector of s (the average of word vectors in the WordNet gloss of s, its hyponyms, and hypernyms) and a lexical cluster of topic words. The presence of unrelated words in topics will be problematic. To deal with this, we create a lexical cluster by the words which their similarity to the target word satisfy a certain threshold. As a result, for each s we calculate $score_1$ based on the similarity of s sense vector with only the words in the created cluster. More formally, if cluster units are w_1,...,w_n and sv_s be s sense vector, $score_1$ of s will be computed as follows:

$$Score_1(s) = \frac{1}{n}\sum_{i=1}^{n} Cos(sv_s, w_i) \tag{2}$$

Using word sense distributions in an English sense-annotated corpus such as SemCor (i.e. the probability of any given sense s for a particular word) has shown to be helpful in some WSD systems. Following Basile et al. (2014), we define the frequency of sense s of word w as the number of occurrence of s divided by the number of occurrence of w. We smooth frequencies by adding one to all counts. Moreover, WordNet senses are ordered based on their frequency, therefore, we add a weight factor to sense distribution equation to take this sense ordering into account. As a result, our system tends to give more priority to the first sense of words which is useful for words not present in SemCor and no information is available about their sense distribution. Score2 for sense s_i is then calculated as follows. Where the f_{s_i} and f_w are respectively frequency of WordNet ith sense of the word w and frequency of w in SemCor.

$$Score_2(s) = \frac{f_{s_i} + 1}{f_w + 1} + \frac{1}{i + 1} \tag{3}$$

3 Result

To evaluate our method, we use the (Raganato et al., 2017) datasets which include: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013) and SemEval-2015 (Moro and Navigli, 2015). Following Moro et al., 2014 we choosed trial dataset of the SemEval-2013 as devset. We fixed our parameters to maximize the f1-measure of our model on the devset. We found 0.9 as the best value for 0.25 as threshold for creating lexical cluster in topics and 100 as the number of topics. Table 1 shows the comparison of our F-Score with other knowledge-based which are Basile14 (Basile et al., 2014), Moro14 (Moro et al., 2014), Chaplot 18 (Chaplot and Salakhutdinov, 2018) and Agirre 18 (Agirre et al., 2018). We can see that our method can achieve a good performance on overall English words compared to the other knowledge-base methods.

4 Conclusion

In this paper, we presented a WSD system that uses LDA topics for semantic expansion of document words. Our system also uses sense frequency information from SemCor to give higher priority to the senses which are more probable to occur.

References

Agirre, E., O. L. de Lacalle, and A. Soroa (2018). The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd. *arXiv preprint arXiv:1805.04277*.

Basile, P., A. Caputo, and G. Semeraro (2014). An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1591–1600.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

Chaplot, D. S. and R. Salakhutdinov (2018). Knowledge-based word sense disambiguation using topic models. In *AAAI*.

Edmonds, P. and S. Cotton (2001). Senseval-2: overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 1–5. Association for Computational Linguistics.

Gale, W. A., K. W. Church, and D. Yarowsky (1992). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pp. 233–237. Association for Computational Linguistics.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Moro, A. and R. Navigli (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 288–297.

Moro, A., A. Raganato, and R. Navigli (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics 2*, 231–244.

Navigli, R., D. Jurgens, and D. Vannella (2013). Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Volume 2, pp. 222–231.

Pradhan, S. S., E. Loper, D. Dligach, and M. Palmer (2007). Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 87–92. Association for Computational Linguistics.

Raganato, A., J. Camacho-Collados, and R. Navigli (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Volume 1, pp. 99–110.

Rehurek, R. and P. Sojka (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Snyder, B. and M. Palmer (2004). The english all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.