Unsupervised Domain Tuning to Improve Word Sense Disambiguation

Judita Preiss and **Mark Stevenson**
j.preiss@sheffield.ac.uk and m.stevenson@dcs.shef.ac.uk
Department of Computer Science, University of Sheffield
211 Portobello, Sheffield, S1 4DP, UK

Abstract

The topic of a document can prove to be useful information for Word Sense Disambiguation (WSD) since certain meanings tend to be associated with particular topics. This paper presents an LDA-based approach for WSD, which is trained using any available WSD system to establish a sense per (Latent Dirichlet allocation based) topic. The technique is tested using three unsupervised and one supervised WSD algorithms within the SPORT and FINANCE domains giving a performance increase each time, suggesting that the technique may be useful to improve the performance of any available WSD system.

1 Introduction

Assigning each word its most frequent sense (MFS) is commonly used as a baseline in Word Sense Disambiguation (WSD). This baseline can be difficult to beat, particularly for unsupervised systems which do not have access to the annotated training data used to determine the MFS. However, it has also been shown that unsupervised methods can be used to identify the most likely sense for each ambiguous word type and this approach can be effective for disambiguation (McCarthy et al., 2004).

Knowledge of the domain of a document has been shown to be useful information for WSD. For example, Khapra et al. (2010) improve the performance of a graph-based WSD system using a small number of hand-tagged examples, but further examples would be required for each new domain. Agirre et al. (2009) automatically construct a thesaurus from texts in a domain which they use for WSD. Unfortunately, performance drops when the thesaurus is combined with information from local context. Stevenson et al. (2011) showed that performance of an unsupervised WSD algorithm can be improved by supplementing the context with domain information. Cai et al. (2007) use LDA to create an additional feature for a supervised WSD algorithm, by inferring topics for labeled training data. Boyd-Graber et al. (2007) integrate a topic model with WordNet and use it to carry out disambiguation and learn topics simultaneously. Li et al. (2010) use sense paraphrases to estimate probabilities of senses and carry out WSD. Koeling et al. (2005) showed that automatically acquiring the predominant sense of a word from a corpus from the same domain increases performance (over using a predominant sense acquired from a balanced corpus), but their work requires a separate thesaurus to be built for each domain under investigation. Navigli et al. (2011) extracted relevant terms from texts in a domain and used them to initialize a random walk over the WordNet graph.

Our approaches rely on a one sense per topic hypothesis (Gale et al., 1992), making use of topics induced using LDA – we present three novel techniques for exploiting domain information that are employable with any WSD algorithm (unsupervised or supervised). Using any WSD algorithm, we create a sense per topic distribution for each LDA topic, and the classification of a new document into a topic determines the sense distribution of the words within. Once a sense per topic distribution is obtained, no further WSD annotation of new texts is required. Instead of fixing domains, our technique

Proceedings of NAACL-HLT 2013, pages 680–684,
Atlanta, Georgia, 9–14 June 2013.

allows these to be dynamically created (using LDA) and we using four existing publicly available WSD algorithms (three unsupervised and one supervised) to show that our technique increases their performance with no changes to the original algorithm.

Section 2 briefly introduces LDA, while Section 3 describes our three techniques for adding domain information to a WSD algorithm. The WSD algorithms employed in the evaluation of our techniques are described in Section 4 with experiments and results in Section 5. Section 6 draws our conclusions and presents avenues for future work.

2 Latent Dirichlet allocation

LDA (Blei et al., 2003) is a widely used topic model, which views the underlying document distribution as having a Dirichlet prior. We employ a publicly available implementation of LDA[1] which has two main execution methods: parameter estimation (model building) and inference for new data (classification of a new document). Both invocation methods produce θ distributions (the topic-document distributions, i.e., $p(t_i|d)$ for t_i topics and d document), and ϕ distributions (word-topic distributions, i.e., $p(w_j|t_i)$ for words w_j). The parameter estimation phase also creates a list of n words most likely to be associated with each topic.

3 Using LDA for WSD

The underlying idea of our approach lies in deriving a document invariant sense distribution for each topic, $p(w, s|t)$. Once this word sense distribution is obtained, the underlying WSD algorithm is never needed again. We make the assumption that while the WSD algorithm may not be able to select the correct sense within an individual text due to insufficient domain information, the topic specific sense will be selected with a greater frequency over all documents pertaining to a topic, and thus the probability distributions over senses generated in this fashion should be more accurate.

Only the distribution $p(w, s|t)$ is dependent on an underlying WSD algorithm – once this distribution is obtained, it can be combined with the LDA derived θ distribution, $p(t|d_{new})$, to compute the desired word sense distribution within the new document d_{new}:

$$p(w, s|d_{new}) = \sum_t p(w, s|t)p(t|d_{new})$$

Sections 3.1, 3.2 and 3.3 describe three different methods for deriving $p(w, s|t)$, and we investigate the performance changes with different WSD algorithms: two versions of Personalized PageRank, described in Section 4.1, a similarity based WSD system outlined in Section 4.2, and a supervised graph based algorithm (Section 4.3).

3.1 Sense-based topic model (SBTM)

In its usual form, the ϕ distribution generated by LDA merely provides a word-topic distribution ($p(w|t)$). However, we modify the approach to directly output $p(w, s|t)$, but we remain able to classify (non WSD annotated) new text. The topic model is built from documents annotated with word senses using the chosen WSD algorithm.[2] The topic model created from this data is based on word-sense combinations and thus ϕ represents $p(w, s|t)$.

To classify new (non sense disambiguated) documents, the model is transformed to a word (rather than word-sense) based for: i.e., the $p(w, s|t)$ probabilities are summed over all senses of w to give resulting probabilities for the wordform. A new document, d_{new}, classified using this system gives rise to a number of distributions, including the probability of a topic given a document distribution ($p(t|d_{new})$).

3.2 Linear equations (LinEq)

If the topic model is created directly from wordforms, we can use the known probabilities $p(s|w, d)$ (obtained from the WSD algorithm), and $p(t|d)$ (from the LDA classifier) to yield an overdetermined system of linear equations of the form

$$p(s|w, d) = \sum_t p(s|w, t)p(t|d)$$

We use an existing implementation of linear least squares to find a solution (i.e. $p(s|w, t)$ for each t)

[1] `http://jgibblda.sourceforge.net/`.

[2] It is not crucial to word sense disambiguate all words in the text – a word can be passed to LDA in either its word-sense, disambiguated, form or in its raw form. While we do not attempt this in our work, it would be possible to build a model specifically for noun senses of a word, by including noun senses of the word and leaving the raw form for any non-noun occurrences.

by minimizing the sum of squared differences between the data values and their corresponding modeled values, i.e., minimizing:

$$\sum_d \left(p(s|w,d) - \sum_t p(s|w,t)p(t|d) \right)^2$$

3.3 Topic words (TopicWord)

The techniques presented in Sections 3.1 and 3.2 both require the WSD algorithm to annotate a reasonably high proportion of the data used to build the topic model. For systems which do not rely on word order, an alternative based on the most likely words per topic is possible: the LDA algorithm generates ϕ, a word-topic distribution. It is therefore possible to extract the most likely words per topic.

To acquire a sense-topic distribution for a topic t, each target word w is included in a bag of words which includes the most likely words for t and the unsupervised WSD algorithm is executed (w is added to the list if t does not already contain it). This technique is not applicable to non bag-of-words WSD algorithms, as structure is absent.

4 Word Sense Disambiguation

Only the topic model documents need to be automatically annotated with the chosen WSD system, after this, the WSD system is never applied again (an LDA classification determines the sense distribution) – this is particularly useful for supervised system which frequently have a long execution time. We explore three different types of WSD system: two versions of a knowledge base based system (Section 4.1), an unsupervised system (Section 4.2) and a supervised system (Section 4.3).

4.1 Personalized PageRank (ppr and w2w)

We use the freely available[3] Personalized PageRank algorithm (Agirre and Soroa, 2009) with WordNet 3.0. In Section 5 we present results from two options of the Personalized PageRank algorithm: *ppr*, which performs one PageRank calculation for a whole content, and *w2w*, which performs one PageRank calculation for every word in the context to be disambiguated.

[3]Available from `http://ixa2.si.ehu.es/ukb/`

4.2 WordNet similarity (sim)

We also evaluated another unsupervised approach, the Perl package WordNet::SenseRelate::AllWords (Pedersen and Kolhatkar, 2009), which finds senses of each word in a text based on senses of the surrounding words. The algorithm is invoked with Lesk similarity (Banerjee and Pedersen, 2002).

4.3 Vector space model (vsm)

An existing vector space model (VSM) based state-of-the-art supervised WSD system with features derived from the text surrounding the ambiguous word (Stevenson et al., 2008) is trained on Semcor (Miller et al., 1993).[4]

5 Experiments

5.1 Data

The approach is evaluated using a domain-specific WSD corpus (Koeling et al., 2005) which includes articles from the FINANCE and SPORTS domains taken from the Reuters corpus (Rose et al., 2002). This corpus contains 100 manually annotated instances (from each domain) for 41 words.[5]

The word-sense LDA topic models are created from 80,128 documents randomly selected from the Reuters corpus (this corresponds to a tenth of the entire Reuters corpus). LDA can abstract a model from a relatively small corpus and a tenth of the Reuters corpus is much more manageable in terms of memory and time requirements, particularly given the need to word sense disambiguate (some part of) each document in this dataset.[6]

[4]A version of Semcor automatically transformed to WordNet 3.0 available from `http://www.cse.unt.edu/~rada/downloads.html#semcor` was used in this work.

[5]Unfortunately, the entire domain-specific sense disambiguated corpus could not be used in the evaluation of our system, as the released corpus does not link each annotated sentence to its source document, and it is not always possible to recover these; approximately 87% of the data could be used. This dataset is available at `http://staffwww.dcs.shef.ac.uk/people/J.Preiss/downloads/source_texts.tgz`

[6]In this work, all 80,128 documents were word sense disambiguated. However, it would be possible to restrict this set to a smaller number, as long as a reliable distribution of word senses per topic could be obtained.

	ppr	w2w	sim	vsm
Baseline	36	41	23	27
SBTM model	39	43	30	31
LinEq	41	44	–	33
TopicWord	38	41	–	–

Table 1: Summary of results based on 150 topics

5.2 Results

Table 1 presents the performance results for the four WSD algorithms based on 150 topics. A range of topic values was explored, and 150 topics yielded highest performance, though the variance between the performance based on different topics (ranging from 50 to 250) was very small (0.4% difference to the average performance with 250 topics, and 3% with 50). The performance shown indicates the precision (number correct / number attempted). Recall is 100% in all cases.

The similarity algorithm (sim) fails on certain documents and therefore the linear equations technique could not be applied. The topic word technique (TopicWord) could not be evaluated using the similarity algorithm, due to the high sensitivity to word order within the test paragraph. In addition, the topic words technique is not applicable to supervised systems, due to its reliance on structured sentences. The best results with this technique were obtained with including all likely words with probabilities exceeding 0.001 and smoothing of 0.1 of the topic document distribution.

Using a Wilcoxon signed-rank test, the results were found to be significantly better over the original algorithms in every case (apart from TopicWords). Both the WordNet similarity (sim) and the VSM approach (vsm) have a lower performance than the two PPR based WSD algorithms (ppt and w2w). For example, sim assigns the same (usually incorrect) sense to all occurrences of the word *tie*, while both PPR based algorithms detect an obvious domain change. The vsm approach suffers from a lack of training data (only a small number of examples of each word appear in Semcor), while sim does not get enough information from the context.

As an interesting aside, the topic models based on word-sense combinations, as opposed to wordforms only, are more informative with less overlap. Examining the word *stake* annotated with the *w2w* WSD algorithm: only topic 1 contains *stake* among the top 12 terms associated with a topic in the word-sense model, while 10 topics are found in the wordform topic model. Table 2 shows the top 12 terms associated with topics containing the word *stake*.

Topic	Word-based model
39	say, will, company, share, deal, european, buy, agreement, **stake**, new, hungary, oil
63	say, share, united, market, offer, stock, union, percent, **stake**, will, point, new
90	say, will, fund, price, london, sell, **stake**, indonesia, court, investment, share, buy
91	say, market, bond, russia, press, party, **stake**, russian, country, indonesia, new, election
97	say, million, bank, uk, percent, share, **stake**, world, will, year, central, british
113	say, will, percent, week, billion, last, italy, plan, **stake**, year, budget, czech
134	say, china, percent, hong, kong, official, **stake**, billion, report, buy, group, year
142	say, percent, market, first, bank, rate, year, dealer, million, money, close, **stake**
145	say, will, new, brazil, dollar, group, percent, **stake**, year, one, make, do
147	say, yen, forecast, million, parent, market, share, will, profit, percent, **stake**, group
	Sense-based model
1	**stake*13286801-n**, share*13285176-n, sell*02242464-v, buy*02207206-v, have*02204692-v, group*00031264-n, company*08058098-n, percent*13817526-n, hold*02203362-v, deal*01110274-n, shareholder, interest*13286801-n

Table 2: The presence of *stake* within the word- and sense-based topic models

6 Conclusion

We present three unsupervised techniques based on acquiring LDA topics which can be used to improve the performance of a number of WSD algorithms. All approaches make use of topic information obtained using LDA and do not require any modification of the underlying WSD system. While the technique is dependent on the accuracy of the WSD algorithm, it consistently outperforms the baselines for all four different algorithms.

Acknowledgments

This research was supported by a Google Research Award. Our thanks also go to the two anonymous reviewers whose comments have made this paper much clearer.

References

Agirre, E., de Lacalle, O. L., and Soroa, A. (2009). Knowledge-based WSD on specific domains: performing better than generic supervised WSD. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1501–1506.

Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of EACL*.

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, pages 135–145.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boyd-Graber, J., Blei, D., and Zhu, X. (2007). A topic model for word sense disambiguation. In *Proceedings of the EMNLP-CoNLL*, pages 1024–1033.

Cai, J. F., Lee, W. S., and Teh, Y. W. (2007). Nus-ml: Improving word sense disambiguation using topic features. In *Proceedings of SEMEVAL*.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237.

Khapra, M., Kulkarni, A., Sohoney, S., and Bhattacharyya, P. (2010). All words domain adapted WSD: Finding a middle ground between supervision and unsupervision. In *Proceedings of ACL 2010*, pages 1532–1541, Uppsala, Sweden.

Koeling, R., Mccarthy, D., and Carroll, J. (2005). Domain specific sense distributions and predominant sense acquisition. In *Proceedings of Joint HLT-EMNLP05*, pages 419–426.

Li, L., Roth, B., and Sporleder, C. (2010). Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287.

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

Navigli, R., Faralli, S., Soroa, A., de Lacalle, O. L., and Agirre, E. (2011). Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In *CIKM*, pages 2317–2320.

Pedersen, T. and Kolhatkar, V. (2009). Wordnet::senserelate::allwords - a broad coverageword sense tagger that maximizes semantic relatedness (demonstration system). In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies Conference*, pages 17–20.

Rose, T. G., Stevenson, M., and Whitehead, M. (2002). The Reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 827–832.

Stevenson, M., Agirre, E., and Soroa, A. (2012). Exploiting domain information for word sense disambiguation of medical documents. *Journal of the American Medical Informatics Association*, 19(2):235–240.

Stevenson, M., Guo, Y., Gaizauskas, R., and Martinez, D. (2008). Knowledge sources for word sense disambiguation of biomedical text. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing at ACL*, pages 80–87.