XLING: Matching Query Sentences to a Parallel Corpus using Topic Models for Word Sense Disambiguation

Liling Tan and **Francis Bond**
Division of Linguistics and Multilingual Studies,
Nanyang Technological University
14 Nanyang Drive, Singapore 637332
alvations@gmail.com, bond@ieee.org

Abstract

This paper describes the XLING system participation in SemEval-2013 Crosslingual Word Sense Disambiguation task. The XLING system introduces a novel approach to skip the sense disambiguation step by matching query sentences to sentences in a parallel corpus using topic models; it returns the word alignments as the translation for the target polysemous words. Although, the topic-model base matching underperformed, the matching approach showed potential in the simple cosine-based surface similarity matching.

1 Introduction

This paper describes the XLING system, an unsupervised Cross-Lingual Word Sense Disambiguation (CLWSD) system based on matching query sentence to parallel corpus using topic models. CLWSD is the task of disambiguating a word given a context by providing the most appropriate translation in different languages (Lefever and Hoste, 2013).

2 Background

Topic models assume that latent topics exist in texts and each semantic topic can be represented with a multinomial distribution of words and each document can be classified into different semantic topics (Hofmann, 1999). Blei et al. (2003b) introduced a Bayesian version of topic modeling using Dirichlet hyper-parameters, Latent Dirichlet Allocation (LDA). Using LDA, a set of topics can be generated to classify documents within a corpus. Each topic will contain a list of all the words in the vocabulary of the corpus where each word is assigned a probability of occurring given a particular topic.

3 Approach

We hypothesized that sentences with different senses of a polysemous word will be classified into different topics during the LDA process. By matching the query sentence to the training sentences by LDA induced topics, the most appropriate translation for the polysemous word in the query sentence should be equivalent to translation of word in the matched training sentence(s) from a parallel corpus. By pursuing this approach, we escape the traditional mode of disambiguating a sense using a sense inventory.

4 System Description

The XLING_TnT system attempts the matching subtask in three steps (1) **Topicalize**: matching the query sentence to the training sentences by the most probable topic. (2) **Rank**: the matching sentences were ranked according to the cosine similarity between the query and matching sentences. (3) **Translate**: provides the translation of the polysemous word in the matched sentence(s) from the parallel corpus.

4.1 Preprocessing

The Europarl version 7 corpus bitexts (English-German, English-Spanish, English-French, English-Italian and English-Dutch) were aligned at word-level with GIZA++ (Och and Ney, 2003). The translation tables from the word-alignments were used to provide the translation of the polysemous word in the **Translate** step.

The English sentences from the bitexts were lemmatized using a dictionary-based lemmatiz-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 167–170, Atlanta, Georgia, June 14-15, 2013.

er: `xlemma`[1]. After the lemmatization, English stopwords[2] were removed from the sentences. The lemmatized and stop filtered sentences were used as document inputs to train the LDA topic model in the **`Topicalize`** step.

Previously, topic models had been incorporated as global context features into a modified naive Bayes network with traditional WSD features (Cai et al. 2007). We try a novel approach of integrating local context (N-grams) by using pseudo-word sentences as input for topic induction. Here we neither lemmatize or remove stops words. For example:

Original Europarl sentence: "*Education and cultural policies are important tools for creating these values*"

Lemmatized and stopped: "*education cultural policy be important tool create these values*"

Ngram pseudo-word: "*education_and_cultural and_cultural_policies cultural_policies_are are_important_tools important_tools_for tools_for_creating for_creating_these creating_these_values*"

4.2 Topicalize and Match

The **`Topicalize`** step of the system first (i) induced a list of topics and trained a topic model for each polysemous word using LDA, then (ii) allocated the topic with the highest probability to each training sentence.

Finally, at evaluation, (iii) the query sentences were assigned the most probable topic inferred using the trained topic models. Then the training sentences allocated with the same topic were considered as matching sentences for the next **`Rank`** step.

4.2.1 Topic Induction

Topic models were trained using Europarl sentences that contain the target polysemous words; one model per target word. The topic models were induced using LDA by setting the number of topics (*#topics*) as 50, and the alpha and beta hyper-parameters were symmetrically set at 1.0/*#topics*. Blei et al. (2003) had shown that the perplexity plateaus when *#topics* $\geq$ 50; higher perplexity means more computing time needed to train the model.

4.2.2 Topic Allocation

Each sentence was allocated the most probable topic induced by LDA. An induced topic contained a ranked list of tuples where the 2nd element in each tuple is a word that associated with the topic, the 1st element is the probability that the associated word will occur given the topic. The probabilities are generatively output using Variational Bayes algorithm as described in Hoffman et al. (2010). For example:

[*(0.0208, 'sport'), (0.0172, 'however'), (0.0170, 'quite'), (0.0166, 'maritime'), (0.0133, 'field'), (0.0133, 'air-transport'), (0.0130, 'appear'), (0.0117, 'arrangement'), (0.0117, 'pertain'), (0.0111, 'supervision')*]

4.2.3 Topic Inference

With the trained LDA model, we inferred the most probable topic of the query sentence. Then we extracted the top-10 sentences from the training corpus that shared the same top ranking topic.

The topic induction, allocation and inference were done separately on the lemmatized and stopped sentences and on the pseudo-word sentence, resulting in two sets of matching sentences. Only the sentences that were in both sets of matches are considered for the **`Rank`** step.

4.3 Rank

Matched sentences from the **`Topicalize`** step were converted into term vectors. The vectors were reweighted using tf-idf and ranked according to the cosine similarity with the query sentences. The top five sentences were piped into the `Translate` step.

4.4 Translate

From the matching sentences, the **`Translate`** step simply checks the `GIZA++` word alignment table and outputs the translation(s) of the target polysemous word. Each matching sentence,

[1] http://code.google.com/p/xlemma/

[2] Using the Page and Article Analyzer stopwords from http://www.ranks.nl/resources/stopwords.html

could output more than 1 translation depending on the target word alignment. As a simple way of filtering stop-words from target European languages, translations with less than 4 characters were removed. This effectively distills misaligned non-content words, such as articles, pronouns, prepositions, etc. To simplify the lemmatization of Spanish and French plural noun suffixes, the '*-es*' and '*-s*' are stemmed from the translation outputs.

The `XLING_TnT` system outputs one translation for each query sentence for the best result evaluation. It output the top 5 translations for the *out-of-five* evaluation.

4.5 Fallback

For the *out-of-five* evaluation, if the query returned less than 5 answers, the first fallback[3] appended the lemma of the Most Frequent Sense (according to Wordnet) of the target polysemous word in their respective language from the Open Multilingual Wordnet.[4] If the first fallback was insufficient, the second fallback appended the most frequent translation of the target polysemous word to the queries' responses.

4.6 Baseline

We also constructed a baseline for matching sentences by cosine similarity between the lemmas of the query sentence and the lemmas of each English sentence in the training corpus.[5] The baseline system is named `XLING_SnT` (Similar and Translate). The cosine similarity is calculated from the division of the vector product of the query and training sentence (i.e. numerator) by the root product of the vector's magnitude squared.

5 Results

Tables 1 and 2 present the results for the XLING system for best and out-of-five evaluation. Our system did worse than the task's baseline, i.e. the Most Frequent Translation (MFT) of the target word for all languages. Moreover the topic model based matching did worse than the cosine similarity matching baseline. The results show that matching on topics did not help. However, Li et al. (2010) and Anaya-Sanchez et al. (2007) had shown that pure topic model based unsupervised system for WSD should perform a little better than Most Frequent Sense baseline in coarse-grain English WSD. Hence it was necessary to perform error analysis and tweaking to improve the XLING system.

BEST	German	Spanish	French	Italian	Dutch
SnT	**8.13**	**19.59**	**17.33**	**12.74**	**9.89**
	(10.36)	(24.31)	(11.57)	(11.27)	(9.56)
TnT	5.28	18.60	16.48	10.70	7.40
	(5.82)	(24.31)	(11.63)	(7.54)	(8.54)
MFT	17.43	23.23	25.74	20.21	20.66
	(15.30)	(27.48)	(20.19)	(19.88)	(24.15)

Table 1: Precision and (Mood) for the best evaluation

OOF	German	Spanish	French	Italian	Dutch
SnT	**23.71**	**44.83**	**38.44**	32.38	**27.11**
	(30.57)	(50.04)	(32.45)	(29.17)	(27.31)
TnT	19.13	39.52	35.3	**33.28**	23.27
	(23.54)	(44.96)	(28.02)	(29.61)	(22.98)
MFT	38.86	53.07	51.36	42.63	43.59
	(44.35)	(57.35)	(47.42)	(41.69)	(41.97)

Table 2: Precision and (Mood) for the oof evaluation

6 Error Analysis and Modifications

Statistically, we could improve the robustness of the topic models in the **`Topicalize`** step by (i) tweaking the Dirichlet hyper-parameters to *alpha = 50/#topics*, *beta = 0.01* as suggested by Wang et al. (2009).

	BEST		OOF	
	Precision	**Mood**	**Precision**	**Mood**
German	6.50	6.71	20.98	25.18
Spanish	**14.77**	**19.43**	**40.22**	**45.67**
French	10.79	7.95	31.26	23.37
Italian	13.10	10.95	36.56	31.94
Dutch	7.42	7.47	21.66	20.42

Table 3: Evaluations on Hyper-parameter tweaks

Although the hyperparameters tweaks improves the scores for German and Dutch evaluations it brings the overall precision and mood precision of the other three languages down. Since the documents from each language are parallel, this

[3] Code sample for the fallback can be found at http://goo.gl/PbdK7

[4] http://www.casta-net.jp/~kuribayashi/multi/

[5] Code-snippet for the baseline can be found at http://pythonfiddle.com/surface-cosine-similarity

suggests that there is some language-dependency for LDA's hyperparameters.

By going through the individual queries and responses, several issues in the **`translate`** step need to be resolved to achieve higher precision; (i) German-English and Dutch-English word alignments containing compound words need to be segmented (e.g. *kraftomnibusverkehr →kraft omnibus verkehr)* and realigned such that the target word *coach* only aligns to *omnibus*, (ii) lemmatization of Italian, German and Dutch is crucial is getting the gold answers of the task (e.g. XLING answers *omnibussen* while the gold answers allowed *omnibus*). The use of target language lemmatizers, such as TreeTagger (Schmid, 1995) would have benefited the system.

7 Discussion

The main advantage of statistical language independent approaches is the ability to scale the system in any possible language. However language dependent processing remains crucial in building an accurate system, especially lemmatization in WSD tasks (e.g. *kraftomnibusverkehr*). We also hypothesize that more context would have improved the results of using topics: disambiguating senses solely from sentential context is artificially hard.

8 Conclusion

Our system has approached the CLWSD task in an unconventional way of matching query sentences to parallel corpus using topic models. Given no improvement from hyper-parameter tweaks, it reiterates Boyd-Graber, Blei and Zhu's (2007) assertion that while topic models capture polysemous use of words, they do not carry explicit notion of senses that is necessary for WSD. Thus our approach to match query sentences by topics did not perform beyond the MFT baseline in the CLWSD evaluation.

However, the surface cosine baseline, without any incorporation of any sense knowledge, had surprisingly achieved performance closer to MFT It provides a pilot platform for future work to approach the CLWSD as a vector-based document retrieval task on parallel corpora and providing the translation from the word alignments.

References

Henry Anaya-S´anchez, Aurora Pons-Porrata, and Rafael Berlanga-Llavori. 2007. Tkb-uo: Using sense clustering for wsd. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 322–325.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A Topic Model for Word Sense Disambiguation. In *Proc. of Empirical Methods in Natural Language Processing(EMNLP)*.

David M. Blei, Andrew Y. Ng, and Michael L. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jun-Fu Cai, Wee-Sun Lee and Yee-Whye Teh. 2007. Improving word sense disambiguation using topic features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1015–1023.

Christiane Fellbaum. (ed.) (1998) *WordNet: An Electronic Lexical Database*, MIT Press

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR '99*, Berkeley, CA, USA.

Matthew Hoffman, David Blei and Francis Bach. 2010. Online Learning for Latent Dirichlet Allocation. In Proceedings of NIPS 2010.

Els Lefever and Véronique Hoste. 2013. SemEval-2013 Task 10: Cross-Lingual Word Sense Disambiguation, In *Proceedings SemEval 2013, in conjunction with *SEM 2013*, Atlanta, USA.

Linlin Li, Benjamin Roth and Caroline Sporleder. Topic Models for Word Sense Disambiguation and Token-based Idiom Detection. In *Proc. of The 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010. Uppsala, Sweden.

Franz Josef Och, Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29:1. pp. 19-51.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, Edward Y. Chang. 2009. Plda: Parallel latent dirichlet allocation for large-scale applications. In *Proc. of 5th International Conference on Algorithmic Aspects in Information and Management*.