COLEUR and COLSLM: A WSD approach to Multilingual Lexical Substitution, Tasks 2 and 3 SemEval 2010

Weiwei Guo and Mona Diab
Center for Computational Learning Systems
Columbia University
{weiwei,mdiab}@ccls.columbia.edu

Abstract

In this paper, we present a word sense disambiguation (WSD) based system for multilingual lexical substitution. Our method depends on having a WSD system for English and an automatic word alignment method. Crucially the approach relies on having parallel corpora. For Task 2 (Sinha *et al.*, 2009) we apply a supervised WSD system to derive the English word senses. For Task 3 (Lefever & Hoste, 2009), we apply an unsupervised approach to the training and test data. Both of our systems that participated in Task 2 achieve a decent ranking among the participating systems. For Task 3 we achieve the highest ranking on several of the language pairs: French, German and Italian.

1 Introduction

In this paper, we present our system that was applied to the cross lingual substitution for two tasks in SEMEVAL 2010, Tasks 2 and 3. We adopt the same approach for both tasks with some differences in the basic set-up. Our basic approach relies on applying a word sense disambiguation (WSD) system to the English data that comes from a parallel corpus for English and a language of relevance to the task, language 2 (l2). Then we automatically induce the English word sense correspondences to l2. Accordingly, for a given test target word, we return its equivalent l2 words assuming that we are able to disambiguate the target word in context.

2 Our Detailed Approach

We approach the problem of multilingual lexical substitution from a WSD perspective. We adopt the hypothesis that the different word senses of ambiguous words in one language probably translate to different lexical items in another language. Hence, our approach relies on two crucial components: a WSD module for the source language (our target test words, in our case these are the English target test words) and an automatic word alignment module to discover the target word sense correspondences with the foreign words in a second language. Our approach to both tasks is unsupervised since we don't have real training data annotated with the target words and their corresponding translations into l2 at the onset of the problem.

Accordingly, at training time, we rely on automatically tagging large amounts of English data (target word instances) with their relevant senses and finding their l2 correspondences based on automatically induced word alignments. Each of these English sense and l2 correspondence pairs has an associated translation probability value depending on frequency of co-occurrence. This information is aggregated in a look-up table over the entire training set. An entry in the table would have a target word sense type paired with all the observed translation correspondences l2 word types. Each of the l2 word types has a probability of translation that is calculated as a normalized weighted average of all the instances of this l2 word type with the English sense aggregated across the whole parallel corpus. This process results in an English word sense translation table (WSTT). The word senses are derived from WordNet (Fellbaum, 1998). We expand the English word sense entry correspondences by adding the translations of the members of target word sense synonym set as listed in WordNet.

For alignment, we specifically use the GIZA++ software for inducing word alignments across the parallel corpora (Och & Ney, 2003). We apply GIZA++ to the parallel corpus in both directions English to l2 and l2 to English then take only the intersection of the two alignment sets, hence fo-

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 129–133,
Uppsala, Sweden, 15-16 July 2010.

cusing more on precision of alignment rather than recall.

For each language in Task 3 and Task 2, we use TreeTagger[1] to do the preprocessing for all languages. The preprocessing includes segmentation, POS tagging and lemmatization. Since TreeTagger is independent of languages, our system does not rely on anything that is language specific; our system can be easily applied to other languages. We run GIZA++ on the parallel corpus, and obtain the intersection of the alignments in both directions. Meanwhile, every time a target English word appears in a sentence, we apply our WSD system on it, using the sentence as context. From this information, we build a WSST from the English sense(s) to their corresponding foreign words. Moreover, we use WordNet as a means of augmenting the translation correspondences. We expand the word sense to its synset from WordNet adding the l2 words that corresponded to all the member senses in the synset yielding more translation variability.

At test time, given a test data target word, we apply the same WSD system that is applied to the training corpus to create the WSTT. Once the target word instance is disambiguated in context, we look up the corresponding entry in the WSTT and return the ranked list of l2 correspondences. We present results for best and for oot which vary only in the cut off threshold. In the BEST condition we return the highest ranked candidate, in the oot condition we return the top 10 (where available).[2]

Given the above mentioned pipeline, Tasks 2 and 3 are very similar. Their main difference lies in the underlying WSD system applied.

3 Task 2

3.1 System Details

We use a relatively simple monolingual supervised WSD system to create the sense tags on the English data. We use the SemCor word sense annotated corpus. SemCor is a subset of the Brown Corpus. For each of our target English words found disambiguated in the SemCor corpus, we create a sense profile for each of its senses. A sense profile is a vector of all the content words that occur in the context of this sense in the SemCor corpus. The dimensions of the vector are word types, as in a bag of words model, and the vector entries are the co-occurrence frequency of the word sense and the word type. At test time, given a a target English word, we create a bag of word types contextual vector for each instance of the word using the surrounding context. We compare the created test vector to the SemCor vectors and choose the highest most similar sense and use that for sense disambiguation. In case of ties, we return more than one sense tag.

Corpus	best		oot	
	P	R	P	R
T2-COLSLM	27.59	25.99	46.61	43.91
T2-COLEUR	19.47	18.15	44.77	41.72

Table 1: Precision and Recall results per corpus on Task 2 test set

3.2 Data

We use both naturally occurring parallel data and machine translation data. The data for our first Task 2 submission, T2-COLEUR, comprises naturally occurring parallel data, namely, the Spanish English portion of the EuroParl data provided by Task 3 organizers. For the machine translation data, we use translations of the source English data pertaining to the following corpora: the Brown corpus, WSJ, SensEval1, SensEval2 datasets as translated by two machine translation systems: Global Link (GL), Systran (SYS) (Guo & Diab, 2010). We refer to the translated corpus as the SALAAM corpus. The intuition for creating SALAAM (an artificial parallel corpus) is to create a balanced translation corpus that is less domain and genre skewed than the EuroParl data. This latter corpus results in our 2nd system for this task T2-COLSLM.

3.3 Results

Table 1 presents our overall results as evaluated by the organizers.

It is clear that the T2-COLSLM outperforms T2-COLEUR.

4 Task 3

4.1 System Details

Contrary to Task 2, we apply a context based unsupervised WSD module to the English side of the parallel data. Our unsupervised WSD method, as described in (Guo & Diab, 2009), is a graph based

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[2] Some of the target word senses had less than 10 l2 word correspondences.

unsupervised WSD method. Given a sequence of words $W = \{w_1, w_2...w_n\}$, each word w_i with several senses $\{s_{i1}, s_{i2}...s_{im}\}$. A graph G = (V,E) is defined such that there exists a vertex v for each sense. Two senses of two different words may be connected by an edge e, depending on their distance. That two senses are connected suggests they should have influence on each other, accordingly a maximum allowable distance is set. They explore 4 different graph based algorithms.We focus on the `In-Degree` graph based algorithm. The `In-Degree` algorithm presents the problem as a weighted graph with senses as nodes and similarity between senses as weights on edges. The `In-Degree` of a vertex refers to the number of edges incident on that vertex. In the weighted graph, the `In-Degree` for each vertex is calculated by summing the weights on the edges that are incident on it. After all the `In-Degree` values for each sense are computed, the sense with maximum value is chosen as the final sense for that word. In our implementation of the `In-Degree` algorithm, we use the JCN similarity measure for both Noun-Noun and Verb-Verb similarity calculation.

4.2 Data

We use the training data from EuroParl provided by the task organizers for the 5 different language pairs. We participate in all the language competitions. We refer to our system as T3-COLEUR.

4.3 Results

Table 2 shows our system results on Task 3, specified by languages.

4.4 Error Analysis and Discussion

As shown in Table 2, our system T3-COLEUR ranks the highest for the French, German and Italian language tasks on both best and oot. However the overall F-measures are very low. Our system ranks last for Dutch among 3 systems and it is middle of the pack for the Spanish language task. In general we note that the results for oot are naturally higher than for BEST since by design it is a more relaxed measure.

5 Related works

Our work mainly investigates the influence of WSD on providing machine translation candidates. Carpuat & Wu (2007) and Chan et al.(2007) show WSD improves MT. However, in (Carpuat & Wu, 2007) classical WSD is missing by ignoring predefined senses. They treat translation candidates as sense labels, then find linguistic features in the English side, and cast the disambiguation process as a classification problem. Of relevance also to our work is that related to the task of English monolingual lexical substitution. For example some of the approaches that participated in the SemEval 2007 excercise include the following. Yuret (2007) used a statistical language model based on a large corpus to assign likelihoods to each candidate substitutes for a target word in a sentence. Martinez et al. (2007) uses WordNet to find candidate substitutes, produce word sequence including substitutes. They rank the substitutes by ranking the word sequence including that substitutes using web queries. In (Giuliano C. *et al.*, 2007), they extract synonyms from dictionaries. They have 2 ways of ranking of the synonyms: by similarity metric based on LSA and by occurrence in a large 5-gram web corpus. Dahl et al. (2007) also extract synonyms from dictionaries. They present two systems. The first one scores substitutes based on how frequently the local context match the target word. The second one incorporates cosine similarity. Finally, Hassan et al. (2007) extract candidates from several linguistic resources, and combine many techniques and evidences to compute the scores such as machine translation, most common sense, language model and so on to pick the most suitable lexical substitution candidates.

6 Conclusions and Future Directions

In this paper we presented a word sense disambiguation based system for multilingual lexical substitution. The approach relies on having a WSD system for English and an automatic word alignment method. Crucially the approach relies on having parallel corpora. For Task 2 we apply a supervised WSD system to derive the English word senses. For Task 3, we apply an unsupervised approach to the training and test data. Both of our systems that participated in Task 2 achieve a decent ranking among the participating systems. For Task 3 we achieve the highest ranking on several of the language pairs: French, German and Italian.

In the future, we would like to investigate the usage of the Spanish and Italian WordNets for the

Language	best			oot		
	P	R	rank	P	R	rank
Dutch	10.71	10.56	3/3	21.47	21.27	3/3
Spanish	19.78	19.59	3/7	35.84	35.46	5/7
French	21.96	21.73	1/7	49.44	48.96	1/5
German	13.79	13.63	1/3	33.21	32.82	1/3
Italian	15.55	15.4	1/3	40.7	40.34	1/3

Table 2: Results of T3-COLEUR per language on Task 3 Test set

task. We would like to also expand our examination to other sources of bilingual data such as comparable corpora. Finally, we would like to investigate using unsupervised clustering of senses (Word Sense Induction) methods in lieu of the WSD approaches that rely on WordNet.

References

CARPUAT M. & WU D. (2007). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 61–72, Prague, Czech Republic: Association for Computational Linguistics.

CHAN Y. S., NG H. T. & CHIANG D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 33–40, Prague, Czech Republic: Association for Computational Linguistics.

DAHL G., FRASSICA A. & WICENTOWSKI R. (2007). SW-AG: Local Context Matching for English Lexical Substitution. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

FELLBAUM C. (1998). "wordnet: An electronic lexical database". MIT Press.

GIULIANO C., GLIOZZO A. & STRAPPARAVA C (2007). FBK-irst: Lexical Substitution Task Exploiting Domain and Syntagmatic Coherence. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

GUO W. & DIAB M. (2009). "Improvements to monolingual English word sense disambiguation". In *ACL Workshop on Semantics Evaluations*.

GUO W. & DIAB M. (2010). "Combining orthogonal monolingual and multilingual sources of evidence for All Words WSD". In *ACL 2010*.

HASSAN S., CSOMAI A., BANEA C., SINHA R. & MIHALCEA R. (2007). UNT: SubFinder: Combining Knowledge Sources for Automatic Lexical Substitution. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

IDE N. & V RONIS J. (1998). Word sense disambiguation: The state of the art. In *Computational Linguistics*, p. 1–40.

JIANG J. & CONRATH. D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan.

LEACOCK C. & CHODOROW M. (1998). Combining local context and wordnet sense similarity for word sense identification. In *WordNet, An Electronic Lexical Database*: The MIT Press.

LEFEVER C. & HOSTE V. (2009). SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, Boulder, Colorado.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *In Proceedings of the SIGDOC Conference*, Toronto.

MARTINEZ D., KIM S. & BALDWIN T. (2007). MELB-MKB: Lexical Substitution system based on Relatives in Context In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

M. PALMER, C. FELLBAUM S. C. L. D. & DANG H. (2001). English tasks: all-words and verb lexical sample. In *In Proceedings of ACL/SIGLEX Senseval-2*, Toulouse, France.

MIHALCEA R. (2005). Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 411–418, Vancouver, British Columbia, Canada: Association for Computational Linguistics.

MILLER G. A. (1990). Wordnet: a lexical database for english. In *Communications of the ACM*, p. 39–41.

NAVIGLI R. (2009). Word sense disambiguation: a survey. In *ACM Computing Surveys*, p. 1–69: ACM Press.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29(1)**, 19–51.

PEDERSEN B. & PATWARDHAN (2005). Maximizing semantic relatedness to perform word sense disambiguation. In *University of Minnesota Supercomputing Institute Research Report UMSI 2005/25*, Minnesotta.

PRADHAN S., LOPER E., DLIGACH D. & PALMER M. (2007). Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, p. 87–92, Prague, Czech Republic: Association for Computational Linguistics.

SINHA R. & MIHALCEA R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA.

SINHA R., MCCARTHY D. & MIHALCEA R. (2009). SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, Irvine, CA.

SNYDER B. & PALMER M. (2004). The english all-words task. In R. MIHALCEA & P. EDMONDS, Eds., *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, p. 41–43, Barcelona, Spain: Association for Computational Linguistics.

YURET D. (2007). KU: Word sense disambiguation by substitution. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.