LIMSI : Cross-lingual Word Sense Disambiguation using Translation Sense Clustering

Marianna Apidianaki
LIMSI-CNRS
Rue John Von Neumann
91403 Orsay Cedex, France
marianna@limsi.fr

Abstract

We describe the LIMSI system for the SemEval-2013 Cross-lingual Word Sense Disambiguation (CLWSD) task. Word senses are represented by means of translation clusters in different languages built by a cross-lingual Word Sense Induction (WSI) method. Our CLWSD classifier exploits the WSI output for selecting appropriate translations for target words in context. We present the design of the system and the obtained results.

1 Introduction

This paper describes the LIMSI system that participated in the Cross-Lingual Word Sense Disambiguation (CLWSD) task of SemEval-2013. The goal of CLWSD is to predict semantically correct translations for ambiguous words in context (Resnik and Yarowsky, 2000; Carpuat and Wu, 2007; Apidianaki, 2009). The CLWSD task of the SemEval-2013 evaluation campaign is a lexical sample task for English nouns and is divided into two subtasks: the *best* subtask where systems are asked to provide a unique good translation for words in context; the *out-of-five* (oof) subtask where systems can propose up to five semantically related translations for each target word instance (Lefever and Hoste, 2013). The CLWSD lexical sample contains 20 nouns and the test set is composed of 50 instances per noun. System performance is evaluated by comparing the system output to a set of gold standard annotations in five languages: French, Spanish, Italian, Dutch and German. Participating systems have to provide contextually appropriate translations for target words in context in each or a subset of the target languages.

We apply the CLWSD method proposed by Apidianaki (2009) to three bilingual tasks: English-Spanish, English-French and English-Italian. The method exploits the translation clusters generated in the three target languages by a cross-lingual Word Sense Induction (WSI) method. The WSI method clusters the translations of target words in a parallel corpus using source language context vectors. The same vectors are exploited during disambiguation in order to select the most appropriate translations for new instances of the target words in context.

2 System Description

2.1 Translation clustering

Contrary to monolingual WSI methods which group the instances of the words into clusters describing their senses, the cross-lingual WSI method used here clusters the translations of words in a parallel corpus. The corpus used for French consists of the English-French parts of Europarl (version 7) (Koehn, 2005) and of the JRC-Acquis corpus (Steinberger et al., 2006), joined together. For English-Spanish and English-Italian we only use the corresponding parts of Europarl. The corpora are first tokenized and lowercased using the Moses scripts, then lemmatized and tagged by part-of-speech (PoS) using the TreeTagger (Schmid, 1994). Words in the corpus are replaced by a lemma and PoS tag pair before word alignment, to resolve categorical ambiguities in context. The corpus is aligned in both translation directions with GIZA++ (Och and Ney, 2000)

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 178–182, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

Target word	French	Spanish	Italian
range	{ensemble, diversité, palette, nombre} {domaine} {portée} {éventail, nombre, gamme, série, ensemble}	{gama, serie, abanico, diversidad, variedad, espectro, conjunto} {cantidad, alcance, àmbito, número, tipo, espectro, rango} {amplitud}	{serie, gamma, spettro, numero, ventaglio} {ampiezza, portata} {settore, ambito} {diversitá, fascia}
mood	{climat, atmosphère}, {esprit, atmosphère, ambiance, humeur} {opinion} {volonté} {attitude}	{clima, atmòsfera, ambiente} {ànimo, sentimiento} {talante} {ànimo, clima, ambiente} {ànimo, humor, ambiente}	{clima} {atmosfera} {chiarezza, predisposizione} {opinione} {atteggiamento}
mission	{opération, mandat} {délégation, commission} {délégation, tâche, voyage, opération}	{función, cometido, objetivo, tarea} {viaje, tarea, delegación} {tarea, mandato, cometido}	{mandato, obiettivo, compito, mission, funzione, operazione,} {viaggio, mission, commissione, delegazione}

Table 1: Sense clusters generated by the WSI method in the three languages.

and three bilingual lexicons are built from the alignment results (one for each language pair) containing intersecting alignments. The lexicons contain noun translations of each English target word in the three languages. We keep French translations that translate the target words at least 10 times in the training corpus; for Spanish and Italian, where the corpus was smaller, the translation frequency threshold was set to 5.

For each translation T_i of a word w, we extract the content words that occur in the same sentence as w whenever it is translated by T_i. These constitute the features of the vector built for the translation. Let N be the number of features retained for each T_i from the corresponding source contexts. Each feature F_j ($1 \leq j \leq N$) receives a total weight $\mathrm{tw}(F_j, T_i)$ defined as the product of the feature's global weight, $\mathrm{gw}(F_j)$, and its local weight with that translation, $\mathrm{lw}(F_j, T_i)$. The global weight of a feature F_j is a function of the number N_i of translations (T_i's) to which F_j is related, and of the probabilities (p_{ij}) that F_j co-occurs with instances of w translated by each of the T_i's:

$$\mathrm{gw}(F_j) = 1 - \frac{\sum_{T_i} p_{ij} \log(p_{ij})}{N_i} \tag{1}$$

Each of the p_{ij}'s is computed as the ratio between the co-occurrence frequency of F_j with w when translated as T_i, denoted as $\mathrm{cooc\_frequency}(F_j, T_i)$, and the total number of features (N) seen with T_i:

$$p_{ij} = \frac{\mathrm{cooc\_frequency}(F_j, T_i)}{N} \tag{2}$$

The local weight $\mathrm{lw}(F_j, T_i)$ between F_j and T_i directly depends on their co-occurrence frequency:

$$\mathrm{lw}(F_j, T_i) = \log(\mathrm{cooc\_frequency}(F_j, T_i)) \tag{3}$$

The pairwise similarity of the translation vectors is calculated using the Weighted Jaccard Coefficient (Grefenstette, 1994). The similarity score of each translation pair is compared to a threshold locally defined for each w, which serves to distinguish strongly related translations from semantically unrelated ones. The semantically related translations of a word w are then grouped into clusters. Translation pairs with a score above the threshold form a set of initial clusters that might be further enriched with other translations through an iterative procedure, provided that there are other translations that are strongly related to the elements in the cluster.[1] The clustering stops when all the translations of w have been clustered and all their relations have been checked. The algorithm performs a soft clustering so translations might be found in different clusters. Final clusters are characterized by global connectivity, meaning that all their elements are linked by pertinent relations. Table 1 gives examples of clusters generated for CLWSD target words in the three languages. The clusters group translations carrying the same sense and their overlaps describe relations between senses. The translation clusters serve as the target words' candidate senses from which one has to be selected during disambiguation.

[1] The thresholding procedure and the clustering algorithm are described in detail in Apidianaki and He (2010).

Subtask	Metric	Spanish			French			Italian		
		LIMSI	Baseline	Best system	LIMSI	Baseline	Best system	LIMSI	Baseline	Best system
Best	P/R	24,7	23,23	32,16	24,56	25,73	30,11	21,2	20,21	25,66
	Mode P/R	32,09	27,48	37,11	22,16	20,19	26,62	23,06	19,88	31,61
OOF	P/R	49,01	53,07	61,69	45,37	51,35	59,8	40,25	42,62	53,57
	Mode P/R	51,41	57,34	64,65	39,54	47,42	57,57	47,21	41,68	56,61
OOF (dupl)	P/R	98,6	-	-	101,75	-	-	90,23	-	-
	Mode P/R	51,41	-	-	39,54	-	-	47,21	-	-

Table 2: Results at the SemEval 2013 CLWSD task.

2.2 Word Sense Disambiguation

The vectors used for clustering the translations also serve for disambiguating new instances of the target words in context. The new contexts are tokenized, lowercased, PoS tagged and lemmatized to facilitate comparison with the vectors. We use the features shared by each pair of clustered translations, or the vector corresponding to the translation in an one-element cluster. If no CFs exist between the new context and a pair of translations, WSD is performed by comparing context information separately to the vector of each clustered translation. Once the common features (CFs) between the vectors and the new context are identified, a score is calculated corresponding to the mean of the weights of the CFs with the translations (weights assigned to the features during WSI). In formula 4, CF_j is the set of CFs and N_{CF} is the number of translations T_i characterized by a CF.

$$wsd\_score = \frac{\sum_{i=1}^{N_{CF}} \sum_j w(T_i, CF_j)}{N_{CF} \cdot |CF_j|} \qquad (4)$$

The cluster containing the highest ranked translation or translation pair is selected and assigned to the new target word instance. If the translations are present in more than one clusters, a new score is calculated using equation 4 and by taking into account the weights of the CFs with the other translations (T_i's) in the cluster.

3 Evaluation

Systems participating to the CLWSD task have to provide the most plausible translation for a word in context in the *best* subtask, and five semantically correct translations in *oof*. The baselines provided by the organizers are based on the output of GIZA++ alignments on Europarl. The *best* baseline corresponds to the most frequent translation of the target word in the corpus and the *oof* baseline to the five most frequent translations. Our CLWSD system makes predictions in three languages for all 1000 test instances. If the selected cluster contains five translations, all of them are proposed in the *oof* subtask while if it is bigger, the five most frequent translations are selected. In case of smaller clusters, the *best* translation is repeated in the output until reaching five suggestions. Duplicate suggestions were allowed in previous cross-lingual SemEval tasks as a means to boost translations with high confidence (Mihalcea et al., 2010). However, as in this year's CLWSD task the *oof* system output has been post-processed by the organizers to keep only unique translations, the number of predictions made by our system for some words has been significantly reduced. This has had a negative impact on the *oof* results, as we will show in the next section.

For selecting *best* translations, each translation of a target word w is scored separately by comparing its vector to the new context. In case the highest-ranked translation has a score lower than 1, the system falls back to using the most frequent translation (MFT). To note that frequency information differs from the one used in the MFT baseline because words in our corpus were replaced by a lemma and PoS tag pair prior to alignment. The discrepancy is more apparent in French where MFT is the most frequent translation of the target word in the joint Europarl and JRC-Acquis corpus. Five teams participated to the CLWSD task with a varying number of systems: twelve systems provided output for Spanish and ten for French and Italian.

4 Results

The results obtained by our system for the *best* and *oof* evaluations in the three languages (Spanish, French and Italian) are presented in Table 2. We contrast them with the baselines provided by the organizers and with the score of the system that performed best in each subtask. Our system made suggestions for all test instances, so recall (R) coincides with precision (P). The baselines are quite challenging, as noted in Lefever and Hoste (2010), especially the *oof* one which contains the five most frequent Europarl translations. These often correspond to the most frequent translations from different sense clusters and cover multiple senses of the target word.

Our system outperforms the *best* baseline in all languages except for French, where the *best* score lies near below the baseline. This is not surprising given that the training corpus for French is the joint Europarl and JRC-Acquis corpus, which causes a discrepancy between the selected *best* translations and the baseline. The mode precision and recall scores reflect the capacity of the system to predict the translations that were most frequently selected by the annotators for each instance and are thus considered as the most plausible ones. Our system outperforms the mode *best* baselines for all languages.

In the *oof* task, the system has been penalized by the elimination of duplicate translations from the output after submission. In previous work, the CLWSD system gave very good results when applied, with some slight variations, to the *out-of-ten* subtask of the SemEval-2010 Cross-Lingual Lexical Substitution task where duplicates served to promote translations with high confidence (Mihalcea et al., 2010; Apidianaki, 2011). Here, after the post-processing step, *oof* suggestions contain in many cases less than five translations which explains the low scores. In Table 2 we provide *oof* results before and after post-processing the output and show how the system was affected by this change in evaluation. By boosting plausible translations, precision and recall scores get higher while mode scores are naturally not affected.[2] As the other systems might have been impacted to different extents by this change, we cannot estimate how this affects the global system ranking.

5 Discussion and future work

We presented a CLWSD system that uses translation clusters as candidate senses. Disambiguation is performed by comparing the feature vectors that served for clustering to the context of new target word instances. We observe that the use of a bigger corpus – as in the case of French – not only does not help in this task but actually has a negative impact on the results. This is due to the inclusion of translations that are not present in the gold standard (built from Europarl) and to the discrepancy between most frequent translations in the large corpus and the Europarl MFT baselines. This discrepancy affects all three languages, as words in the training corpora were replaced by lemma and PoS tag pairs prior to alignment.

It is important to note that our CLWSD method exploits the output of another unsupervised semantic analysis method (WSI) which groups the translations into clusters. This is an important feature of the system and affects the results in two ways. First, the translation clusters of a word constitute its candidate senses from which the CLWSD method selects the most appropriate one for a given context. This means that no variation regarding the contents of a cluster is permitted and that different instances are tagged by the same set of translations, contrary to the gold standard annotations which might, at the same time, be very close and contain some variations. In the system output, this is the case only when overlapping clusters are selected for different instances. Moreover, given that the WSI method is automatic and that the clusters are not manually validated, the noise that might be introduced during clustering is propagated and reflected in the disambiguation results. So, if a cluster contains one or more noisy translations, these occur in the disambiguation output and naturally count as wrong predictions. However, in an application setting like Machine Translation (MT), the translation clusters could be filtered using information from the target language context. Future work will focus on integrating this method into MT systems and examining ways for optimally taking advantage of CLWSD predictions in this context.

[2] Precision scores might be inflated, as in the case of French, because the credit for each item is not divided by the number of predictions and the annotation frequencies are used.

References

Marianna Apidianaki and Yifan He. 2010. An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT-10)*, pages 219–226, Paris, France.

Marianna Apidianaki. 2009. Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, pages 77–85, Athens, Greece.

Marianna Apidianaki. 2011. Unsupervised Cross-Lingual Lexical Substitution. In *Proceedings of the First workshop on Unsupervised Learning in NLP in conjunction with EMNLP*, pages 13–23, Edinburgh, Scotland, July. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*, pages 61–72.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 15–20, Uppsala, Sweden.

Els Lefever and Véronique Hoste. 2013. SemEval-2013 Task 10: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (*SEM 2013)*, pages 63–72, Atlanta, USA.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 9–14, Uppsala, Sweden.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, pages 440–447, Hongkong, China.

Philip Resnik and David Yarowsky. 2000. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(3):113–133.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147.