SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation

Els Lefever[1,2] **and Véronique Hoste**[1,3]

[1]LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
[2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium
[3]Department of Linguistics, Ghent University
Blandijnberg 2, 9000 Gent, Belgium

{Els.Lefever,Veronique.Hoste}@hogent.be

Abstract

The goal of the Cross-lingual Word Sense Disambiguation task is to evaluate the viability of multilingual WSD on a benchmark lexical sample data set. The traditional WSD task is transformed into a multilingual WSD task, where participants are asked to provide contextually correct translations of English ambiguous nouns into five target languages, viz. French, Italian, English, German and Dutch. We report results for the 12 official submissions from 5 different research teams, as well as for the ParaSense system that was developed by the task organizers.

1 Introduction

Lexical ambiguity remains one of the major problems for current machine translation systems. In the following French sentence "Je cherche des idées pour manger de l'avocat"[1], the word "avocat" is clearly referring to the fruit, whereas both Google Translate[2] as well as Babelfish[3] translate the word as "lawyer". Although "lawyer" is a correct translation of the word "avocat", it is the wrong translation in this context. Other language technology applications, such as Question Answering (QA) systems or information retrieval (IR) systems, also suffer from the poor contextual disambiguation of word senses. Word sense disambiguation (WSD) is still considered one of the most challenging problems within language technology today. It requires the construction of an artificial text understanding as the system should detect the correct word sense based on the context of the word. Different methodologies have been investigated to solve the problem; see for instance Agirre and Edmonds (2006) and Navigli (2009) for a detailed overview of WSD algorithms and evaluation.

This paper reports on the second edition of the "Cross-Lingual Word Sense Disambiguation" (CLWSD) task, that builds further on the insights we gained from the SemEval-2010 evaluation (Lefever and Hoste, 2010b) and for which new test data were annotated. The task is an unsupervised Word Sense Disambiguation task for English nouns, the sense label of which is composed of translations in different target languages (viz. French, Italian, Spanish, Dutch and German). The sense inventory is built up on the basis of the Europarl parallel corpus; all translations of a polysemous word were manually grouped into clusters, which constitute different senses of that given word. For the test data, native speakers assigned a translation cluster(s) to each test sentence and gave their top three translations from the predefined list of Europarl translations, in order to assign weights to the set of gold standard translations.

The decision to recast the more traditional monolingual WSD task into a cross-lingual WSD task was motivated by the following arguments. Firstly, using multilingual unlabeled parallel corpora contributes to clearing the data acquisition bottleneck for WSD, because using translations as sense labels excludes the need for manually created sense-tagged corpora

[1]English translation: "I'm looking for ideas to eat avocado".
[2]http://translate.google.com
[3]http://be.bing.com/translator/

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166, Atlanta, Georgia, June 14-15, 2013.

and sense inventories such as WordNet (Fellbaum, 1998) or EuroWordNet (Vossen, 1998). Moreover, as there is fairly little linguistic knowledge involved, the framework can be easily deployed for a variety of different languages. Secondly, a cross-lingual approach also deals with the sense granularity problem; finer sense distinctions are only relevant as far as they get lexicalized in different translations of the word. If we take the English word "head" as an example, we see that this word is always translated as "hoofd" in Dutch (both for the "chief" and for the 'body part" sense of the word). At the same time, the subjectivity problem is tackled that arises when lexicographers have to construct a fixed set of senses for a particular word that should fit all possible domains and applications. In addition, the use of domain-specific corpora allows to derive sense inventories that are tailored towards a specific target domain or application and to train a dedicated CLWSD system using these particular sense inventories. Thirdly, working immediately with translations instead of more abstract sense labels allows to bypass the need to map abstract sense labels to corresponding translations. This makes it easier to integrate a dedicated WSD module into real multilingual applications such as machine translation (Carpuat and Wu, 2007) or information retrieval (Clough and Stevenson, 2004).

Many studies have already shown the validity of a cross-lingual approach to Word Sense Disambiguation (Brown et al., 1991; Gale and Church, 1993; Ng et al., 2003; Diab, 2004; Tufiş et al., 2004; Chan and Ng, 2005; Specia et al., 2007; Apidianaki, 2009). The Cross-lingual WSD task contributes to this research domain by the construction of a dedicated benchmark data set where the ambiguous words were annotated with the senses from a multilingual sense inventory extracted from a parallel corpus. This benchmark data sets allows a detailed comparison between different approaches to the CLWSD task.

The remainder of this paper is organized as follows. Section 2 focuses on the task description and briefly recapitalizes the construction of the sense inventory and the annotation procedure of the test sentences. Section 3 presents the participating systems to the task, whereas Section 4 gives an overview of the experimental setup and results. Section 5 concludes this paper.

2 Task set up

The "Cross-lingual Word Sense Disambiguation" (CLWSD) task was organized for the first time in the framework of SemEval-2010 (Lefever and Hoste, 2010b) and resulted in 16 submissions from five different research teams. Many additional research teams showed their interest and downloaded the trial data, but did not manage to finish their systems in time. In order to gain more insights into the complexity and the viability of cross-lingual WSD, we proposed a second edition of the task for SemEval-2013 for which new test data were annotated.

The CLWSD task is an unsupervised Word Sense Disambiguation task for a lexical sample of twenty English nouns. The sense label of the nouns is composed of translations in five target languages (viz. Spanish, French, German, Italian and Dutch) and the sense inventory is built up on the basis of the Europarl parallel corpus[4]. This section briefly describes the data construction process for the task. For a more detailed description of the gold standard creation and data annotation process, we refer to Lefever and Hoste (2010a; 2010b).

2.1 Sense inventory

The starting point for the gold standard sense inventory creation was the parallel corpus Europarl. We selected six languages from Europarl (English and the five target languages) and only considered the 1-1 sentence alignments between English and the five target languages[5]. In order to obtain the multilingual sense inventory we:

1. performed word alignment on the parallel corpus in order to find all possible translations for our set of ambiguous focus nouns

2. clustered the resulting translations by meaning and manually lemmatized all translations

The resulting sense inventory was then used to annotate the sentences in the test set that was developed for the SemEval-2013 CLWSD task.

[4] http://www.statmt.org/europarl/

[5] This six-lingual sentence-aligned subcorpus of Europarl can be downloaded from http://lt3.hogent.be/semeval/.

2.2 Test data

For the creation of the test data set, we manually selected 50 sentences per ambiguous focus word from the part of the ANC corpus that is publicly available[6]. In total, 1000 sentences were annotated using the sense inventory that was described in Section 2.1. Three annotators per target language were asked to first select the correct sense cluster and next to choose the three contextually most appropriate translations from this sense cluster. They could also provide fewer translations in case they could not find three good translations for this particular occurrence of the test word. These translations were used to (1) compose the set of gold standard translations per test instance and (2) to assign frequency weights to all translations in the gold standard (e.g. translations that were chosen by all three annotators get a frequency weight of 3 in the gold standard).

2.3 Evaluation tasks

Two subtasks were proposed for the Cross-lingual WSD task: a *best evaluation* and an *Out-of-five* evaluation task. For the *best* evaluation, systems can propose as many guesses as the system believes are correct, but the score is divided by the number of guesses. In case of the *Out-of-five* evaluation, systems can propose up to five guesses per test instance without being penalized for wrong translation suggestions. Both evaluation tasks are explained in more detail in Section 4.1.

3 Systems

3.1 Systems participating to the official CLWSD evaluation campaign

Five different research teams participated to the CLWSD task and submitted up to three different runs of their system, resulting in 12 different submissions for the task. All systems took part in both the *best* and the *Out-of-five* evaluation tasks. These systems took very different approaches to solve the task, ranging from statistical machine translation, classification and sense clustering to topic model based approaches.

The XLING team (Tan and Bond, 2013) submitted three runs of their system for all five target languages. The first version of the system presents a topic matching and translation approach to CLWSD (*TnT* run), where LDA is applied on the Europarl sentences containing the ambiguous focus word in order to train topic models. Each sentence in the training corpus is assigned a topic that contains a list of associated words with the topic. The topic of the test sentence is then inferred and compared to the matching training sentences by means of the cosine similarity between the training and test vectors. WordNet (WN) is used as a fallback in case the system returns less than 5 answers. The second - and best performing - flavor of the system (*SnT* run) calculates the cosine similarity between the context words of the test and training sentences. The output of the system then contains the translation that results from running word alignment on the focus word in the training corpus. As a fallback, WordNet is again used. The WN senses are sorted by frequency in the SemCor corpus and the corresponding translation is selected from the aligned WordNet in the target language. The third run of the system (*merged*) combines the output from the other two flavors of the system.

The LIMSI system (Apidianaki, 2013) applies an unsupervised CLWSD method that was proposed in (Apidianaki, 2009) for three target languages, viz. Spanish, Italian and French. First, word alignment is applied on the parallel corpus and three bilingual lexicons are built, containing for each focus word the translations in the three target languages. In a next step, a vector is built for each translation of the English focus word, using the cooccurrences of the word in the sentences in which it gets this particular translation. A clustering algorithm then groups the feature vectors using the Weighted Jaccard measure. New instances containing the ambiguous focus word are then compared to the training feature vectors and assigned to one of the sense clusters. In case the highest-ranked translation in the cluster has a score below the threshold, the system falls back to the most frequent translation.

Two very well performing systems take a classification-based approach to the CLWSD task: the HLTDI and WSD2 systems. The HLTDI system (Rudnick et al., 2013) performs word alignment on the intersected Europarl corpus to locate training instances containing the ambiguous focus words. The first flavor of the system (*l1*) uses a maxent clas-

[6] http://www.americannationalcorpus.org/

sifier that is trained over local context features. The L2 model (*l2* run) also adds translations of the focus word into the four other target languages to the feature vector. To disambiguate new test instances, these translations into the four other languages are estimated using the classifiers built in the first version of the system (l1). The third system run (*mrf*) builds a Markov network of L1 classifiers in order to find the best translation into all five target languages jointly. The nodes of this network correspond to the distribution produced by the L1 classifiers, while the edges contain pairwise potentials derived from the joint probabilities of translation labels occurring together in the training data.

Another classification-based approach is presented by the WSD2 system (van Gompel and van den Bosch, 2013), that uses a k-NN classifier to solve the CLWSD task. The first configuration of the system (*c1l*) uses local context features for a window of three words containing the focus word. Parameters were optimized on the trial data. The second flavor of the system (*c1lN*) uses the same configuration of the system, but without parameter optimization. The third configuration of the system (*var*) is heavily optimized on the trial data, selecting the winning configuration per trial word and evaluation metric. In addition to the local context features, also global bag-of-word context features are considered for this version of the system.

A completely different approach is taken by the NRC-SMT system (Carpuat, 2013), that uses a statistical machine translation approach to tackle the CLWSD task. The baseline version of the system (*SMTbasic*) represents a standard phrase-based SMT baseline, that is trained only on the intersected Europarl corpus. Translations for the test instances are extracted from the top hypothesis (for the *best* evaluation) or from the 100-best list (for the *Out-of-five* evaluation). The optimized version of the system (*SMTadapt2*) is trained on the Europarl corpus and additional news data, and uses mixture models that are developed for domain adaptation in SMT.

In addition to the five systems that participated to the official evaluation campaign, the organizers also present results for their ParaSense system, which is described in the following section.

3.2 ParaSense system

The ParaSense system (Lefever et al., 2013) is a multilingual classification-based approach to CLWSD. A combination of both local context information and translational evidence is used to discriminate between different senses of the word, the underlying hypothesis being that using multilingual information should be more informative than only having access to monolingual or bilingual features. The local context features contain the word form, lemma, part-of-speech and chunk information for a window of seven words containing the ambiguous focus word. In addition, a set of bag-of-words features is extracted from the aligned translations that are not the target language of the classifier. Per ambiguous focus word, a list of all content words (nouns, adjectives, adverbs and verbs) that occurred in the linguistically preprocessed aligned translations of the English sentences containing this word, were extracted. Each content word then corresponds to exactly one binary feature per language. For the construction of the translation features for the training set, we used the Europarl aligned translations. As we do not dispose of similar aligned translations for the test instances for which we only have the English test sentences at our disposal, we used the Google Translate API[7] to automatically generate translations for all English test instances in the five target languages.

As a classifier, we opted for the k Nearest neighbor method as implemented in TIMBL (Daelemans and van den Bosch, 2005). As most classifiers can be initialized with a wide range of parameters, we used a genetic algorithm to optimize the parameter settings for our classification task.

4 Results

4.1 Experimental set up

Test set The lexical sample contains 50 English sentences per ambiguous focus word. All instances were manually annotated per language, which resulted in a set of gold standard translation labels per instance. For the construction of the test dataset, we refer to Section 2.

[7] http://code.google.com/apis/language/

Evaluation metric The BEST precision and recall metric was introduced by (McCarthy and Navigli, 2007) in the framework of the SemEval-2007 competition. The metric takes into account the frequency weights of the gold standard translations: translations that were picked by different annotators received a higher associated frequency which is incorporated in the formulas for calculating precision and recall. For the BEST precision and recall evaluation, the system can propose as many guesses as the system believes are correct, but the resulting score is divided by the number of guesses. In this way, systems that output many guesses are not favored and systems can maximize their score by guessing the most frequent translation from the annotators. We also calculate Mode precision and recall, where precision and recall are calculated against the translation that is preferred by the majority of annotators, provided that one translation is more frequent than the others.

The following variables are used for the BEST precision and recall formulas. Let H be the set of annotators, T the set of test words and h_i the set of translations for an item $i \in T$ for annotator $h \in H$. Let A be the set of words from T where the system provides at least one answer and a_i the set of guesses from the system for word $i \in A$. For each i, we calculate the multiset union (H_i) for all h_i for all $h \in H$ and for each unique type (res) in H_i that has an associated frequency ($freq_{res}$). Equation 1 lists the BEST precision formula, whereas Equation 2 lists the formula for calculating the BEST recall score:

$$Precision = \frac{\sum_{a_i : i \in A} \frac{\frac{\sum_{res \in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|A|} \quad (1)$$

$$Recall = \frac{\sum_{a_i : i \in T} \frac{\frac{\sum_{res \in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|T|} \quad (2)$$

Most Frequent translation baseline As a baseline, we selected the most frequent lemmatized translation that resulted from the automated word alignment (GIZA++) for all ambiguous nouns in the training data. This baseline is inspired by the most frequent sense baseline often used in WSD evaluations. The main difference between the most frequent sense baseline and our baseline is that the latter is corpus-dependent: we do not take into account the overall frequency of a word as it would be measured based on a large general purpose corpus, but calculate the most frequent sense (or translation in this case) based on our training corpus.

4.2 Experimental results

For the system evaluation results, we show precision and Mode precision figures for both evaluation types (*best* and *Out-of-five*). In our case, precision refers to the number of correct translations in relation to the total number of translations generated by the system, while recall refers to the number of correct translations generated by the classifier. As all participating systems predict a translation label for all sentences in the test set, precision and recall will give identical results. As a consequence, we do not list the recall and Mode recall figures that are in this case identical to the corresponding precision scores.

Table 1 lists the averaged *best* precision scores for all systems, while Table 2 gives an overview of the *best* Mode precision figures for all five target languages, viz. Spanish (Es), Dutch (Nl), German (De), Italian (It) and French (Fr). We list scores for all participating systems in the official CLWSD evaluation campaign, as well as for the organizers' system *ParaSense*, that is not part of the official SemEval competition. The best results for the *best* precision evaluation are achieved by the NRC-SMTadapt2 system for Spanish and by the WSD2 system for the other four target languages, closely followed by the HLTDI system. The latter two systems also obtain the best results for the *best* Mode precision metric.

Table 3 lists the averaged *Out-of-five* precision scores for all systems, while Table 4 gives an overview of the *Out-of-five* Mode precision figures for all five target languages, viz. Spanish (Es), Dutch (Nl), German (De), Italian (It) and French (Fr). For the *Out-of-five* evaluation, where systems are allowed to generate up to five unique translations without being penalized for wrong translations, again the HLTDI and WSD2 systems obtain the best classification performance.

Although the winning systems use different approaches (statistical machine translation and classi-

fication algorithms), they have in common that they only use a parallel corpus to extract disambiguating information, and do not use external resources such as WordNet. As a consequence, this makes the systems very flexible and language-independent. The ParaSense system, that incorporates translation information from four other languages, outperforms all other systems, except for the *best* precision metric in Spanish, where the NRC-SMT system obtains the overall best results. This confirms the hypothesis that a truly multilingual approach to WSD, which incorporates translation information from multiple languages into the feature vector, is more effective than only using monolingual or bilingual features. A possible explanation could be that the differences between the different languages that are integrated in the feature vector enable the system to refine the obtained sense distinctions. We indeed see that the ParaSense system outperforms the classification-based bilingual approaches which exploit similar information (e.g. training corpora and machine learning algorithms).

	Es	Nl	De	It	Fr
Baseline					
	23.23	20.66	17.43	20.21	25.74
results for the HLTDI **system**					
hltdi-l1	29.01	21.53	19.50	24.52	27.01
hltdi-l2	28.49	22.36	19.92	23.94	28.23
hltdi-mrf	29.36	21.61	19.76	24.62	27.46
results for the XLING **system**					
merged	11.09	4.91	4.08	6.93	9.57
snt	19.59	9.89	8.13	12.74	17.33
tnt	18.60	7.40	5.29	10.70	16.48
results for the LIMSI **system**					
limsi	24.70			21.20	24.56
results for the NRC-SMT **system**					
basic	27.24				
adapt2	**32.16**				
results for the WSD2 **system**					
c1l	28.40	23.14	20.70	25.43	29.88
c1lN	28.65	**23.61**	**20.82**	**25.66**	**30.11**
var	23.31	17.17	16.20	20.38	25.89
results for the PARASENSE **system**					
	31.72	25.29	24.54	28.15	31.21

Table 1: BEST precision scores averaged over all twenty test words for Spanish (Es), Dutch (Nl), German (De), Italian (It) and French (Fr).

	Es	Nl	De	It	Fr
Baseline					
	27.48	24.15	15.30	19.88	20.19
results for the HLTDI **system**					
hltdi-l1	36.32	25.39	24.16	26.52	21.24
hltdi-l2	**37.11**	25.34	**24.74**	26.65	21.07
hltdi-mrf	36.57	25.72	24.01	26.26	21.24
results for the XLING **system**					
merged	24.31	8.54	5.82	7.54	11.63
snt	21.36	9.56	10.36	11.27	11.57
tnt	24.31	8.54	5.82	7.54	11.63
results for the LIMSI **system**					
limsi	32.09			23.06	22.16
results for the NRC-SMT **system**					
basic	32.28				
adapt2	36.2				
results for the WSD2 **system**					
c1l	33.89	26.32	24.73	**31.61**	**26.62**
c1lN	33.70	**27.96**	24.27	30.67	25.27
var	27.98	18.74	21.74	20.69	16.71
results for the PARASENSE **system**					
	40.26	30.29	25.48	30.11	26.33

Table 2: BEST Mode precision scores averaged over all twenty test words for Spanish (Es), Dutch (Nl), German (De), Italian (It) and French (Fr).

	Es	Nl	De	It	Fr
Baseline					
	53.07	43.59	38.86	42.63	51.36
results for the HLTDI **system**					
hltdi-l1	**61.69**	46.55	**43.66**	**53.57**	57.76
hltdi-l2	59.51	46.36	42.32	53.05	58.20
hltdi-mrf	9.89	5.69	4.15	3.91	7.11
results for the XLING **system**					
merged	43.76	24.30	19.83	33.95	38.15
snt	44.83	27.11	23.71	32.38	38.44
tnt	39.52	23.27	19.13	33.28	35.30
results for the LIMSI **system**					
limsi	49.01			40.25	45.37
results for the NRC-SMT **system**					
basic	37.98				
adapt2	41.65				
results for the WSD2 **system**					
c1l	58.23	**47.83**	43.17	52.22	59.07
c1lN	57.62	47.62	43.24	52.73	**59.80**
var	55.70	46.85	41.46	51.18	59.19

Table 3: OUT-OF-FIVE precision scores averaged over all twenty test words for Spanish (Es), Dutch (Nl), German (De), Italian (It) and French (Fr).

	Es	Nl	De	It	Fr
Baseline					
	57.35	41.97	44.35	41.69	47.42
results for the HLTDI **system**					
hltdi-l1	**64.65**	**47.34**	53.50	**56.61**	51.96
hltdi-l2	62.52	44.06	49.03	54.06	53.57
hltdi-mrf	11.39	5.09	3.14	3.87	7.79
results for the XLING **system**					
merged	48.63	23.64	24.64	31.74	30.11
snt	50.04	27.30	30.57	29.17	32.45
tnt	44.96	22.98	23.54	29.61	28.02
results for the LIMSI **system**					
limsi	51.41			47.21	39.54
results for the NRC-SMT **system**					
basic	42.92				
adapt2	45.38				
results for the WSD2 **system**					
c1l	63.75	45.27	50.11	54.13	**57.57**
c1lN	63.80	44.53	**50.26**	54.37	56.40
var	61.51	41.82	49.23	54.73	54.97

Table 4: OUT-OF-FIVE Mode precision scores averaged over all twenty test words for Spanish (Es), Dutch (Nl), German (De), Italian (It) and French (Fr).

In general, we notice that French and Spanish have the highest scores, while Dutch and German seem harder to tackle. Italian is situated somewhere in between the Romance and Germanic languages. This trend confirms the results that were obtained during the first SemEval Cross-lingual WSD task (Lefever and Hoste, 2010b). As pointed out after the first competition, the discrepancy between the scores for the Romance and Germanic languages can probably be explained by the number of classes (or translations in this case) the systems have to choose from. Germanic languages are typically characterized by a very productive compounding system, where compounds are joined together in one orthographic unit, which results in a much higher number of different class labels. As the Romance languages typically write compounds in separate orthographic units, they dispose of a smaller number of different translations for each ambiguous noun.

We can also notice large differences between the scores for the individual words. Figure 1 illustrates this by showing the *best* precision scores in Spanish for the different test words for the best run per system. Except for some exceptions (e.g. *coach* in the NRC-SMT system), most system performance scores follow a similar curve. Some words (e.g. *match, range*) are particularly hard to disambiguate, while others obtain very high scores (e.g. *mission, soil*). One possible explanation for the very good scores for some words (e.g. *soil*) can be attributed to a very generic translation which accounts for all senses of the word even though there might be more suitable translations for each of the senses depending on the context. Because the manual annotators were able to select three good translations for each test instance, the most generic translation is often part of the gold standard translations. This is also reflected in the high baseline scores for these words. For the words performing badly in most systems, an inspection of the training data properties revealed two possible explanations for these poor classification results. Firstly, there seems to be a link with the number of training instances, corresponding to the frequency of the word in the training corpus. Both for *coach* and *match*, two words consistently performing bad in all systems, there are very few training examples in the corpus (66 and 109 respectively). This could also explain why the NRC-SMT system, that also uses additional parallel data, achieves better results for *coach* than all other systems. Secondly, the ambiguity or number of valid translations per word in the training data also seems to play a role in the classification results. Both *job* and *range* appear very hard to classify correctly, and both words are very ambiguous, with no fewer than 121 and 125 translations, respectively, to choose from in Spanish.

5 Conclusion

The Cross-lingual Word Sense Disambiguation task attempts to address three important challenges for WSD, namely (1) the data acquisition bottleneck, which is caused by the lack of manually created resources, (2) the sense granularity and subjectivity problem of the existing sense inventories and (3) the need to make WSD more suited for practical applications. The task contributes to the WSD research domain by the construction of a dedicated benchmark data set that allows to compare different approaches to the Cross-lingual WSD task.

The evaluation results lead to the following observations. Firstly, languages which make exten-

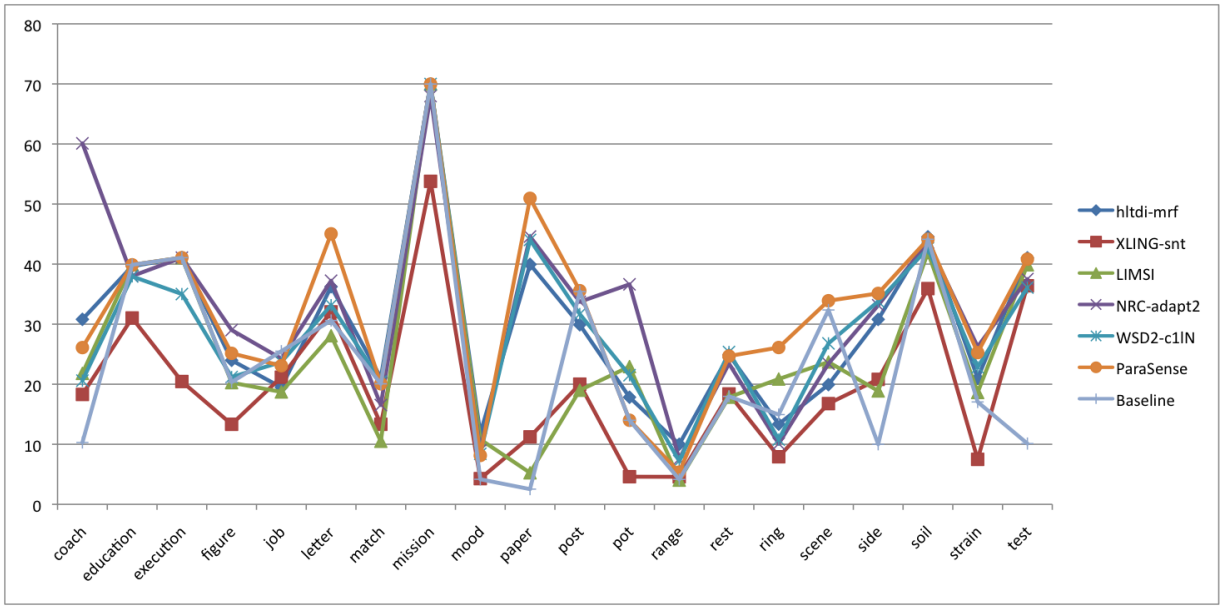

Figure 1: Spanish *best* precision scores for all systems per ambiguous focus word.

sive use of single word compounds seem harder to tackle, which can probably be explained by the higher number of translations these classifiers have to choose from. Secondly, we can notice large differences between the performances of the individual test words. For the words that appear harder to disambiguate, both the number of training instances as well as the ambiguity of the word seem to play a role for the classification performance. Thirdly, both the ParaSense system as well as the two winning systems from the competition extract all disambiguating information from the parallel corpus and do not use any external resources. As a result, these systems are very flexible and can be easily extended to other languages and domains. In addition, the good scores of the ParaSense system, that incorporates information from four additional languages, confirms the hypothesis that a truly multilingual approach is an effective way to tackle the CLWSD task.

Acknowledgments

We would like to thank all annotators for their hard work.

References

Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation. Algorithms and Applications*. Text, Speech and Language Technology. Springer.

M. 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)*, pages 77–85, Athens, Greece.

Marianna Apidianaki. 2013. LIMSI : Cross-lingual Word Sense Disambiguation using Translation Sense Clustering. 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)*, Atlanta, USA.

P.F. Brown, S.A.D. Pietra, V.J.D. Pietra, and R.L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270, Berkeley, California.

M. Carpuat and D. Wu. 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)*, pages 61–72, Prague, Czech Republic.

Marine Carpuat. 2013. NRC: A Machine Translation Approach to Cross-Lingual Word Sense Disambiguation (SemEval-2013 Task 10). 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)*, Atlanta, USA.

Y.S. Chan and H.T. Ng. 2005. Scaling Up Word Sense Disambiguation via Parallel Texts. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, pages 1037–1042, Pittsburgh, Pennsylvania, USA.

P. Clough and M. Stevenson. 2004. Cross-language information retrieval using eurowordnet and word sense disambiguation. In *Advances in Information Retrieval, 26th European Conference on IR Research (ECIR)*, pages 327–337, Sunderland, UK.

W. Daelemans and A. van den Bosch. 2005. *Memory-based Language Processing*. Cambridge University Press.

M. Diab. 2004. *Word Sense Disambiguation within a Multilingual Framework*. Phd, University of Maryland, USA.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W.A. Gale and K.W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

E. Lefever and V. Hoste. 2010a. Construction of a Benchmark Data Set for Cross-Lingual Word Sense Disambiguation. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

E. Lefever and V. Hoste. 2010b. SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 15–20, Uppsala, Sweden.

E. Lefever, V. Hoste, and M. De Cock. 2013. Five languages are better than one: an attempt to bypass the data acquisition bottleneck for wsd. In *In Alexander Gelbukh (ed.), CICLing 2013, Part I, LNCS 7816*, pages 343–354. Springer-Verlag Berlin Heidelberg.

D. McCarthy and R. Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

R. Navigli. 2009. Word Sense Disambiguation: a Survey. *ACM Computing Surveys*, 41(2):1–69.

H.T. Ng, B. Wang, and Y.S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 455–462, Sapporo, Japan.

Alex Rudnick, Can Liu, and Michael Gasser. 2013. HLTDI: CL-WSD Using Markov Random Fields for SemEval-2013 Task 10. 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)*, Atlanta, USA.

L. Specia, M.G.V. Nunes, and M. Stevenson. 2007. Learning Expressive Models for Word Sense Disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 41–48, Prague, Czech Republic.

Liling Tan and Francis Bond. 2013. XLING: Matching Query Sentences to a Parallel Corpus using Topic Models for WSD. 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)*, Atlanta, USA.

D. Tufiş, R. Ion, and N. Ide. 2004. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1312–1318, Geneva, Switzerland, August. Association for Computational Linguistics.

Maarten van Gompel and Antal van den Bosch. 2013. Parameter optimisation for Memory-based 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)*, Atlanta, USA.

P. Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.