SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation

Els Lefever[1,2] **and Veronique Hoste**[1,2]

[1]LT3, Language and Translation Technology Team, University College Ghent, Belgium
[2]Department of Applied Mathematics and Computer Science, Ghent University, Belgium

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

Abstract

The goal of this task is to evaluate the feasibility of multilingual WSD on a newly developed multilingual lexical sample data set. Participants were asked to automatically determine the contextually appropriate translation of a given English noun in five languages, viz. Dutch, German, Italian, Spanish and French. This paper reports on the sixteen submissions from the five different participating teams.

1 Introduction

Word Sense Disambiguation, the task of selecting the correct sense of an ambiguous word in a given context, is a well-researched NLP problem (see for example Agirre and Edmonds (2006) and Navigli (2009)), largely boosted by the various Senseval and SemEval editions. The SemEval-2010 Cross-lingual Word Sense Disambiguation task focuses on two bottlenecks in current WSD research, namely the scarcity of sense inventories and sense-tagged corpora (especially for languages other than English) and the growing tendency to evaluate the performance of WSD systems in a real application such as machine translation and cross-language information retrieval (see for example Agirre et al. (2007)).

The Cross-lingual WSD task aims at the development of a multilingual data set to test the feasibility of multilingual WSD. Many studies have already shown the validity of this cross-lingual evidence idea (Gale et al., 1993; Ide et al., 2002; Ng et al., 2003; Apidianaki, 2009), but until now no benchmark data sets have been available. For the SemEval-2010 competition we developed (i) a sense inventory in which the sense distinctions were extracted from the multilingual corpus Europarl[1] and (ii) a data set in which the ambiguous words were annotated with the senses from the multilingual sense inventory. The Cross-Lingual WSD task is a lexical sample task for English nouns, in which the word senses are made up of the translations in five languages, viz. Dutch, French, Italian, Spanish and German. Both the sense inventory and the annotated data set were constructed for a sample of 25 nouns. The data set was divided into a trial set of 5 ambiguous nouns and a test set of 20 nouns. The participants had to automatically determine the contextually appropriate translation for a given English noun in each or a subset of the five target languages. Only translations present in Europarl were considered as valid translations.

The remainder of this article is organized as follows. Section 2 focuses on the task description and gives a short overview of the construction of the sense inventory and the annotation of the benchmark data set with the senses from the multilingual sense inventory. Section 3 clarifies the scoring metrics and presents two frequency-based baselines. The participating systems are presented in Section 4, while the results of the task are discussed in Section 5. Section 6 concludes this paper.

2 Task setup

2.1 Data sets

Two types of data sets were used in the Cross-lingual WSD task: (a) a parallel corpus on the basis of which the gold standard sense inventory was created and (b) a collection of English sentences containing the lexical sample words annotated with their contextually appropriate translations in five languages.

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

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 15–20,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

Below, we provide a short summary of the complete data construction process. For a more detailed description, we refer to Lefever and Hoste (2009; 2010).

The gold standard sense inventory was derived from the Europarl parallel corpus[2], which is extracted from the proceedings of the European Parliament (Koehn, 2005). We selected 6 languages from the 11 European languages represented in the corpus, viz. English (our target language), Dutch, French, German, Italian and Spanish. All data were already sentence-aligned using a tool based on the Gale and Church (1991) algorithm, which was part of the Europarl corpus. We only considered the 1-1 sentence alignments between English and the five other languages. These sentence alignments were made available to the task participants for the five trial words. The sense inventory extracted from the parallel data set (Section 2.2) was used to annotate the sentences in the trial set and the test set, which were extracted from the JRC-ACQUIS Multilingual Parallel Corpus[3] and BNC[4].

2.2 Creation of the sense inventory

Two steps were taken to obtain a multilingual sense inventory: (1) word alignment on the sentences to find the set of possible translations for the set of ambiguous nouns and (2) clustering by meaning (per target word) of the resulting translations.

GIZA++ (Och and Ney, 2003) was used to generate the initial word alignments, which were manually verified by certified translators in all six involved languages. The human annotators were asked to assign a "NULL" link to words for which no valid translation could be identified. Furthermore, they were also asked to provide extra information on compound translations (e.g. the Dutch word *Investeringsbank* as a translation of the English multiword *Investment Bank*), fuzzy links, or target words with a different PoS (e.g. the verb *to bank*).

The manually verified translations were clustered by meaning by one annotator. In order to do so, the translations were linked across languages on the basis of unique sentence IDs. After the selection of all unique translation combinations, the translations were grouped into clusters. The clusters were organized in two levels, in which the top level reflects the main sense categories (e.g. for the word *coach* we have (1) (sports) manager, (2) bus, (3) carriage and (4) part of a train), and the subclusters represent the finer sense distinctions. Translations that correspond to English multiword units were identified and in case of non-apparent compounds, i.e. compounds which are not marked with a "-", the different compound parts were separated by §§ in the clustering file (e.g. the German *Post§§kutsche*). All clustered translations were also manually lemmatized.

2.3 Sense annotation of the test data

The resulting sense inventory was used to annotate the sentences in the trial set (20 sentences per ambiguous word) and the test set (50 sentences per ambiguous word). In total, 1100 sentences were annotated. The annotators were asked to (a) pick the contextually appropriate sense cluster and to (b) choose their three preferred translations from this cluster. In case they were not able to find three appropriate translations, they were also allowed to provide fewer. These potentially different translations were used to assign frequency weights (shown in example (2)) to the gold standard translations per sentence. The example (1) below shows the annotation result in both German and Dutch for an English source sentence containing *coach*.

(1) SENTENCE 12. STRANGELY , the national coach of the Irish teams down the years has had little direct contact with the four provincial coaches .

German 1: Nationaltrainer
German 2: Trainer
German 3: Coach

Dutch 1: trainer
Dutch 2: coach
Dutch 3: voetbaltrainer

For each instance, the gold standard that results from the manual annotation contains a set of translations that are enriched with

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

[3] http://wt.jrc.it/lt/Acquis/

[4] http://www.natcorp.ox.ac.uk/

frequency information. The format of both the input file and gold standard is similar to the format that will be used for the SemEval Cross-Lingual Lexical Substitution task (Sinha and Mihalcea, 2009). The following example illustrates the six-language gold standard format for the trial sentence in (1). The first field contains the target word, PoS-tag and language code, the second field contains the sentence ID and the third field contains the gold standard translations in the target language, enriched with their frequency weight:

(2) coach.n.nl 12 :: coach 3; speler-trainer 1; trainer 3; voetbaltrainer 1;
coach.n.fr 12 :: capitaine 1; entraîneur 3;
coach.n.de 12 :: Coach 1; Fußbaltrainer 1; Nationaltrainer 2; Trainer 3;
coach.n.it 12 :: allenatore 3;
coach.n.es 12 :: entrenador 3;

3 Evaluation

3.1 Scoring

To score the participating systems, we use an evaluation scheme which is inspired by the English lexical substitution task in SemEval 2007 (McCarthy and Navigli, 2007). We perform both a *best result* evaluation and a more relaxed evaluation for the *top five results.* The evaluation is performed using precision and recall ($Prec$ and Rec in the equations below), and Mode precision (M_P) and Mode recall (M_R), where we calculate precision and recall against the translation that is preferred by the majority of annotators, provided that one translation is more frequent than the others.

For the precision and recall formula we use the following variables. Let H be the set of annotators, T the set of test items and h_i the set of responses for an item $i \in T$ for annotator $h \in H$. For each $i \in T$ we calculate the mode (m_i) which corresponds to the translation with the highest frequency weight. For a detailed overview of the M_P and M_R calculations, we refer to McCarthy and Navigli (2007). Let A be the set of items from T (and TM) where the system provides at least one answer and $a_i : i \in A$ the set of guesses from the system for item i. 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 ($freq_{res}$). In order to assign frequency weights to our gold standard translations, we asked our human annotators to indicate their top 3 translations, which enables us to also obtain meaningful associated frequencies ($freq_{res}$) viz. “1” in case a translation is picked by 1 annotator, “2” if picked by two annotators and “3” if chosen by all three annotators.

Best result evaluation For the *best result* evaluation, systems 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 a lot of guesses are not favoured.

$$Prec = \frac{\sum_{a_i:i\in A} \frac{\frac{\sum_{res\in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|A|} \tag{1}$$

$$Rec = \frac{\sum_{a_i:i\in T} \frac{\frac{\sum_{res\in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|T|} \tag{2}$$

Out-of-five (Oof) evaluation For the more relaxed evaluation, systems can propose up to five guesses. For this evaluation, the resulting score is not divided by the number of guesses.

$$Prec = \frac{\sum_{a_i:i\in A} \frac{\sum_{res\in a_i} freq_{res}}{|H_i|}}{|A|} \tag{3}$$

$$Rec = \frac{\sum_{a_i:i\in T} \frac{\sum_{res\in a_i} freq_{res}}{|H_i|}}{|T|} \tag{4}$$

3.2 Baselines

We produced two frequency-based baselines:

1. For the *Best result* evaluation, we select the most frequent lemmatized translation that results from the automated word alignment process (GIZA++).
2. For the *Out-of-five* or *more relaxed* evaluation, we select the five most frequent (lemmatized) translations that result from the GIZA++ alignment.

Table 1 shows the baselines for the *Best* evaluation, while Table 2 gives an overview per language of the baselines for the *Out-of-five* evaluation.

	Prec	Rec	M_P	M_R
Spanish	18.36	18.36	23.38	23.38
French	20.71	20.71	15.21	15.21
Italian	14.03	14.03	11.23	11.23
Dutch	15.69	15.69	8.71	8.71
German	13.16	13.16	6.95	6.95

Table 1: *Best* Baselines

	Prec	Rec	M_P	M_R
Spanish	48.41	48.41	42.62	42.62
French	45.99	45.99	36.45	36.45
Italian	34.51	34.51	29.70	29.70
Dutch	37.43	37.43	24.58	24.58
German	32.89	32.89	29.80	29.80

Table 2: *Out-of-five* Baselines

4 Systems

We received sixteen submissions from five different participating teams. One group tackled all five target languages, whereas the other groups focused on four (one team), two (one team) or one (two teams) target language(s). For both the *best* and the *Out-of-five* evaluation tasks, there were between three and seven participating systems per language.

The OWNS system identifies the nearest neighbors of the test instances from the training data using a pairwise similarity measure (weighted sum of the word overlap and semantic overlap between two sentences). They use WordNet similarity measures as an additional information source, while the other teams merely rely on parallel corpora to extract all lexical information. The UvT-WSD systems use a k-nearest neighbour classifier in the form of one word expert per lemma–Part-of-Speech pair to be disambiguated. The classifier takes as input a variety of local and global context features. Both the FCC-WSD and T3-COLEUR systems use bilingual translation probability tables that are derived from the Europarl corpus. The FCC-WSD system uses a Naive Bayes classifier, while the T3-COLEUR system uses an unsupervised graph-based method. Finally, the UHD systems build for each target word a multilingual co-occurrence graph based on the target word's aligned contexts found in parallel corpora. The cross-lingual nodes are first linked by translation edges, that are labeled with the translations of the target word in the corresponding contexts. The graph is transformed into a minimum spanning tree which is used to select the most relevant words in context to disambiguate a given test instance.

5 Results

For the system evaluation results, we show precision (*Prec*), recall (*Rec*), Mode precision (M_P) and Mode recall (M_R). We ranked all system results according to recall, as was done for the Lexical Substitution task. Table 3 shows the system ranking on the *best* task, while Table 4 shows the results for the *Oof* task.

	Prec	Rec	M_P	M_R
Spanish				
UvT-v	23.42	24.98	24.98	24.98
UvT-g	19.92	19.92	24.17	24.17
T3-COLEUR	19.78	19.59	24.59	24.59
UHD-1	20.48	16.33	28.48	22.19
UHD-2	20.2	16.09	28.18	22.65
FCC-WSD1	15.09	15.09	14.31	14.31
FCC-WSD3	14.43	14.43	13.41	13.41
French				
T3-COLEUR	21.96	21.73	16.15	15.93
UHD-2	20.93	16.65	17.78	14.15
UHD-1	20.22	16.21	17.59	14.56
OWNS2	16.05	16.05	14.21	14.21
OWNS1	16.05	16.05	14.21	14.21
OWNS3	12.53	12.53	14.21	14.21
OWNS4	10.49	10.49	14.21	14.21
Italian				
T3-COLEUR	15.55	15.4	10.2	10.12
UHD-2	16.28	13.03	14.89	9.46
UHD-1	15.94	12.78	12.34	8.48
Dutch				
UvT-v	17.7	17.7	12.05	12.05
UvT-g	15.93	15.93	10.54	10.54
T3-COLEUR	10.71	10.56	6.18	6.16
German				
T3-COLEUR	13.79	13.63	8.1	8.1
UHD-1	12.2	9.32	11.05	7.78
UHD-2	12.03	9.23	12.91	9.22

Table 3: *Best* System Results

Beating the baseline seems to be quite challenging for this WSD task. While the best systems outperform the baseline for the *best* task,

	Prec	Rec	M_P	M_R
Spanish				
UvT-g	43.12	43.12	43.94	43.94
UvT-v	42.17	42.17	40.62	40.62
FCC-WSD2	40.76	40.76	44.84	44.84
FCC-WSD4	38.46	38.46	39.49	39.49
T3-COLEUR	35.84	35.46	39.01	38.78
UHD-1	38.78	31.81	40.68	32.38
UHD-2	37.74	31.3	39.09	32.05
French				
T3-COLEUR	49.44	48.96	42.13	41.77
OWNS1	43.11	43.11	38.29	38.29
OWNS2	38.74	38.74	37.73	37.73
UHD-1	39.06	32	37.00	26.79
UHD-2	37.92	31.38	37.66	27.08
Italian				
T3-COLEUR	40.7	40.34	38.99	38.70
UHD-1	33.72	27.49	27.54	21.81
UHD-2	32.68	27.42	29.82	23.20
Dutch				
UvT-v	34.95	34.95	24.62	24.62
UvT-g	34.92	34.92	19.72	19.72
T3-COLEUR	21.47	21.27	12.05	12.03
German				
T3-COLEUR	33.21	32.82	33.60	33.56
UHD-1	27.62	22.82	25.68	21.16
UHD-2	27.24	22.55	27.19	22.30

Table 4: *Out-of-five* System Results

this is not always the case for the *Out-of-five* task. This is not surprising though, as the *Oof* baseline contains the five most frequent Europarl translations. As a consequence, these translations usually contain the most frequent translations from different sense clusters, and in addition they also contain the most generic translation that often covers multiple senses of the target word.

The best results are achieved by the UvT-WSD (Spanish, Dutch) and ColEur (French, Italian and German) systems. An interesting feature that these systems have in common, is that they extract all lexical information from the parallel corpus at hand, and do not need any additional data sources. As a consequence, the systems can easily be applied to other languages as well. This is clearly illustrated by the ColEur system, that participated for all supported languages, and outperformed the other systems for three of the five languages.

In general, we notice that Spanish and French have the highest scores, followed by Italian, whereas Dutch and German seem to be more challenging. The same observation can be made for both the *Oof* and *Best* results, except for Italian that performs worse than Dutch for the latter. However, given the low participation rate for Italian, we do not have sufficient information to explain this different behaviour on the two tasks. The discrepancy between the performance figures for Spanish and French on the one hand, and German and Dutch on the other hand, seems more readily explicable. A likely explanation could be the number of classes (or translations) the systems have to choose from. As both Dutch and German are characterized by a rich compounding system, these compound translations also result in a higher number of different translations. Figure 1 illustrates this by listing the number of different translations (or classes in the context of WSD) for all trial and test words. As a result, the broader set of translations makes the WSD task, that consists in choosing the most appropriate translation from all possible translations for a given instance, more complicated for Dutch and German.

6 Concluding remarks

We believe that the Cross-lingual Word Sense Disambiguation task is an interesting contribution to the domain, as it attempts to address two WSD problems which have received a lot of attention lately, namely (1) the scarcity of hand-crafted sense inventories and sense-tagged corpora and (2) the need to make WSD more suited for practical applications.

The system results lead to the following observations. Firstly, languages which make extensive use of single word compounds seem harder to tackle, which is also reflected in the baseline scores. A possible explanation for this phenomenon could lie in the number of translations the systems have to choose from. Secondly, it is striking that the systems with the highest performance solely rely on parallel corpora as a source of information. This would seem very promising for future multilingual WSD research; by eliminating the need

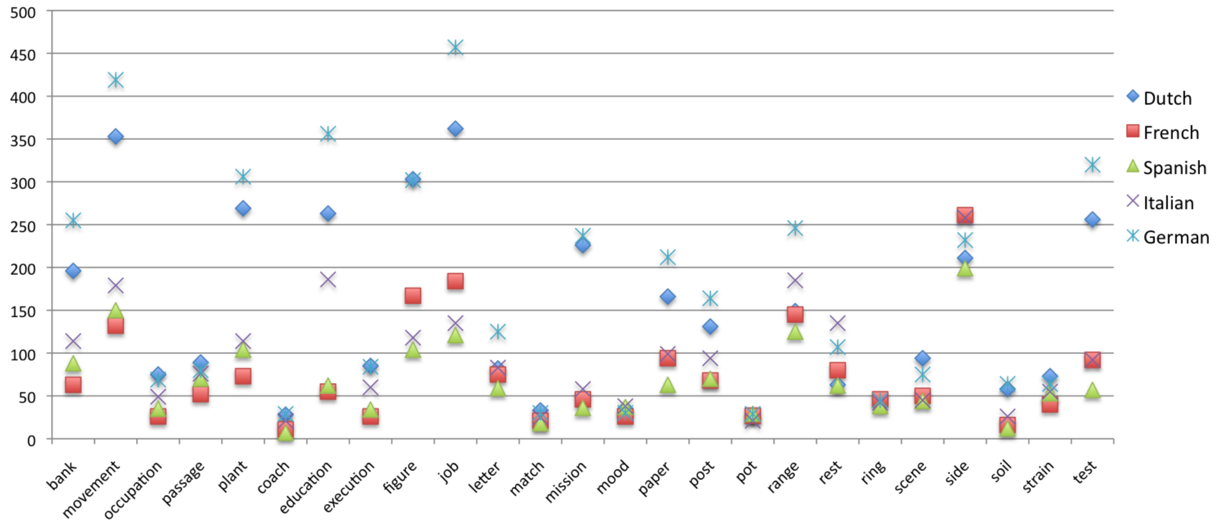

Figure 1: Number of different translations per word for Dutch, French, Spanish, Italian and German.

for external information sources, these systems present a more flexible and language-independent approach to WSD.

References

E. Agirre and P. Edmonds, editors. 2006. *Word Sense Disambiguation.* Text, Speech and Language Technology. Springer, Dordrecht.

E. Agirre, B. Magnini, O. Lopez de Lacalle, A. Otegi, G. Rigau, and P. Vossen. 2007. Semeval-2007 task01: Evaluating wsd on cross-language information retrieval. In *Proceedings of CLEF 2007 Workshop, pp. 908 - 917. ISSN: 1818-8044. ISBN: 2-912335-31-0.*

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)*, Athens, Greece.

W.A. Gale and K.W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Computational Linguistics*, pages 177–184.

W.A. Gale, K.W. Church, and D. Yarowsky. 1993. A method for disambiguating word senses in a large corpus. In *Computers and the Humanities*, volume 26, pages 415–439.

N. Ide, T. Erjavec, and D. Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit.*

E. Lefever and V. Hoste. 2009. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the NAACL-HLT 2009 Workshop: SEW-2009 - Semantic Evaluations*, pages 82–87, Boulder, Colorado.

E. Lefever and V. Hoste. 2010. Construction of a benchmark data set for cross-lingual word sense disambiguation. In *Proceedings of the seventh international conference on Language Resources and Evaluation.*, Malta.

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. In *ACM Computing Surveys*, volume 41, pages 1–69.

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

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

McCarthy D. Sinha, R. D. and R. Mihalcea. 2009. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the NAACL-HLT 2009 Workshop: SEW-2009 - Semantic Evaluations*, Boulder, Colorado.