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

Els Lefever\textsuperscript{1,2} and Veronique Hoste\textsuperscript{1,2}

\textsuperscript{1}LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittannielaan 45, 9000 Gent, Belgium
\textsuperscript{2}Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium

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

Abstract

We propose a multilingual unsupervised Word Sense Disambiguation (WSD) task for a sample of English nouns. Instead of providing manually sense-tagged examples for each sense of a polysemous noun, our sense inventory is built up on the basis of the Europarl parallel corpus. The multilingual setup involves the translations of a given English polysemous noun in five supported languages, viz. Dutch, French, German, Spanish and Italian.

The task targets the following goals: (a) the manual creation of a multilingual sense inventory for a lexical sample of English nouns and (b) the evaluation of systems on their ability to disambiguate new occurrences of the selected polysemous nouns. For the creation of the hand-tagged gold standard, all translations of a given polysemous English noun are retrieved in the five languages and clustered by meaning. Systems can participate in 5 bilingual evaluation subtasks (English - Dutch, English - German, etc.) and in a multilingual subtask covering all language pairs.

As WSD from cross-lingual evidence is gaining popularity, we believe it is important to create a multilingual gold standard and run cross-lingual WSD benchmark tests.

1 Introduction

The Word Sense Disambiguation (WSD) task, which consists in selecting the correct sense of a given word in a given context, has been widely studied in computational linguistics. For a recent overview of WSD algorithms, resources and applications, we refer to Agirre and Edmonds (2006) and Navigli (2009). Semantic evaluation competitions such as Senseval\textsuperscript{1} and its successor Semeval revealed that supervised approaches to WSD usually achieve better results than unsupervised methods (Márquez et al., 2006). The former use machine learning techniques to induce a classifier from manually sense-tagged data, where each occurrence of a polysemous word gets assigned a sense label from a predefined sense inventory such as WordNet (Fellbaum, 1998). These supervised methods, however, heavily rely on large sense-tagged corpora which are very time consuming and expensive to build. This phenomenon, well known as the knowledge acquisition bottleneck (Gale et al., 1992), explains the modest use and success of supervised WSD in real applications.

Although WSD has long time been studied as a stand-alone NLP task, there is a growing feeling in the WSD community that WSD should preferably be integrated in real applications such as Machine Translation or multilingual information retrieval (Agirre and Edmonds, 2006). Several studies have demonstrated that for instance Statistical Machine Translation (SMT) benefits from incorporating a dedicated WSD module (Chan et al., 2007; Carpuat and Wu, 2007). Using translations from a corpus instead of human-defined sense labels is one way of facilitating the integration of WSD in multilingual applications. It also implic-
itly deals with the granularity problem as finer sense distinctions are only relevant as far as they are lexicalized in the translations. Furthermore, this type of corpus-based approach is language-independent, which makes it a valid alternative for languages lacking sufficient sense inventories and sense-tagged corpora, although one could argue that the lack of parallel corpora for certain language pairs might be problematic as well. The methodology to deduce word senses from parallel corpora starts from the hypothesis that the different sense distinctions of a polysemous word are often lexicalized cross-linguistically. For instance, if we query the English noun “bill” in the English-Dutch Europarl, the following top four translations are retrieved: “rekening” (Eng.: “invoice”) (198 occurrences), “kosten” (Eng.: “costs”) (100 occ.), “Bill” (96 occ.) and “wetsvoorstel” (Eng.: “piece of legislation”) (77 occ.). If we make the simplifying assumption for our example that (i) these are the only Dutch translations of our focus word and that (ii) all sense distinctions of “bill” are lexicalized in Dutch, we can infer that the English noun “bill” has at most four different senses. These different senses in turn can be grouped in case of synonymy. In the Dutch-French Europarl, for example, both “rekening” and “kosten”, are translated by the French “frais”, which might indicate that both Dutch words are synonymous.

Several WSD studies are based on the idea of cross-lingual evidence. Gale et al. (1993) use a bilingual parallel corpus for the automatic creation of a sense-tagged data set, where target words in the source language are tagged with their translation of the word in the target language. Diab and Resnik (2002) present an unsupervised approach to WSD that exploits translational correspondences in parallel corpora that were artificially created by applying commercial MT systems on a sense-tagged English corpus. Ide et al. (2002) use a multilingual parallel corpus (containing seven languages from four language families) and show that sense distinctions derived from translation equivalents are at least as reliable as those made by human annotators. Moreover, some studies present multilingual WSD systems that attain state-of-the-art performance in all-words disambiguation (Ng et al., 2003). The proposed Cross-lingual Word Sense Disambiguation task differs from earlier work (e.g. Ide et al. (2002)) through its independence from an externally defined sense set.

The remainder of this paper is organized as follows. In Section 2, we present a detailed description of the cross-lingual WSD task. It introduces the parallel corpus we used, informs on the development and test data and discusses the annotation procedure. Section 3 gives an overview of the different scoring strategies that will be applied. Section 4 concludes this paper.

2 Task set up

The cross-lingual Word Sense Disambiguation task involves a lexical sample of English nouns. We propose two subtasks, i.e. systems can either participate in the bilingual evaluation task (in which the answer consists of translations in one language) or in the multilingual evaluation task (in which the answer consists of translations in all five supported languages). Table 1 shows an example of the bilingual sense labels for two test occurrences of the English noun bank in our parallel corpus which will be further described in Section 2.1. Table 2 presents the multilingual sense labels for the same sentences.

... giving fish to people living on the [bank] of the river

<table>
<thead>
<tr>
<th>Language</th>
<th>Sense label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (NL)</td>
<td>oever/dijk</td>
</tr>
<tr>
<td>French (F)</td>
<td>rives/rivage/bord/bords</td>
</tr>
<tr>
<td>German (D)</td>
<td>Ufer</td>
</tr>
<tr>
<td>Italian (I)</td>
<td>riva</td>
</tr>
<tr>
<td>Spanish (ES)</td>
<td>orilla</td>
</tr>
</tbody>
</table>

Table 1: Example of bilingual sense labels for the English noun bank

The [bank] of Scotland ...

<table>
<thead>
<tr>
<th>Language</th>
<th>Sense label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (NL)</td>
<td>bank/krediinstelling</td>
</tr>
<tr>
<td>French (F)</td>
<td>banque/établissement de crédit</td>
</tr>
<tr>
<td>German (D)</td>
<td>Bank/Kreditinstitut</td>
</tr>
<tr>
<td>Italian (I)</td>
<td>banca</td>
</tr>
<tr>
<td>Spanish (ES)</td>
<td>banco</td>
</tr>
</tbody>
</table>

Table 2: Example of multilingual sense labels for the English noun bank
... giving fish to people living on the [bank] of the river

<table>
<thead>
<tr>
<th>Language</th>
<th>Sense label</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL,F,D,I,ES</td>
<td>oever/dijk, rives/rivage/bord/bords, Ufer, riva, orilla</td>
</tr>
</tbody>
</table>

The [bank] of Scotland ...

<table>
<thead>
<tr>
<th>Language</th>
<th>Sense label</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL,F,D,I,ES</td>
<td>bank/kredietinstelling, banque/établissement de crédit, Bank/Kreditinstitut, banca, banco</td>
</tr>
</tbody>
</table>

Table 2: Example of multi-lingual sense labels for the English noun bank

2.1 Corpus and word selection

The document collection which serves as the basis for the gold standard construction and system evaluation is 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: English (our target language), Dutch, French, German, Italian and Spanish. All sentences are aligned using a tool based on the Gale and Church (1991) algorithm. We only consider the 1-1 sentence alignments between English and the five other languages (see also Tufis et al. (2004) for a similar strategy). These 1-1 alignments will be made available to all task participants. Participants are free to use other training corpora, but additional translations which are not present in Europarl will not be included in the sense inventory that is used for evaluation.

For the competition, two data sets will be developed. The development and test sentences will be selected from the JRC-ACQUIS Multilingual Parallel Corpus\(^3\). The development data set contains 5 polysemous nouns, for which we provide the manually built sense inventory based on Europarl and 50 example instances, each annotated with one sense label (cluster that contains all translations that have been grouped together for that particular sense) per target language. The manual construction of the sense inventory will be discussed in Section 2.2. The test data contains 50 instances for 20 nouns from the test data as used in the Cross-Lingual Lexical Substitution Task\(^4\). In this task, annotators and systems are asked to provide as many correct Spanish translations as possible for an English target word. They are not bound to a predefined parallel corpus, but can freely choose the translations from any available resource. Selecting the target words from the set of nouns that will be used for the Lexical Substitution Task should make it easier for systems to participate in both tasks.

2.2 Manual annotation

The sense inventory for the 5 target nouns in the development data and the 20 nouns in the test data is manually built up in three steps.

1. In the first annotation step, the 5 translations of the English word are identified per sentence ID. In order to speed up this identification, GIZA++ (Och and Ney, 2003) is used to generate the initial word alignments for the 5 languages. All word alignments are manually verified.

In this step, we might come across multiword translations, especially in Dutch and German which tend to glue parts of compounds together in one orthographic unit. We decided to keep these translations as such, even if they do not correspond exactly to the English target word. Although we will not remove these compound translations from our sense inventory, we will make sure that the development and test sentences do not contain target words that are part of compound translations.

\(^2\)http://www.statmt.org/europarl/

\(^3\)http://wt.jrc.it/lt/Acquis/

\(^4\)http://lit.csci.unt.edu/index.php/Semeval_2010
of a larger multiword unit, in order not to dis-
advantage systems that do not deal with decom-
pounding.

2. In the second step, three annotators per lan-
guage will cluster the retrieved translations per target language. On the basis of the sentence IDs, the translations in all languages will be au-
tomatically coupled. Only translations above a predefined frequency threshold are considered for inclusion in a cluster. Clustering will hap-
pen in a trilingual setting, i.e. annotators always cluster two target languages simultane-
ously (with English being the constant source language).

After the clustering of the translations, the annotators perform a joint evaluation per lan-
guage in order to reach a consensus clustering for each target language. In case the annota-
tors do not reach a consensus, we apply soft-
clustering for that particular translation, i.e. we assign the translation to two or more different clusters.

3. In a last step, there will be a cross-lingual conflict resolution in which the resulting cluster-
ings are checked cross-lingually by the human annotators.

The resulting sense inventory is used to annotate the sentences in the development set and the test set. This implies that a given target word is annotated with the appropriate sense cluster. This annotation is done by the same native annotators as in steps 2 and 3. The goal is to reach a consensus cluster per sentence. But again, if no consensus is reached, soft-clustering is applied and as a consequence, the correct answer for this particular test instance consists of one of the clusters that were considered for soft-clustering.

The resulting clusters are used by the three native annotators to select their top 3 translations per sentence. These potentially different translations are kept to calculate frequency information for all answer translations (discussed in section 3).

Table 3 shows an example of how the translation clusters for the English noun “paper” could look like in a trilingual setting.

3 System evaluation

As stated before, systems can participate in two tasks, i.e. systems can either participate in one or more bilingual evaluation tasks or they can participate in the multilingual evaluation task incorporating the five supported languages. The evaluation of the multilingual evaluation task is simply the average of the system scores on the five bilingual evaluation tasks.

3.1 Evaluation strategies

For the evaluation of the participating systems we will use an evaluation scheme which is inspired by the English lexical substitution task in SemEval 2007 (McCarthy andNavigli, 2007). The evaluation will be performed using precision and recall ($P$ and $R$ in the equations that follow). We perform both a best result evaluation and a more relaxed evaluation for the top five results.

Let $H$ be the set of annotators, $T$ be the set of test items and $h_i$ be the set of responses for an item $i \in T$ for annotator $h \in H$. Let $A$ be the set of items from $T$ where the system provides at least one answer and $a_i : i \in A$ be 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 the formula of (McCarthy andNavigli, 2007), the associated frequency ($freq_{res}$) is equal to the number of times an item appears in $H_i$. As we define our answer clusters by consensus, this frequency would always be “1”. In order to overcome this, we ask our human annotators to indicate their top 3 translations, which enables us to also obtain meaningful associated frequencies ($freq_{res}$) (“1” in case the translation is not chosen by any annotator, “2” in case a translation is picked by 1 annotator, “3” if picked by two annotators and “4” 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

5The annotators will be selected from the master students at the “University College Ghent – Faculty of Translation” that trains certified translators in all six involved languages.
divided by the number of guesses. In this way, systems that output a lot of guesses are not favoured.

\[
P = \frac{\sum_{i \in A} \sum_{res \in a_i} freq_{res}}{|A|} \quad (1)
\]

\[
R = \frac{\sum_{i \in T} \sum_{res \in a_i} freq_{res}}{|T|} \quad (2)
\]

Relaxed 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.

\[
P = \frac{\sum_{i \in A} \sum_{res \in a_i} freq_{res}}{|H_i|} \quad (3)
\]

\[
R = \frac{\sum_{i \in T} \sum_{res \in a_i} freq_{res}}{|H_i|} \quad (4)
\]

3.2 Baseline

We will produce two, both frequency-based, baselines. The first baseline, which will be used for the best result evaluation, is based on the output of the GIZA++ word alignments on the Europarl corpus and just returns the most frequent translation of a given word. The second baseline outputs the five most frequent translations of a given word according to the GIZA++ word alignments. This baseline will be used for the relaxed evaluation. As a third baseline, we will consider using a baseline based on EuroWordNet\(^6\), which is available in the five target languages.

4 Conclusions

We presented a multilingual unsupervised Word Sense Disambiguation task for a sample of English nouns. The lack of supervision refers to the construction of the sense inventory, that is built up on the basis of translations retrieved from the Europarl corpus in five target languages. Systems can participate in a bilingual or multilingual evaluation and are asked to provide correct translations in one or five target languages for new instances of the selected polysemous target nouns.

References


\(^6\)http://www.illc.uva.nl/EuroWordNet
Table 3: translation clusters for the English noun "paper"