Cross-lingual Word Sense Disambiguation: Documentation on Data and Evaluation

Els Lefever and Veronique Hoste

March 2, 2010

1 Introduction

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.

Organizing this task consists in: (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.

There are two types of scoring:

1. *best*: scoring the best substitutes for an ambiguous target word in context

2. *oof*: scoring the best 5 substitutes for an ambiguous target word in context

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).

2.1 Corpus

The document collection which serves as the basis for the gold standard construction and system evaluation is the Europarl parallel corpus[1], which is ex-

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

tracted from the proceedings of the European Parliament [2]. 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 [1] algorithm. We only consider the 1-1 sentence alignments between English and the five other languages (see also [7] for a similar strategy). These 1-1 alignments will be made available to all task participants[2]. 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.

2.2 Word alignment and clustering

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 the following way:

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++ [5] 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.

   Even for the other languages we sometimes obtain multiword alignments for a single target word (e.g *occupation* is sometimes translated in Spanish as *actividad profesional*). In these cases we have also kept the multiword as such as a valid translation suggestion.

   > SOURCE: This monitoring committee , which is part of OLAF , comprises five independent experts who continue to pursue their normal **occupations** however .
   >
   > SPANISH: Este comité supervisor de la OLAF está formado por cinco expertos independientes pero que realizan su **actividad profesional** normal .
   >
   > DUTCH: Het Comité van toezicht van OLAF bestaat uit onafhankelijke experts , die ook hun normale **beroepsbezigheden** blijven voortzetten .
   >
   > GERMAN: Dieser Überwachungsausschuss des OLAF besteht aus fünf unabhängigen Experten , die aber ihrer normalen **Berufstätigkeit** nachgehen .

[2] http://lt3.hogent.be/semeval

> FRENCH: Ce comité de suivi de l' OLAF est composé de cinq experts indépendants qui poursuivent cependant leur **activité professionnelle** normale .
>
> ITALIAN: Detto comitato è composto da cinque esperti indipendenti , che tuttavia non cessano di svolgere la loro **attività lavorativa** precedente .

The word alignment links for *occupation* in this sentence are:

- Spanish: actividad profesional
- Dutch: beroepsbezigheden
- German: Berufstätigkeit
- French: activité professionnelle
- Italian: attività lavorativa

For the evaluation we take both the entire compound as well as the part that translates the ambiguous word into consideration. If we take for instance a sentence containing *sugar plant*, the translation in Dutch will be *suikerfabriek*, and in this particular case both *suikerfabriek* as *fabriek* (which is the part of the compound that refers to *plant*), will be considered as valid translations for evaluation.

2. In the second step, all translations that result from the word alignment process are clustered per target language. On the basis of the sentence IDs, the translations in all languages are automatically coupled. Three human annotators validate this multilingual clustering. Translations that correspond to English multiword units are identified, and the different compound parts are separated by §§ in the clustering file.

   Table 1 shows the cluster for *bank* that refers to the *West Bank* meaning, where all compound parts are separated by §§.

More information about the manual construction of the sense inventory can be found in [3].

3 Data sets

For the competition, two data sets will be developed: one trial data set containing 20 manually labeled instances for five ambiguous words and one test set containing 50 instances for 20 ambiguous English nouns.

The sense inventory that results from the clustering (see Section 2.2) is used to annotate the sentences in the development and test set. This implies that a given target word is annotated with the appropriate sense cluster. The goal is to reach a consensus cluster per sentence. But again, if no consensus is

Dutch	**Italian**	**French**	**German**	**Spanish**
Bank	Cisgiordania	Cisjordanie	West§§jordanland	Cisjordania
Cisjordanië	sponda	rive	West§§bank-Umweltprojekt	río Jordán
Jordaan-oever	cisgiordano	bande	West-Bank	Franja
Jordaan§§oever	riva occidentale del Giordano	cisjordanien	West§§bank	costa
Transjordanië	sponda occidentale del Giordano	Bank	Bank	orilla
West§§bank	Bank	Banque	West§§jordangebiet	Ribera
West§§oever	striscia di Gaza		West§§jordanien	junto
deel	riva		West §§jordan§§ufer	
oever			West§§küste	
			West§§ufer	
			Ufer	
			Jordan§§ufer	

Table 1: translation cluster for the English noun *bank* in the *Cisjordan* meaning

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.

These chosen clusters are then used by four native annotators to select their top 3 translations per sentence. These potentially different translations are kept to construct the gold standard and to calculate frequency information for all answer translations.

The example below shows one annotation result for French and Italian for a sentence containing *bank*. The human annotators (a) pick the right cluster and (b) select their top 3 translations from this cluster:

> SENTENCE 3. Considering the importance of the existing links between the Community and the Palestinian people of the West Bank and the Gaza Strip, and the common values that they share
>
> French Cluster: 4
>
> French 1 Cisjordanie
> French 2 rive
> French 3 bande
>
> Italian Cluster: 4
>
> Italian 1 Cisgiordania
> Italian 2 riva

Italian 3 sponda

3.1 Development data

The development data set contains 5 polysemous nouns (*bank, occupation, passage, movement, plant*), for which we provide:

- the manually built sense inventory (see trial_clustering.xls) based on Europarl. This Excel sheet contains one worksheet per ambiguous word, where all translations have been clustered over the five languages per meaning. The coarse-grained sense clusters have been further split up into more fine-grained sub clusters.

- 20 example instances that are selected from the JRC-ACQUIS Multilingual Parallel Corpus[3], and annotated with all translations provided by the human annotators. Human annotators are asked (1) to pick the most appropriate cluster and (2) pick the three translations from this cluster that are best suited for the occurrence of the ambiguous word.

3.2 Test data

The test data contains 50 instances for 20 nouns from the test data as used in the Cross-Lingual Lexical Substitution Task[4]. Selecting the target words from the set of nouns thats will be used for the Lexical Substitution Task should make it easier for systems to participate in both tasks.
The nouns that will be used for the test set are *coach, education, execution, figure, job, letter, match, mission, mood, paper, post, pot, range, rest, ring, scene, side, soil, strain and test*).

For this test set, we will also:

- build up the sense inventory manually. The resulting clustering will be used for the evaluation, and thus not be made available to systems that are participating to the task.

- annotate all instances with the translations provided by the human annotators and use this as the gold standard for the evaluation.

3.3 Format of the trial and test sentences: see [word].data

The format of both the input file and gold standard is similar to the format that is used for the Cross-Lingual Lexical Substitution task [6]. This should allow teams to easily participate to both tasks. All data files will be delivered in UTF-8 format.

The input files to systems for evaluation will comply with following format:

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

[4] http://lit.csci.unt.edu/index.php/Semeval_2010

```
<corpus lang="english">
<lexelt item="{lemma}.{pos}">

<instance id="{id}">
<context> ... <head>{target word}</head> ...</context>
</instance>

</lexelt>
</corpus>
```

Each $< lexelt >$ tag focuses on a particular lemma and part of speech, as specified in *item*. As we focus on nouns for this task, the part of speech tag will always be ”n”. Each sentence starts with an *instance* tag, that specifies the unique ID of the sentence ($id = ”x”$). The sentence itself is enclosed in the $< context >$ tags, and contains an instance of the target lemma between the $< head >$ tags. The syntax is illustrated in following example for bank;

```
<corpus lang="english">
<lexelt item="bank.n">

<instance id="16">
<context>If one or more branches of a participating NCB are closed on an NCB
 business day owing to local or regional <head>bank</head> holidays,
 the relevant participating NCB shall inform the institutions in advance
 of the arrangements to be made for transactions involving those branches.</context>
</instance>

<instance id="17">
<context>1. A bearer of electronic money may, during the period of validity,
ask the issuer to redeem it at par value in coins and <head>bank</head> notes
or by a transfer to an account free of charges other than those strictly necessary
to carry out that operation.</context>
</instance>

</lexelt>
</corpus>
```

3.4 Format of the gold standard: see [word].[language].gold

The gold standard format is the same for both the *best* and *oof* evaluations, but the system output files differ for both scoring methods. All clustered translations are manually lemmatised, so systems should ensure that their answers are lemmatized as well. A particular case for German is the *Eszett*, where both the ß and double *ss* are accepted as orthographical variants.

The format of the gold standard is the following:

```
{lexelt}{.language} \s {id} \s :: \s {list of translations}
```

lexelt contains the "lemma.pos" combination, whereas *language* contains the language code. The five language codes that are used are:

- de: German
- fr: French
- nl: Dutch
- es: Spanish
- it: Italian

Each item in the list of possible translations is separated by ";" and consists of the lemmatized word and the frequency count that reflects the number of times a translation has been chosen by the human annotators.

Example of the gold standard format:

```
bank.n.fr 1 :: bank 1;banque 2;institution 1;
bank.n.fr 10 :: berge 1;bord 2;rivage 1;rive 2;
bank.n.fr 11 :: bande 1;cisjordanie 2;rive 1;
```

3.5 System format for best

The system output files should follow the same format as the gold standard without the frequency information:

```
{lexelt}{.language} \s {id} \s :: \s {list of translation suggestions}
```

The best guess of the system should appear first in the list as in:

```
bank.n.fr 10 :: bord;
bank.n.fr 11 :: cisjordanie;rive;
```

The *best* system output can contain as many guesses as the system believes are appropriate for a given test instance, but the credit for each correct guess will be divided by the number of guesses. The first guess in the list is taken as the best answer, and will get more weight. For a detailed discussion of the evaluation strategy, we refer to [6].

3.6 System format for oof

The system output files for the *oof* evaluation should stick to following format:

```
{lexelt}{.language} \s {id} \s ::: \s {list of translation suggestions}
```

The only difference with the *best* evaluation are the three colons instead of two colons (best):

```
movement.n.it 3 ::: circolare;fluttuazione;mosso;movimento;movimiento;muoversi;
movement.n.it 4 ::: circolazione;libera circolazione;traffico;trasferimento;
```

For this evaluation measure, systems can provide up to 5 substitutes. The credit for each correct guess is not divided by the number of guesses and the order of the guesses is not taken into account.

4 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.

For the evaluation of the participating systems we will use a minor adapted version of the evaluation scheme which is inspired by the English lexical substitution task in SemEval 2007 [4]. The evaluation will be performed using precision and recall (P and R in the equations that follow). We perform both a:

- *best result* evaluation: any number of guesses, with the very best guess (bg) first
- a more relaxed evaluation for the *top five results*: no penalization for multiple guesses, systems are only allowed to provide 5 guesses

We will probably also perform an additional evaluation where we do not only take into account the translations that are picked by the human annotators, but also take into account all translations belonging to the cluster that has been chosen by the human annotators.

4.1 scoring formula

We use the same evaluation formula as described in [4].
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 [4], 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 divided by the number of guesses. In this way, systems that output a lot of guesses are not favoured.

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

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

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.

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

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

4.2 Scoring script

Use following command in order to run the evaluation script (which is a perl script):

```
perl ScorerTask3.pl system_output gold_standard [-t best|oof ] [-v]
```

where

- system_output: the file that contains the system output for a particular language (required)
- gold_standard: the file that contains the gold standard provided by the human annotators (required)
- -t specifies the scoring type: *best* or *oof* (out of five), with *best* as the default (optional)
- -v creates line-by-line evaluation scores in the output file (optional)

5 Baselines

We will produce three – frequency-based – baselines:

1. 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.
2. 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.
3. As a third baseline, we will consider using the most frequent sense baseline based on EuroWordNet[5], which is available in the five target languages.

References

[1] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In *Computational Linguistics*, pages 177–184, 1991.

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

[3] E. Lefever and V. Hoste. 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, 2009.

[4] D. McCarthy and R. Navigli. 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, 2007.

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

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

[5] http://www.illc.uva.nl/EuroWordNet

[7] Dan Tufiş, Radu Ion, and Nancy Ide. 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 2004. Association for Computational Linguistics.