UHD: Cross-Lingual Word Sense Disambiguation Using Multilingual Co-occurrence Graphs

Carina Silberer and **Simone Paolo Ponzetto**
Department of Computational Linguistics
Heidelberg University
{silberer,ponzetto}@cl.uni-heidelberg.de

Abstract

We describe the University of Heidelberg (UHD) system for the Cross-Lingual Word Sense Disambiguation SemEval-2010 task (CL-WSD). The system performs CL-WSD by applying graph algorithms previously developed for monolingual Word Sense Disambiguation to multilingual co-occurrence graphs. UHD has participated in the BEST and out-of-five (OOF) evaluations and ranked among the most competitive systems for this task, thus indicating that graph-based approaches represent a powerful alternative for this task.

1 Introduction

This paper describes a graph-based system for Cross-Lingual Word Sense Disambiguation, i.e. the task of disambiguating a word in context by providing its most appropriate translations in different languages (Lefever and Hoste, 2010, CL-WSD henceforth). Our goal at SemEval-2010 was to assess whether graph-based approaches, which have been successfully developed for monolingual Word Sense Disambiguation, represent a valid framework for CL-WSD. These typically transform a knowledge resource such as WordNet (Fellbaum, 1998) into a graph and apply graph algorithms to perform WSD. In our work, we follow this line of research and apply graph-based methods to *multilingual co-occurrence graphs* which are automatically created from parallel corpora.

2 Related Work

Our method is heavily inspired by previous proposals from Véronis (2004, Hyperlex) and Agirre et al. (2006). Hyperlex performs graph-based WSD based on *co-occurrence graphs*: given a monolingual corpus, for each target word a graph is built where nodes represent content words co-occurring with the target word in context, and edges connect the words which co-occur in these contexts. The second step iteratively selects the node with highest degree in the graph (root hub) and removes it along with its adjacent nodes. Each such selection corresponds to isolating a high-density component of the graph, in order to select a sense of the target word. In the last step the root hubs are linked to the target word and the Minimum Spanning Tree (MST) of the graph is computed to disambiguate the target word in context. Agirre et al. (2006) compare Hyperlex with an alternative method to detect the root hubs based on PageRank (Brin and Page, 1998). PageRank has the advantage of requiring less parameters than Hyperlex, whereas the authors ascertain equal performance of the two methods.

3 Graph-based Cross-Lingual WSD

We start by building for each target word a multilingual co-occurrence graph based on the target word's aligned contexts found in parallel corpora (Sections 3.1 and 3.2). Multilingual nodes are linked by translation edges, labeled with the target word's translations observed in the corresponding contexts. We then use an adapted PageRank algorithm to select the nodes which represent the target word's different senses (Section 3.3) and, given these nodes, we compute the MST, which is used to select the most relevant words in context to disambiguate a given test instance (Section 3.4). Translations are finally given by the incoming translation edges of the selected context words.

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 134–137,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

3.1 Monolingual Graph

Let C_s be all contexts of a target word w in a source language s, i.e. English in our case, within a (PoS-tagged and lemmatized) monolingual corpus. We first construct a monolingual co-occurrence graph $G_s = \langle V_s, E_s \rangle$. We collect all pairs (cw_i, cw_j) of co-occurring nouns or adjectives in C_s (excluding the target word itself) and add each word as a node into the initially empty graph. Each co-occurring word pair is connected with an edge $(v_i, v_j) \in E_s$, which is assigned a weight $w(v_i, v_j)$ based on the strength of association between the respective words cw_i and cw_j:

$$w(v_i, v_j) = 1 - \max\,[p(cw_i|cw_j), p(cw_j|cw_i)].$$

The conditional probability of word cw_i given word cw_j is estimated by the number of contexts in which cw_i and cw_j co-occur divided by the number of contexts containing cw_j.

3.2 Multilingual Graph

Given a set of target languages L, we then extend G_s to a labeled multilingual graph $G_{ML} = \langle V_{ML}, E_{ML} \rangle$ where:

1. $V_{ML} = V_s \cup \bigcup_{l \in L} V_l$ is a set of nodes representing content words from either the source (V_s) or the target (V_l) languages;

2. $E_{ML} = E_s \cup \bigcup_{l \in L} \{E_l \cup E_{s,l}\}$ is a set of edges. These include (a) *co-occurrence edges* $E_l \subseteq V_l \times V_l$ between nodes representing words in a target language (V_l), weighted in the same way as the edges in the monolingual graph; (b) labeled *translation edges* $E_{s,l}$ which represent translations of words from the source language into a target language. These edges are assigned a complex label $t \in \mathcal{T}_{w,l}$ comprising a translation of the word w in the target language l and its frequency of translation, i.e. $E_{s,l} \subseteq V_s \times \mathcal{T}_{w,l} \times V_l$.

The multilingual graph is built based on a word-aligned multilingual parallel corpus and a multilingual dictionary. The pseudocode is presented in Algorithm 1. We start with the monolingual graph from the source language (line 1) and then for each target language $l \in L$ in turn, we add the translation edges $(v_s, t, v_l) \in E_{s,l}$ of each word in the source language (lines 5-15). In order to include the information about the translations of w in the different target languages, each translation edge (v_s, t, v_l) receives a translation label t. Formally, let $C_{v_s} \subseteq C_s$ be the contexts where v_s and w co-occur, and C_{v_l} the word-aligned contexts in language l of C_{v_s}, where v_s is translated as v_l. Then each edge between nodes v_s and v_l is labeled with a translation label t (lines 14-15): this includes a translation of w in C_{v_l}, its frequency of translation and the information of whether the translation is monosemous, as found in a multilingual dictionary, i.e. EuroWordNet (Vossen, 1998) and PanDictionary (Mausam et al., 2009). Finally, the multilingual graph is further extended by inserting all possible co-occurrence edges $(v_i, v_j) \in E_l$ between the nodes for the target language l (lines 16-19, i.e. we apply the step from Section 3.1 to l and C_l). As a result of the algorithm, the multilingual graph is returned (line 20).

Algorithm 1 Multilingual co-occurrence graph.

```
Input:  target word $w$ and its contexts $C_s$
        monolingual graph $G_s = \langle V_s, E_s \rangle$
        set of target languages $L$
Output: a multilingual graph $G_{ML}$

 1: $G_{ML} = \langle V_{ML}, E_{ML} \rangle \leftarrow G_s = \langle V_s, E_s \rangle$
 2: for each $l \in L$
 3:   $V_l \leftarrow \emptyset$
 4:   $C_l$ := aligned sentences of $C_s$ in lang. $l$
 5:   for each $v_s \in V_s$
 6:     $T_{v_s,l}$ := translations of $v_s$ found in $C_l$
 7:     $C_{v_s} \subseteq C_s$ := contexts containing $w$ and $v_s$
 8:     for each translation $v_l \in T_{v_s,l}$
 9:       $C_{v_l}$ := aligned sentences of $C_{v_s}$ in lang. $l$
10:       $\mathcal{T}_{w,C_{v_l}} \leftarrow$ translation labels of $w$ from $C_{v_l}$
11:       if $v_l \notin V_{ML}$ then
12:         $V_{ML} \leftarrow V_{ML} \cup v_l$
13:         $V_l \leftarrow V_l \cup v_l$
14:       for each $t \in \mathcal{T}_{w,C_{v_l}}$
15:         $E_{ML} \leftarrow E_{ML} \cup (v_s, t, v_l)$
16:   for each $v_i \in V_l$
17:     for each $v_j \in V_l,\ i \neq j$
18:       if $v_i$ and $v_j$ co-occur in $C_l$ then
19:         $E_{ML} \leftarrow E_{ML} \cup (v_i, v_j)$
20: return $G_{ML}$
```

3.3 Computing Root Hubs

We compute the root hubs in the multilingual graph to discriminate the senses of the target word in the source language. Hubs are found using the adapted PageRank from Agirre et al. (2006):

$$PR(v_i) = (1-d) + d \sum_{j \in deg(v_i)} \frac{w_{ij}}{\sum_{k \in deg(v_j)} w_{jk}} PR(v_j)$$

where d is the so-called damping factor (typically set to 0.85), $deg(v_i)$ is the number of adjacent nodes of node v_i and w_{ij} is the weight of the co-occurrence edge between nodes v_i and v_j.

Since this step aims to induce the senses for the target word, only nodes referring to words in English can become root hubs. However, in order to use additional evidence from other languages, we furthermore include in the computation of PageRank co-occurrence edges from the target languages, as long as these occur in contexts with ‘safe’, i.e. *monosemous*, translations of the target word. Given an English co-occurrence edge $(v_{s,i}, v_{s,j})$ and translation edges $(v_{s,i}, v_{l,i})$ and $(v_{s,j}, v_{l,j})$ to nodes in the target language l, labeled with monosemous translations, we include the co-occurrence edge $(v_{l,i}, v_{l,j})$ in the PageRank computation. For instance, animal and biotechnology are translated in German as Tier and Biotechnologie, both with edges labeled with the monosemous Pflanze: accordingly, we include the edge (Tier, Biotechnologie) in the computation of $PR(v_i)$, where v_i is either animal or biotechnology.

Finally, following Véronis (2004), a MST is built with the target word as its root and the root hubs of G_{ML} forming its first level. By using a multilingual graph, we are able to obtain MSTs which contain translation nodes and edges.

3.4 Multilingual Disambiguation

Given a context W for the target word w in the source language, we use the MST to find the most relevant words in W for disambiguating w. We first map each content word $cw \in W$ to nodes in the MST. Since each word is dominated by exactly one hub, we can find the relevant nodes by computing the correct hub $disHub$ (i.e. sense) and then only retain those nodes linked to $disHub$. Let W_h be the set of mapped content words dominated by hub h. Then, $disHub$ can be found as:

$$disHub = \operatorname*{argmax}_h \sum_{cw \in W_h} \frac{d(cw)}{dist(cw, h) + 1}$$

where $d(cw)$ is a function which assigns a weight to cw according to its distance to w, i.e. the more words occur between w and cw within W, the smaller the weight, and $dist(cw, h)$ is given by the number of edges between cw and h in the MST. Finally, we collect the translation edges of the retained context nodes W_{disHub} and we sum the translation counts to rank each translation.

4 Results and Analysis

Experimental Setting. We submitted two runs for the task (UHD-1 and UHD-2 henceforth). Since we were interested in assessing the impact of using different resources with our methodology, we automatically built multilingual graphs from different sentence-aligned corpora, i.e. Europarl (Koehn, 2005) for UHD-1, augmented with the JRC-Acquis corpus (Steinberger et al., 2006) for UHD-2[1]. Both corpora were tagged and lemmatized with TreeTagger (Schmid, 1994) and word aligned using GIZA++ (Och and Ney, 2003). For German, in order to avoid the sparseness deriving from the high productivity of compounds, we performed a morphological analysis using Morphisto (Zielinski et al., 2009).

To build the multilingual graph (Section 3.2), we used a minimum frequency threshold of 2 occurrences for a word to be inserted as a node, and retained only those edges with a weight less or equal to 0.7. After constructing the multilingual graph, we additionally removed those translations with a frequency count lower than 10 (7 in the case of German, due to the large amount of compounds). Finally, the translations generated for the BEST evaluation setting were obtained by applying the following rule onto the ranked answer translations: add translation tr_i while $count(tr_i) \geq count(tr_{i-1})/3$, where i is the i-th ranked translation.

Results and discussion. The results for the BEST and out-of-five (OOF) evaluations are presented in Tables 1 and 2 respectively. Results are computed using the official scorer (Lefever and Hoste, 2010) and no post-processing is applied to the system's output, i.e. we do not back-off to the baseline most frequent translation in case the system fails to provide an answer for a test instance. For the sake of brevity, we present the results for UHD-1, since we found no statistically significant difference in the performance of the two systems (e.g. UHD-2 outperforms UHD-1 only by $+0.7\%$ on the BEST evaluation for French).

[1] As in the case of Europarl, only 1-to-1-aligned sentences were extracted.

Language	P	R	Mode P	Mode R
FRENCH	20.22	16.21	17.59	14.56
GERMAN	12.20	9.32	11.05	7.78
ITALIAN	15.94	12.78	12.34	8.48
SPANISH	20.48	16.33	28.48	22.19

Table 1: BEST results (UHD-1).

Language	P	R	Mode P	Mode R
FRENCH	39.06	32.00	37.00	26.79
GERMAN	27.62	22.82	25.68	21.16
ITALIAN	33.72	27.49	27.54	21.81
SPANISH	38.78	31.81	40.68	32.38

Table 2: OOF results (UHD-1).

Overall, in the BEST evaluation our system ranked in the middle for those languages where the majority of systems participated – i.e. second and fourth out of 7 submissions for FRENCH and SPANISH. When compared against the baseline, i.e. the most frequent translation found in Europarl, our method was able to achieve in the BEST evaluation a higher precision for ITALIAN and SPANISH (+1.9% and +2.1%, respectively), whereas FRENCH and GERMAN lie near below the baseline scores (−0.5% and −1.0%, respectively). The trade-off is a recall always below the baseline. In contrast, we beat the Mode precision baseline for all languages, i.e. up to +5.1% for SPANISH. The fact that our system is strongly precision-oriented is additionally proved by a low performance in the OOF evaluation, where we always perform below the baseline (i.e. the five most frequent translations in Europarl).

5 Conclusions

We presented in this paper a graph-based system to perform CL-WSD. Key to our approach is the use of a co-occurrence graph built from multilingual parallel corpora, and the application of well-studied graph algorithms for monolingual WSD (Véronis, 2004; Agirre et al., 2006). Future work will concentrate on extensions of the algorithms, e.g. computing hubs in each language independently and combining them as a joint problem, as well as developing robust techniques for unsupervised tuning of the graph weights, given the observation that the most frequent translations tend to receive too much weight and accordingly crowd out more appropriate translations. Finally, we plan to investigate the application of our approach directly to multilingual lexical resources such as PanDictionary (Mausam et al., 2009) and BabelNet (Navigli and Ponzetto, 2010).

References

Eneko Agirre, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art WSD. In *Proc. of EMNLP-06*, pages 585–593.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proc. of SemEval-2010*.

Mausam, Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proc. of ACL-IJCNLP-09*, pages 262–270.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL-10*.

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

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP '94)*, pages 44–49.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. of LREC '06*.

Jean Véronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands.

Andrea Zielinski, Christian Simon, and Tilman Wittl. 2009. Morphisto: Service-oriented open source morphology for German. In *State of the Art in Computational Morphology*, volume 41 of *Communications in Computer and Information Science*, pages 64–75. Springer.