Applying cross-lingual WSD to wordnet development

Marianna Apidianaki, Benoît Sagot

Applying cross-lingual WSD to wordnet development

Marianna Apidianaki[1,2], **Benoît Sagot**[1]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 30 rue du Château des Rentiers, 75013 Paris, France
2. LIMSI, France
benoit.sagot@inria.fr, marianna.apidianaki@limsi.fr

Abstract

The automatic development of semantic resources constitutes an important challenge in the NLP community. The methods used generally exploit existing large-scale resources, such as Princeton WordNet, eventually combined with information extracted from multilingual resources and parallel corpora. In this paper, we show how cross-lingual Word Sense Disambiguation can be applied to wordnet development. We apply the proposed method to WOLF, a free wordnet for French still under development, in order to fill synsets that did not contain any literal yet.

1. Introduction

The need for lexical and semantic knowledge in NLP applications has steered several initiatives for resource development in recent years. Some of these attempts aimed the development of multilingual semantic resources on the basis of Princeton WordNet (PWN) (Fellbaum, 1998): its structure was generally preserved while its contents were imported in the newly built resources by various translation-based methods (Vossen, 1998; Tufis et al., 2004).

Despite the importance of these projects, the high cost of the manual methods employed and the limited coverage of the obtained resources have motivated the advent of automatic wordnet development methods (Dyvik, 1998). In this stream of research, semantic information is acquired from parallel corpora on the basis of the assumption that the translations of words in texts offer insights into their semantics (Resnik and Yarowsky, 1997; Diab and Resnik, 2002). More recently, a multilingual semantic network, BabelNet, has been automatically built by jointly exploiting PWN, Wikipedia and the output of Statistical Machine Translation systems (Navigli and Ponzetto, 2010). Following this line of research, our aim is to show how cross-lingual Word Sense Disambiguation (WSD) (Apidianaki, 2009) can be applied to wordnet development, for creating new resources or enriching existing ones. We illustrate the approach by integrating new information, in the form of sense clusters, into the French semantic resource WOLF (Sagot and Fišer, 2008) and analyzing the obtained results.

2. WOLF

WOLF is a freely available wordnet for French, created on the basis of PWN (version 2.0) (Sagot and Fišer, 2008). This wordnet was automatically built following the *expand* model (Vossen, 1998; Tufis et al., 2004). Monosemous literals in the PWN were translated using a bilingual French-English lexicon built from various multilingual resources. Polysemous PWN literals were handled by an *alignment* approach based on the multilingual parallel corpus SEE-ERA.NET (**?**). The synsets obtained from both approaches were then merged. The resulting network, WOLF, preserves the hierarchy and structure of PWN 2.0 and contains the definitions and usage examples provided in PWN for each synset. Nevertheless, as information was not found for all PWN synsets by these automatic methods, WOLF is rather sparse. In total, it contains 32,351 non-empty synsets including 37,991 unique literals (against 115,424 synsets with 145,627 literals in PWN 2.0).

3. Enriching the WOLF

Filling empty synsets in a wordnet can be achieved by creating clusters of synonyms and defining the place where they should be located in the hierarchy. The cross-lingual WSD method proposed by (Apidianaki, 2009) is well adapted to this task. It exploits the results of a Word Sense Induction (WSI) method that clusters the translations of words in a parallel corpus according to their similarity (Apidianaki, 2008). The translation clusters are characterized by feature vectors (see 3.1.) that can be used for assessing the similarity of a cluster and a synset, thanks to information extracted from the PWN (see 3.2.).

3.1. Word Sense Induction

The WSI method is trained on the sentence aligned FR−EN part of the EUROPARL corpus (release v6) (Koehn, 2005). The corpus is lemmatized, POS-tagged (Schmid, 1994) and word-aligned (Och and Ney, 2003). Bilingual lexicons are then extracted for each translation direction (EN−FR/FR−EN) and filtered according to alignment scores and POS, while an intersection filter discards any correspondences not found in both lexicons. The translations used for clustering are the ones that translate *w* more than 10 times in the training corpus.

Each translation is characterized by a vector built from the lemmas of the content words in the corresponding source language sentences. A similarity score is computed for each translation pair by a variation of the Weighted Jaccard measure (Grefenstette, 1994; Apidianaki, 2008). Translations with a score above a threshold, defined locally for each *w*, are considered as semantically related (Apidianaki and He, 2010). The clustering algorithm groups the translations according to their similarity and the obtained sense-clusters describe the senses of the corresponding source language words. The clusters

generated, for instance, for the English noun *stage*, and which contain the translations retained for the word from the training corpus, describe its two senses: {stade, phase, étape} and {scène}.

3.2. Cross-lingual WSD

The generated EN–FR sense cluster inventory contains entries for English words of different POS. In this paper, as a first experiment, we focus on word meanings corresponding to empty synsets in WOLF.

The unsupervised WSD classifier used (Apidianaki, 2009) exploits the WSI results. In a classic WSD task, the clusters constitute the candidate senses of a word from which the most adequate one has to be selected for each instance of the word in context. This selection is performed by comparing the vectors of the clusters to information in the new context. In the current setting, where the WSD method aims at assigning clusters to empty synsets in WOLF, the information used for WSD consists of the words found in the corresponding PWN synsets and their related synsets, their definitions and usage examples. The retained information is lemmatized (Schmid, 1994) and gathered in a bag of words. The adequacy of a cluster for filling a given synset is estimated by comparing the cluster's vector with the PWN information retained for the synset. If common features (CF)s are found with just one cluster, this cluster is selected. Otherwise, each 'cluster-synset' association is assigned a score, corresponding to the mean of the weights of the CFs relatively to the clustered translations (weights assigned to each feature during WSI). In formula 1, j is the number of CFs and i is the number of translations in the cluster characterized by a CF. The highest scored cluster is selected and assigned to the empty synset.

$$assoc\_score = \frac{\sum_i \sum_j w(T_i, CF_j)}{i * j} \quad (1)$$

The empty synset 'odd#a#2' (definition: "not easily explained"; usage: "it is odd that his name is never mentioned"), for instance, is correctly filled by the FR cluster {curieux, bizarre}. The other clusters of *odd*, which were scored less, are: {contradictoire, singulier, bizarre} and {curieux, étrange}.

4. Evaluation

Overall, 3,904 previously empty synsets have been filled by our approach (2,333 nominal, 576 verbal, 709 adjectival and 286 adverbial synsets). We have examined manually 10% of them for each POS, for evaluating the quality of the proposed clusters (i.e. do the clustered words share a common meaning?), and the correctness of their association to some synset in WOLF – which can only happen if the clusters are good. Both aspects have been evaluated by two annotators. The inter-annotator agreement was measured at κ = 0.67, for cluster quality, and 0.59 for the WSD results, which is conventionally interpreted as "good" agreement (Cohen, 1960).

According to the evaluation results obtained for all POS, the clusters group semantically similar words in 75.5% of the cases, with significant variations for different POS, as shown in Table 1. This is due to the restrictive cluster quality criterion used, according to which one incorrect word in an otherwise correct cluster, turns the whole cluster into an incorrect one. This strict criterion has a strong effect on clusters containing many translations, as is often the case for verb clusters.

	Nouns	**Verbs**	**Adjs**	**Advs**
Clusters	72.1	62.9	81.0	86.2
WSD	64.6	53.0	75.1	73.7

Table 1: Evaluation Results (%)

The error analysis indicates some other cases of problematic clustering: cases where multiword units were not considered during word alignment ({considération, compte}, cluster of *consideration*, instead of {prise en compte, prise en considération}); clustering of topically related words ({raisin, moût} corresponding to *grape*); tagging problems ({énergétique, énergie} corresponding to *energy*); clustering of antonymous, but distributionally similar, words ({sain, malsain}, cluster of *unhealthy*).

Given that only good clusters can be correctly integrated into WOLF, we calculate the performance of the WSD method by reference to the number of good clusters. The score obtained for the WSD insertions by averaging the scores provided by the two annotators is 67%, which is very encouraging. We should highlight the difficulty of this task as the WSD method is asked to fill synsets that were left empty by the methods initially employed for creating WOLF. These empty synsets often correspond to rare senses in PWN, that may not exist in the training corpus, or to senses for which little information is provided.

In order to more fairly estimate the performance of the WSD method in this setting, we also tested it on the whole resource. In this case, the method was asked to select the most adequate synset for each cluster from *all* the synsets in WOLF (not only the empty ones). In this case, the method reaches a performance of 80.13%, which shows that it is particularly well adapted to the wordnet development task.

5. Conclusion

We have shown that a cross-lingual WSD method, based on unsupervised WSI, can be efficiently used for wordnet development. We integrated sense-clusters of translations into a French wordnet resource, the WOLF, by exploiting information found in PWN. The results of our experiment on empty synsets indicate that the proposed unsupervised methods are particularly useful for the construction, or enrichment, of wordnets in languages other than English.

We now intend to use the proposed methods in order to enrich other, non-empty, synsets in WOLF. Moreover, we will apply measures of semantic similarity on PWN in order to merge closely related synsets and, consequently, reduce the number of empty ones.

6. References

Marianna Apidianaki and Yifan He. 2010. An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. In *Proceedings of the 7th International*

Workshop on Spoken Language Translation (IWSLT), pages 219–226, Paris, France.

Marianna Apidianaki. 2008. Translation-oriented sense induction based on parallel corpora. In *Language Resources and Evaluation Conference (LREC)*, pages 3269–3275, Marrakech, Morocco.

Marianna 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.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mona Diab and Philip Resnik. 2002. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 255–262, Philadelphia.

Helge Dyvik. 1998. Translations as semantic mirrors: from parallel corpus to wordnet. In *Proceedings of the Workshop Multilinguality in the lexicon II at the 13th biennial European Conference on Artificial Intelligence (ECAI'98)*, pages 24–44, Brighton, UK.

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

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Dordrecht : Kluwer Academic Publisher.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Philip Resnik and David Yarowsky. 1997. A perspective on Word Sense Disambiguation Methods and their Evaluation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics : Why, What, and How?*, Washington, D.C.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Ontolex 2008*, Marrakech, Morocco.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal on Information Science and Technology. Special Issue on BalkaNet*, volume 7, pages 9–34.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.