Word Sense Disambiguation as a Wordnets’ Validation Method in Balkanet

Dan Tufiș1,2, Radu Ion1, Nancy Ide3

1Research Institute for Artificial Intelligence Calea „13 Septembrie”, no. 13, PO 050711, Bucharest
2University „A.I. Cuza” of Iași,
3Department of Computer Science, Vassar College, Poughkeepsie, NY 12604-0520 USA

tufis@racai.ro, radu@racai.ro, ide@cs.vassar.edu

Abstract

BalkaNet is a European project which aims at the development of monolingual wordnets for five languages in the Balkans area (Bulgarian, Greek, Romanian Serbia, and Turkish) and at improvement of the Czech wordnet developed in the EuroWordNet project. The wordnets are aligned to the Princeton Wordnet, according to the principles established by EuroWordNet. One of the main concerns of this project was the semantic validation of the wordnets alignment. To this end, we developed a WSD system based on parallel corpora which exploits the common intuition according to which words that are reciprocal translations in a parallel texts should have the same (or closely related) interlingual meanings. With wordnets under construction our WSD system is mainly a validation tool, pinpointing wrong interlingual alignments, incomplete or missing synsets in one or another of the wordnets. The evaluation of the WSD system showed very encouraging results.

Introduction

In previous papers (Erjavec et al., 2001, Ide et al. (2002), Tufis and Ion (2003) we reported on our sense clustering work, based on translation equivalents (Tufis (2002), Tufis and Barbù (2002)) extracted from parallel corpora. In (Tufis and Ion (2003)) we described a neutral method for labeling the sense clusters which considers only two languages Romanian and English). We say the method is neutral because it is not bound to any sense inventory name convention. Each cluster in the target language (Romanian) was numbered taking into account its size (sense number 1 being assigned to the largest cluster). Additionally, each cluster was labeled with the most frequent translation in one of the source languages (in our case, English). For instance, in our “1984” multilingual corpus the 45 occurrences of the Romanian word “deget” which were translated into the English part, were clustered into 3 classes containing 35, 6 and 4 occurrences and sense-numbered 1, 2 and 3 respectively. The occurrences in the three clusters were differently translated as “finger”, “toe” and “thumb” and therefore the sense labels of the three clusters were “1(finger)”, “2(toe)” and “3(thumb)”. However this is a favorable show-case when the three senses of the target word are differently lexicalized in the source language. In many cases, although the target word was clustered in different sets its translation in English was the same. For instance the word “miscare” was clustered into 3 classes which were labeled as “1(movement)”, “2(movement)” and “3(jerk)”. The meaning of the occurrences in the first cluster corresponded to a change of position that does not entail a change of location (cf. Princeton Wordnet) while the meaning of occurrences in the second cluster corresponded to the social sense. The granularity of the word sense clustering based on translation equivalents depends on the languages in the parallel corpus and the language register of the parallel text. This method is rather short-sighted to sense distinctions in one language which are not realized in at least one of the other languages as different lexemes. That is to say that if two senses of a word in the target language are not differently lexicalized in the source languages, there is no way to differentiate which occurrences of the target word are used in one sense and which are used in the other one. As the number and diversity of source languages grows, decreases the chance of not lexicalizing the two senses of the target word in at least one language. With six source languages, from three language families (Romance, Slavic and Ugro-Finic), sense clustering of English words was reasonable precise (around 75%). The labeling of sense clusters, although useful in many possible applications, was not only language-pair dependent (by the translation equivalent) but also text dependent (by the frequency rank). Therefore, we tried to compensate these drawbacks by using additional resources such as the aligned wordnets of the BalkaNet (or EuroWordNet) type. The new method for word sense disambiguation (WSD) uses the Princeton Wordnet 2.0 sense inventory and relies on the previous clustering algorithm as a back-off mechanism. The underlying hypothesis exploits the common intuition according to which words that are reciprocal translations in a parallel texts should have the same (or closely related) interlingual meanings (in terms of BalkaNet, ILI record-projections or simply ILI codes). This hypothesis is reasonable if the monolingual wordnets are reliable and they are correctly linked to the interlingual index (ILI). Quality assurance of the wordnets is one main concern in the BalkaNet project and to this end, the consortium developed several methods and tools for wordnets validation have been constructed. These validation methods are largely described in various papers authored by BalkaNet consortium members (see Proceedings of the Global WordNet Conference, Brno, 2004). We implemented a language independent WSD program, called WSDtool, which has been extended so that it can work also as a multilingual wordnet checker and as a specialized editor for correcting the spotted errors. We described in (Tufis et al. 2004) this system in the validation mode and showed that we were able to spot some interlingual alignment errors very hard to note just by human introspective analysis. Here we will discuss the word sense disambiguation based on the corrected wordnets.

The Basic Methodology

The methodology of the WSD based on parallel corpora and interlingually aligned wordnets assumes the following basic steps:
A) given a bitext \( T_{L1L2} \) in languages L1 and L2 for which there are aligned wordnets, one extracts the pairs of lexical items that are reciprocal translations: \( \{<W'_{L1}, W'_{L2}>\} \)

B) for each lexical alignment of interest, \( <W'_{L1}, W'_{L2}> \), one extracts, for each language, the ILI codes for the synsets that contain \( W'_{L1} \) and \( W'_{L2} \) respectively; thus, one gets two lists of ILI codes, \( L_{1}^{\text{ILI}}(W'_{L1}) \) and \( L_{2}^{\text{ILI}}(W'_{L2}) \), one for each language. The WSD of the lexical items under consideration comes to identify one ILI code common to the intersection \( L_{1}^{\text{ILI}}(W'_{L1}) \cap L_{2}^{\text{ILI}}(W'_{L2}) \) or a pair of ILI codes \( \text{ILI}1 \in L_{1}^{\text{ILI}}(W'_{L1}) \) and \( \text{ILI}2 \in L_{2}^{\text{ILI}}(W'_{L2}) \) so that \( \text{ILI}1 \) and \( \text{ILI}2 \) are the most similar ILI codes (below we elaborate on this issue) among the candidate pairs \( (L_{1}^{\text{ILI}}(W'_{L1}) \odot L_{2}^{\text{ILI}}(W'_{L2})) \) with \( \odot \) representing the Cartesian product among the two sets.

The A) processing step is crucial and its accuracy is essential for the success of the validation method. A recent shared task evaluation (www.cs.unt.edu/~rada/wpt) of different word aligners, organized on the occasion of the Conference of the NAACL showed that step A) may be solved quite reliably. Our system, best performing on the Romanian-English alignment exercise (Tufis et al. 2003) produced lexicons, relevant for wordnets evaluation, with an aggregated F-measure as high as 84.26%. Meanwhile, the word-aligner was further improved (Barbu, 2004) so that the current performances (on the same data) are about 1% better on all scores in word alignment and about 2% better in wordnet-relevant dictionaries (containing only translation equivalents of the same POS).

The B) step is the one where the aligned wordnets come to work. The correctness of the interlingual alignment is essential in finding a pair of ILI codes which would disambiguate the words that are found to be translation equivalents.

However, both in the EuroWordNet and in BalkaNet the ILI records are not structured, so we need to clarify what “closely related ILI” means. In the context of this research, we assume that the hierarchy preservation principle (Tufiș and Cristea, 2002) is sound. Under this assumption, we take the relatedness of two ILI records R1 and R2 as a measure for the semantic-similarity between the synsets Syn1 and Syn2 in PWN2.0 that correspond to R1 and R2. We used a very simple definition of the semantic similarity between two synsets: semantic-similarity (Syn1, Syn2)=1/1+k where the k is the number of oriented links from one synset to another or from the two synsets to the nearest common ancestor. The score is 1 when the two synsets are identical (or have the same ILI code), is 0.33 for two sister synsets and is 0.5 for mother/daughter or whole/part or any single link related synsets. Two ILI records R1 and R2 will be considered closely related if relatedness(R1, R2)=semantic-similarity (Syn1, Syn2)≥k, where k is an empirical threshold. In our experiments we considered it 0.33 (i.e. we allowed at most two link traversal between what we consider two closely related synsets).

Having a parallel corpus, containing texts in k+1 languages \( (T, L_1, L_2, \ldots L_k) \) and having monolingual wordnets for all of them, interlinked via an ILI-like structure, let us call the T language as the target language and \( L_1, L_2, \ldots L_k \) as source languages. The parallel corpus is encoded as a sequence of translation units (TU). A translation unit contains aligned sentences from each language, with tokens tagged and lemmatized as exemplified below (for details on encoding see http://nl.ijs.si/ME/V2/msd/html/):

\[
\begin{align*}
&\text{<tu id=“Ozz.113”>} \\
&\text{<seg lang=“en”>} \\
&\text{<s id=“Oen.1.1.24_2”>} \\
&\text{<w lemma=“Winston” ana=“Np”>Winston</w>} \\
&\text{<w lemma=“be” ana=“Vais3s”>was</w>} \\
&\ldots \\
&\text{<seg>} \\
&\text{<seg lang=“ro”>} \\
&\text{<s id=“Oro.1.2.23_2”>} \\
&\text{<w lemma=“Winston” ana=“Np”>Winston</w>} \\
&\text{<w lemma=“fi” ana=“Vmit3s”>era</w>} \\
&\ldots \\
&\text{<seg>} \\
&\text{<seg lang=“cs”>} \\
&\text{<s id=“Ocs.1.1.24_2”>} \\
&\text{<w lemma=“Winston” ana=“Np”>Winston</w>} \\
&\text{<w lemma=“se” ana=“Pxx--d--ypn--n”>s1</w>} \\
&\ldots \\
&\text{<tu>}
\end{align*}
\]

Table 1. A partial translation unit from the parallel corpus

| Occ #1 | Occ #2 | … | Occ #n |
|-------|-------|---|--------|
| \(L_1\) | \(eq_{11}\) | \(eq_{12}\) | … | \(eq_{1n}\) |
| \(L_2\) | \(eq_{21}\) | \(eq_{22}\) | … | \(eq_{2n}\) |
| \(L_k\) | \(eq_{k1}\) | \(eq_{k2}\) | … | \(eq_{kn}\) |

Table 2. The translation equivalents matrix (EQ matrix)

If a specific occurrence of the target word is not translated in language \( L_i \), then \( eq_{ij} \) is represented by the null string. This table is generated as a result of step A) discussed in the previous section. In the clustering system mentioned in the introductory section, the columns of this table were the vectors used by the agglomerative algorithm presented in (Ide et all. 2002). The step B) of the basic methodology (see previous section) transforms the matrix shown in Table 2 in a matrix of the same dimensions, as shown in Table 3, called VSA (Validation and Sense Assignment):
synsets in which occurs the translation equivalent for the j-th occurrence of $W_{EN}$. If no translation equivalent was found in language $L_i$ for the j-th occurrence of $W_{EN}$ VSA(i,j) is undefined, otherwise it is a set containing 0, 1 or more ILI codes. For undefined VSAs, this algorithm cannot make any decision with respect to the sense number of the respective occurrence of the target word. However, it is very unlikely that an entire column in Table 3 become undefined and as such the lack of information from one source language could be compensated from other source languages. When the cell VSA(i,j) contains a single ILI code this is the common sense in the two considered language for the j-th occurrence of the target word and its translation equivalence. Considering the English-Romanian example before <toe deget>, a corresponding VSA should contain only the ILI-code ENG20-0528265-n corresponding to sense 1 of toe in PWN and to sense number 3 of deget in Romanian wordnet. Thus the disambiguation of this translation pair would be <toe(1) deget(3)>.

The case for an empty set in a VSA cell is needs the notion of the semantic similarity introduced before. The situation appears when none of the senses of the target word corresponds to an ILI code to which a sense of the translation equivalent was linked. By using the semantic-similarity score presented earlier, the algorithm selects the pair in $L_{EN}^i(W_{EN}) \otimes L_{IL}^i(W_{IL})$ which shows the highest similarity. In case of ties, the pair corresponding to the most frequent sense (as seen in the English part of current bitext) of the target word is selected. If this heuristics cannot make the difference, choice is made in favour of the pair corresponding to the smallest PWN2.0 sense number for the target word. If may happen that no pair in $L_{EN}^{L_i}(W_{EN}) \otimes L_{IL}^{L_i}(W_{IL})$ meets the semantic similarity requirement. In this case, neither the occurrence of the target word nor its translation equivalent can be semantically disambiguated. As in the previous case, it is not likely that in all other languages the situation would be identical. Thus, again, the lack of information from one source language could be compensated from other source languages.

Finally, the case when in a VSA cell there are two or more ILI-codes exemplifies what we call cross-lingual ambiguity, i.e. two or more senses are common to the target word and the corresponding translation equivalent in the i-th language. For instance at least two senses of the English word movement are identical to the senses carried by the Romanian word misca. The tie heuristics discussed before are applied also in this situation.

**Back-off mechanism**

In the previous section we mentioned that in case a VSA cell corresponding to a pair of languages is undefined, or is empty and the semantic similarity score does not provide a solution, the occurrence of the target word (and by side effect its translation equivalent when the target word was translated) the lack of information could be compensated from other VSA corresponding to the same occurrence of the target word but a different language. This is a back-off mechanism which requires that for all languages in the parallel corpus aligned wordnets are available. Also, it requires that the aligned wordnets for all the languages be of similar quality as far as interlingual linking is concerned. Each of these requirements is very strong and asking for both of them is even stronger. The back-off mechanism relying on sense clustering based on translation equivalents as discussed in (Ide et al. 2002) removes the requirement for source language wordnets because the sense clustering relies only on translation equivalents. This back-off mechanism is described below. The occurrences of different target words which could not be assigned a sense label by the previous method, will receive one through the sense clustering algorithm mentioned in the first section of the paper (see (Ide et al. 2002) for details). Since the sense clustering occurs after the wordnet based WSD was performed, each word cluster containing a non-labeled occurrence will also contain occurrences already disambiguated. The majority sense label in the cluster will contaminate the unlabeled occurrences. In the unlikely event that all the unlabeled occurrences of a given word will be classified in sets containing no previously labeled occurrences of the word in case, we use two heuristics: the first one is a direct consequence of Zipf sense distribution law according to which senses of a word observe a skewed distribution with most of the words used with the same sense. According to this heuristics the clusters of unlabeled occurrences are considered subsets of the first, second, etc., largest clusters containing occurrences already disambiguated (sense-assigned). The second heuristics assumes that the occurrences not disambiguated are used with a sense that was already used in the text. The second heuristics is applied only when no other occurrence of the word in case has been disambiguated, one way or another. This could happen for instance when that there is a single occurrence of the word and it couldn’t be disambiguated by the WSDtool. The rationale for this heuristics is that the sense numbering in wordnets is based on sense frequencies in a balanced corpus (at least PWN2.0 is compliant with such an assumption). These heuristics are very similar to the one used for dealing with ties in the basic WSD algorithm.

**Test Data and WSD Evaluation**

In order to evaluate both the performance of the WSDtool and to assess the accuracy of the interlingual linking of the Balkanet wordnets we selected a bag of English target nouns, verbs and adjectives. The set of English targeted words were extracted from the parallel corpus “1984” so that all their senses (at least two per POS) defined in PWN2.0 were also implemented (and interlingually aligned) in all BalkaNet wordnets. There resulted 211 words with 1810 occurrences in the English part of the parallel corpus. We manually assigned senses to all these 1810 occurrences of the target words, building the Gold Standard (GS). In this experiment we also engaged the students enrolled in the Computational Linguistics Master program. An extraction script generated for each student a set of sentences containing occurrences of the targeted words. The extraction process ensured that the same sentence was in at least three student-sets. The context for sense disambiguation exercise was defined by the sentence containing the targeted word. Out of the students’ hand disambiguated a simple majority sense was computed (MAJ). Finally, the same targeted words were automatically disambiguated by the WSDtool algorithm (ALG). Out of the entire set of targeted words, the system could not make a decision for 398 occurrences, mainly
because they were not translated in the Romanian text. Another reason for failure was that translation of the target English, as found by the underlying word-aligner, was wrong. In (Barbu, 2004) it is shown that the error rate of our last version of the word aligner (for non-null alignments) is less than 11.5%. This error rate is largely due to English words occurring only once, or English words which are translated each time differently so that the corresponding translation pairs are hapax legomena. However, most of the time hapax legomena pairs which are cognates are correctly found. The evaluation program generated a file containing detailed information for each occurrence of the targeted word: - the sense number in the gold standard; - a majority voting sense number as resulted from the students’ sense assignments. - the sense assigned by the algorithm - the names of the students that evaluated the occurrence and the sense they assigned; In order to compare the results we took into account only the 1412 occurrences that were sense disambiguated by the algorithm (without the back-off mechanism). The table below summarizes the results in terms of agreement between GS and MAJ, GS and ALG, ALG and MAJ and GS, ALG and MAJ:

|            | GS=MAJ | GS=ALG | MAJ=ALG | GS=MAJ=ALG |
|------------|--------|--------|---------|------------|
| GS=MAJ     | 73.22% |        |         |            |
| GS=ALG     |        | 78.68% |         |            |
| MAJ=ALG    | 67.13% |        |         |            |
| GS=MAJ=ALG |        |        | 62.32%  |            |

Table 4. WSD agreements (without back-off mechanism)

It is interesting to note that the ALG agreement with GS is superior to the agreement between the majority of students and the GS (although we noticed a student who if considered instead of majority, her agreement with the GS was slightly better than GS=ALG score (78.71%). At the time of this writing the integration of the clustering algorithm with the WSD tool and back-off mechanism evaluation is not finished. A rough worst-case estimation for the GS=ALG could be done on the basis of the clustering accuracy we reported before (~75%) and therefore, the accuracy should not be lower than 77%-78%. We found this result extremely encouraging as it shows that the tedious hand-made WSD in building word-sense disambiguated corpora (presumably done by an expensive expert) can be avoided.

Conclusions

The WSD challenge is solving it in a monolingual context. The WSD accuracy we obtained is not surprisingly superior to the state of the art in monolingual WSD because in our case the knowledge imbedded by the human translators into the parallel text is of a tremendous help. The real challenge of the WSD problem is solving it in a monolingual context, because this is by far the most frequent and useful setting. The main problem for the monolingual WSD is the lack of enough training data. One way to improve this drawback is offered by the method we discussed here:

a) collect as many and as large parallel corpora in which the language of interest is paired at least to a language that could offer a hub wordnet (as is English with PWN2.0); the more languages (other than the hub) present in the parallel corpus, the better the final result;

b) apply the WSD algorithm as presented here for each parallel corpora collected as in a).

c) extract the language of interest from each parallel corpora processed in b) and concatenate the results.

The Internet becomes each day a richer repository of documents published in several languages (see for instance [http://www.balkantimes.com](http://www.balkantimes.com) where the same news is published in 10 languages). By following an approach in the spirit of the work reported here, one could incrementally build larger and larger sense-annotated corpora in his/her own language and approach the monolingual WSD problem much better equipped in terms of training data.

Acknowledgements

The work reported here was carried with within the European project Balkanet, no. IST-2000 29388 and support from the Romanian Ministry of Education and Research under the CORINT programme.

References

Barbu A.M. (2004). A word alignment system based on a translation equivalence extractor. In this volume.

Erjavec, T., Lawson A., Romary, L.(eds.) (1998). East Meet West: A Compendium of Multilingual Resources. TELRI-MULTEXT EAST CD-ROM, 1998, ISBN: 3-922641-46-6.

Erjavec T., Ide N., Tufiş, D. (2001). Automatic Sense Tagging Using Parallel Corpora. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, (pp. 212—219), Tokyo.

Ide, N., Erjavec, T., Tufiş, D. (2002). Sense Discrimination with Parallel Corpora. In Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, (pp. 56–60), Philadelphia.

Tufiş, D., Cristea, D. (2002). Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet. In Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation, (pp. 35–41), Las Palmas.

Tufiş, D., Barbu, A.M. (2002). Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing. In International Journal of Speech Technology, Kluwer Academic Publishers, 5(3), 199-209.

Tufiş, D. (2002). A cheap and fast way to build useful translation lexicons. In Proceedings of the 19th International Conference on Computational Linguistics, (pp. 1030—1036) Taipei.

Tufiş D., Ion R (2003). Word sense clustering based on translation equivalence in parallel texts; a case study in Romanian. In Proceedings of the International Conference on Speech and Dialog - SPED, (pp.13--26), Bucharest.

Tufiş D., Barbu A.M., Ion R. (2003). A word-alignment system with limited language resources. In Proceedings of the NaACL 2003 Workshop on Building and Using Parallel Texts; Romanian-English Shared Task, (pp.36—39), Edmonton.

Tufiş, D., Ion, R., Barbu, E. Barbu, V. (2004). Cross-Lingual Validation of Wordnets. In Proceedings of the 2nd International Wordnet Conference, (pp. 332-340) Brno.