Word Sense Disambiguation as a Wordnets’ Validation Method in Balkanet

Dan Tufis,1,2, Radu Ion1, Nancy Ide3

1Research Institute for Artificial Intelligence Calea „13 Septembrie”, no. 13, PO 050711, Bucharest
2University „A.I. Cuza” of Iasi,
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 the EuroWordNet consortium. One of the main concerns of this project is the interlingual validation of the wordnets alignment. To this end, we have 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.

Introduction
In previous papers (Ide et al., 2001, 2002) we reported on our sense clustering work, which is based on translation equivalents extracted from parallel corpora (Tufis (2002), Tufis and Barbu (2002)). Tufis and Ion (2003) build on this work and further describe a method to accomplish a “neutral” labeling for the sense clusters in Romanian and this work and further describe a method to accomplish a “neutral” labeling for the sense clusters in Romanian and

Tufis and Barbu (2002)). Tufis and Ion (2003) build on this work and further describe a method to accomplish a “neutral” labeling for the sense clusters in Romanian and

this work and further describe a method to accomplish a “neutral” labeling for the sense clusters in Romanian and

and/or languages from less diverse families are used, approximately 75% accurate; when fewer languages and Finno-Ugric), sense clustering of English words was

languages from three language families (Romance, Slavic

languages from three language families (Romance, Slavic

and 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 the EuroWordNet consortium. One of the main concerns of this project is the interlingual validation of the wordnets alignment. To this end, we have 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 Basic Methodology

Our methodology consists of the following basic steps:

1. given a bitext T L1L2 in languages L1 and L2 for which there are aligned wordnets, extract all pairs of lexical items that are reciprocal translations: {<W1, W2>}2.

2. for each lexical alignment <W1, W2>, extract the ILI codes for the synsets that contain W1 and W2 respectively to yield two lists of ILI codes, L1(L1) and L2(L2).

3. identify one ILI code common to the intersection L1(L1) (W1) ∩ L2(L2) or a pair of ILI codes I1I2 = L1(L1) (W1) and I1I2 = L2(L2), so that I1I2 are the most similar ILI codes (defined below) among the candidate pairs (L1(L1) (W1) ⊗ L2(L2)) [⊗ = Cartesian product]

The accuracy of step 1 is essential for the success of the validation method. For this step, we rely an alignment system that turned in the best performance on English-Romanian in a recent shared task evaluation of word aligners (Tufis, et al., 2003) and has since been further improved (Barbu, 2004) to produce the lexicons. The success of step 2 is dependent on the accuracy of the wordnets interlingual alignment to find a pair of ILI codes that can disambiguate the translation equivalents.

Our measure of ILI similarity is based on the principle of hierarchy preservation (Tufis & Cristea, 2002), which asserts that the relatedness (rel) of two ILI records R1 and R2 is a measure of semantic-similarity (ss) between two

Tufis & Cristea, 2002), which asserts that the relatedness (rel) of two ILI records R1 and R2 is a measure of semantic-similarity (ss) between two
synsets Syn₁ and Syn₂ in PWN2.0 that correspond to R₁ and R₂. We compute semantic-similarity by

\[ ss(Syn₁, Syn₂) = 1/1+k \]

where \( k \) is the number of links from Syn₁ to Syn₂ or from both Syn₁ and Syn₂ to the nearest common ancestor. The semantic similarity is 1 when the two synsets are identical (or have the same ILI code), 0.33 for two sister synsets, and 0.5 for mother/daughter, whole/part, or synsets related by a single link.

Two ILI records R₁ and R₂ are considered closely related if \( rel(R₁, R₂) = ss(Syn₁, Syn₂) \geq t \), where \( t \) is an empirical threshold, which in our experiments was set to 0.33 (i.e., we allowed at most two link traversals between what we consider two closely related synsets).

We use a parallel corpus containing texts in \( n+1 \) languages (T, L₁, L₂…Lₖ), where for the purposes of disambiguation T is the target language and L₁, L₂…Lₖ are the source languages. We also use monolingual wordnets for all \( n+1 \) languages, interlinked via an ILI-like structure. The parallel corpus is encoded as a sequence of translation units (TU), each containing aligned sentences from each language with tokens tagged and lemmatized as follows:

\[
\text{<tu id="Ozz.113">}
\text{<seg lang="en">}
\text{<w lemma="Winston" ana="Np">Winston</w>}
\text{<w lemma="be" ana="Vais3s">was</w>}
\text{... </seg>}
\text{<seg lang="ro">}
\text{<w lemma="Winston" ana="Np">Winston</w>}
\text{<w lemma="fi" ana="Vmi13s">era</w>}
\text{... </seg>}
\text{<seg lang="cs">}
\text{<w lemma="Winston" ana="Np">Winston</w>}
\text{<w lemma="se" ana="Px---d--ypn--n">si</w>}
\text{... </seg>}
\text{<seg lang="en">}
\text{... </seg>}
\text{<seg lang="cs">}
\text{... </seg>}
\text{</tu>}

For each source language and for all occurrences of a specific word in the target language T, we build a matrix of translation equivalents as shown in Table 1 (eqij represents the translation equivalent in the \( i \)th source language of the \( j \)th occurrence of the word in the target language):

<table>
<thead>
<tr>
<th>Occ #1</th>
<th>Occ #2</th>
<th>...</th>
<th>Occ #n</th>
</tr>
</thead>
<tbody>
<tr>
<td>L₁</td>
<td>eq₁₁</td>
<td>eq₁₂</td>
<td>...</td>
</tr>
<tr>
<td>L₂</td>
<td>eq₂₁</td>
<td>eq₂₂</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Lₖ</td>
<td>eqₖ₁</td>
<td>eqₖ₂</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1. The translation equivalents matrix (EQ matrix)

If a specific occurrence of the target word is not translated in language Lᵢ, eqᵢⱼ is represented by the null string. The table is generated as a result of step 1, as described in the previous section. Step 2 transforms the matrix in Table 1 to a matrix with the same dimensions (Table 2) called VSA (Validation and Sense Assignment):

<table>
<thead>
<tr>
<th>Occ #1</th>
<th>Occ #2</th>
<th>...</th>
<th>Occ #n</th>
</tr>
</thead>
<tbody>
<tr>
<td>L₁</td>
<td>VSA₁₁</td>
<td>VSA₁₂</td>
<td>...</td>
</tr>
<tr>
<td>L₂</td>
<td>VSA₂₁</td>
<td>VSA₂₂</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Lₖ</td>
<td>VSAₖ₁</td>
<td>VSAₖ₂</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 2. The VSA matrix

with VSAᵢⱼ = \( Lᵢ \otimes Lⱼ \), where \( Lᵢ \) and \( Lⱼ \) represent the ILI-codes of all synsets in which the target word \( W_{EN} \) occurs, and \( Lⱼ \) is the list of ILI-codes for all synsets in which the translation equivalent for the \( j \)th occurrence of \( W_{EN} \) occurs.

If no translation equivalent is found in language \( Lᵢ \) for the \( j \)th occurrence of \( W_{EN} \), VSAᵢⱼ is undefined; otherwise, it is a set containing 0, 1 or more ILI codes. For undefined VSAs, the algorithm cannot determine the sense number for the corresponding occurrence of the target word. However, it is very unlikely that an entire column in Table 2 is undefined, i.e., that there is no translation equivalent for an occurrence of the target word in any of the source languages.

When VSAᵢⱼ contains a single ILI code, the target word occurrence and its translation equivalent are assigned the same sense. For example, the VSA for the English-Romanian translation pair \(<\text{toe deget}>\) should contain the single ILI-code ENG20-0528265-n corresponding to sense 1 of toe in PWN and sense number 3 of deget in the Romanian wordnet. Thus the disambiguation of this translation pair would be \(<\text{toe(1) deget(3)}>\).

When the VSA set is empty—i.e., when none of the senses of the target word corresponds to an ILI code to which a sense of the translation equivalent was linked—the algorithm selects the pair in \( Lᵢ \otimes Lⱼ \) with the highest ss score. In case of ties, the pair corresponding to the most frequent sense of the target word in the current bitext pair is selected. If this heuristic in turn fails, the choice is made in favor of the pair corresponding to the lowest PWN2.0 sense number for the target word, since PWN senses are ordered by frequency. If no pair in \( Lᵢ \otimes Lⱼ \) meets the semantic similarity requirement, neither the occurrence of the target word nor its translation equivalent can be semantically disambiguated; but once again, it is extremely rare that there is no translation equivalent for an occurrence of the target word in any of the source languages.

When the VSA contains two or more ILI-codes, we have the case of cross-lingual ambiguity, i.e., two or more senses are common to the target word and the corresponding translation equivalent in the \( j \)th language. For example, at least two senses of the English word movement are identical to senses of the Romanian word miscare. In these cases, the heuristics applied in the case of ties are applied.

2 For details on the encoding of the corpus, see http://nl.ijs.si/ME/V2/msd/html/

3 In our earlier approach based on clustering, the columns of this table were the vectors used by the agglomerative clustering algorithm (see Ide, et al., 2002).
Back-off Method

For the method described in the previous section to succeed, aligned wordnets must be available for all languages in the parallel corpus. Furthermore, it is essential that the quality of the inter-lingual linking is high for all languages concerned. In cases where we cannot fulfill these requirements, we rely on a “back-off” method involving sense clustering based on translation equivalents, as discussed in (Ide, et al., 2002). This method clusters occurrences based on translation equivalents alone, thus eliminating reliance only on high-quality, aligned wordnets. We apply the clustering method after the wordnet-based method has been applied, and therefore each cluster containing an undisambiguated occurrence of the target word will also typically contain several occurrences that have already been assigned a sense. We can therefore assign the most frequent sense assignment in the cluster to previously unlabeled occurrences within the same cluster.

In the unlikely event that all occurrences of a given word which could not be disambiguated by the basic method are grouped together in clusters containing no previously labeled occurrences, we apply two heuristics: the first 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 only unlabeled occurrences, in decreasing order of their size, are joined to the first, second, etc., largest clusters containing occurrences already disambiguated (sense-assigned). The first heuristics assumes that the occurrences not disambiguated are used with a sense that was already used in the text. The second heuristic is applied only when no occurrence of the word in question has been disambiguated at all, as, for instances, in cases where there exists a single occurrence of the target word in the text, and the wordnet-based method has failed to disambiguate it. In this case the selected sense of the target word is the most frequent as recorded in PWN2.0. These heuristics are similar to those used for dealing with ties in the wordnet-based WSD algorithm.

Test Data and WSD Evaluation

In order to both evaluate the performance of the WSDtool and assess the accuracy of the interlingual linking of the Romanian wordnet to PWN2.0, we selected a bag of English target nouns, verbs, and adjectives extracted from the parallel corpus of George Orwell’s 1984 so that all their senses (at least two per POS) defined in PWN2.0 were also included and interlingually aligned in the Romanian wordnet. This set contained 211 words which had 1810 occurrences in 1385 sentences of the English part of the parallel corpus. To create a “gold standard” sense tagging for evaluation purposes, we manually sense-tagged all the occurrences of the 211 target words. We then enlisted 13 students enrolled in the Computational Linguistics Masters program at the University “A.I. Cuza” of Iași to manually assign senses to the same occurrences of the target words. An extraction script generated for each student a subset of the 1385 sentences containing occurrences of the targeted words. The extraction process ensured that the same sentence was in at least three student-sets. With each of the 1810 occurrences of the target words disambiguatated by at least three students, we computed a simple majority sense (MAJ) for each occurrence of the target words.

Disambiguation results for the same set of words were then generated by the WSDtool algorithm. The system was unable to make a decision for 398 of the 1810 occurrences, primarily in cases where the occurrence was not translated in the Romanian text or was incorrectly determined by the word-aligner. An evaluation program was then applied that generated a file containing detailed information for each of the 1810 occurrences, including

- the sense number for that occurrence in the gold standard (GS)
- the majority sense assigned by the student annotators (MAJ)
- the sense assigned by the algorithm (ALG)
- the names of the students who evaluated the occurrence and the sense(s) they assigned

For comparison purposes, we took into account only the 1412 occurrences that were sense disambiguated by the algorithm (without the back-off mechanism). Table 3 summarizes the results. It is interesting to note that the agreement between the algorithm and the gold standard is higher than between the majority vote of the students and the gold standard.

<table>
<thead>
<tr>
<th>GS=MAJ</th>
<th>GS=ALG</th>
<th>MAJ=ALG</th>
<th>GS=MAJ=ALG</th>
</tr>
</thead>
<tbody>
<tr>
<td>73.22%</td>
<td>78.68%</td>
<td>67.13%</td>
<td>62.32%</td>
</tr>
</tbody>
</table>

Table 3. WSD agreements (without back-off mechanism)

At the present time, the integration of the clustering algorithm with the WSD tool and back-off mechanism evaluation is not completed, and we therefore cannot report results for the fully-implemented method. A rough worst-case estimation of GS=ALG for the full implementation could be conjectured on the basis of the clustering accuracy we reported previously (~75%); therefore, we may assume that the accuracy for the combined WSDtool-cluster method would not be lower than 77%–78%.

Conclusions

Our disambiguation results, at the WN2.0 granularity level, using parallel resources, are (not surprisingly) superior to the state of the art in monolingual WSD because the knowledge embedded by the human translators into the parallel texts is of a tremendous help. Yet, 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. However, more and more parallel resources are becoming available, in particular on the World Wide Web (see for instance http://www.balkantimes.com where the same news is published in 10 languages), as well as a result of

In (Barbu, 2004) there are discussed later developments of our underlying word aligner that (for non-null alignments) has an error rate less than 11.5%. This error rate is largely due to English words occurring only once, or English words that are translated differently in each occurrence, so that the corresponding translation pairs are hapax legomena that are also not cognates.
the development of wordnets for an increasing number of languages. This opens up the possibility for application of our and similar methods to large amounts of parallel data in the not-too-distant future. One of the greatest advantages of applying such methods to parallel data is that it may be used to automatically sense-tag corpora in not only one language, but rather several at once. The resulting resources could provide substantial training data for monolingual WSD.

Acknowledgements

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

References