Word Sense Disambiguation: A Case Study on the Granularity of Sense Distinctions

DAN TUFIŞ
Research Institute for Artificial Intelligence
13, "13 Septembrie", Bucharest, 050711
ROMANIA

Abstract: The paper presents a method for word sense disambiguation (WSD) based on parallel corpora. The method exploits recent advances in word alignment and word clustering based on automatic extraction of translation equivalents and is supported by a lexical ontology made of aligned wordnets for the languages in the corpora. The wordnets are aligned to the Princeton Wordnet, according to the principles established by EuroWordNet. The evaluation of the WSD system was performed using three different granularity sense inventories.

Key-Words: Alignment, Lexical ontology, Parallel corpora, Sense inventories, Word sense disambiguation, wordnet

1 Introduction

Understanding natural language assumes one way or another, being able to associate to an ambiguous word (w) in a text or discourse the sense (s_k) which is distinguishable from other senses (s_1, …, s_{k-1}, s_{k+1}, …, s_n) potentially attributable to that word in a given context (c_i). Word-sense disambiguation is, by far, the most difficult part of the semantic processing required for natural language understanding. In a limited domain of discourse this problem is alleviated by considering only coarse-grained sense distinctions, relevant for the given domain. Such a solution, although computationally motivated with respect to the universe of discourse considered, has the disadvantage of reduced portability and is fallible when the meanings of words cross the boundaries of the prescribed universe of discourse. A general semantic lexicon, such as Princeton WordNet2.0[1] (henceforth PWN2.0), with word-senses labeled for specialized domains offers much more expressivity and power, reducing application dependency but, on the other hand posing the hard and challenging problem of contextual word-sense disambiguation. We describe a multilingual environment, relying on several monolingual wordnets, aligned to PWN2.0 used as an interlingual index (ILI), for word-sense disambiguation in parallel texts. The words of interest, irrespective of the language in the multilingual documents are disambiguated by using the same sense-inventory labels. The aligned wordnets were constructed in the context of the BalkaNet project [15]. BalkaNet is an European project that developed monolingual wordnets for five Balkan languages (Bulgarian, Greek, Romanian Serbian, and Turkish) and extended the Czech wordnet initially developed in the EuroWordNet project. The wordnets are aligned to the Princeton Wordnet (PWN2.0), taken as an interlingual index, following the principles established by the EuroWordNet consortium [8]. The version of the PWN2.0 used as ILI is an enhanced XML version where each synset is mapped onto a SUMO conceptual category [7] and is classified under one of the IRST domains [6].

The basic text we use for the WSD task is Orwell's novel "Ninety Eighty Four" (1984) and its translations in several languages (Bulgarian, Czech, Estonian, Greek, Hungarian, Romanian, Serbian, Slovene, and Turkish). All the translations were sentence-aligned to English, POS tagged and lemmatized. From the 9 bilingual alignments (EN-ZZ, where ZZ is one of the 9 translated versions of the original) we kept only the 1-1 sentence alignments. We discarded less than 7% from each monolingual text, but in return, due to the transitivity of the 1-1 sentence alignment we could sentence align (using the English hub) all the 10 texts. Thus, it is possible to extract any combination of two, three, up to 9 language sub-corpora of the parallel corpus with the initial alignment still valid (there are 1012 such parallel sub-corpora). Out of this sentence-aligned 10-language parallel corpus we retained only the English original, plus the Bulgarian, Czech and Romanian translations since for these languages we had wordnets with maximal

[1] http://www.cogsci.princeton.edu/~wn/

cross-lingual lexical coverage and the POS tagging and lemmatization in the respective language texts were consistent with Multext-East specification[2] [3].

The WSD experiments described in this paper start with the above mentioned sub-corpus, part-of-speech tagged, lemmatized and sentence aligned. All these preliminary processing are largely described elsewhere [10,12, 16, 16, 2].

The word sense disambiguation method described here has four steps:

a) word alignment of the parallel corpus and translation pairs extraction; this step results in good translation pairs (GT), bad translation pairs (BT) and word occurrences without identified translations (NT); many NTs are either happax legomena words or word occurrences that were not translated in the other language;
b) wordnet-based sense disambiguation of the translation pairs found (GT+BT) in step a); this step results in sense-assigned words (SAWO) for GT and sense-unasigned word occurrences (SUWO) for BT.
c) word sense clustering for NT and SUWO; this phase takes advantage of the sense assignment in step b).
d) generating the XML disambiguated parallel corpus with every content word (in each language) annotated with a single sense label. Sense label inventory can be one of the three available in the BalkaNet lexical ontology: PWN2.0 unique synset identifiers, SUMO conceptual categories and IRST-Domains.

For the evaluation purposes, we selected a set of fairly frequent English literals for which all of their senses (i.e., all of their synsets) are represented in the BalkaNet wordnets. They were disambiguated by three independent experts who negotiated the disagreements and thus created a gold-standard annotation for the evaluation of precision and recall of the automatic procedure.

2 Word Alignment and Translation Lexicon Extraction

Word alignment is a hard NLP problem which can be simply stated as follows: given $<T_{L1}\ T_{L2}>$ a pair of reciprocal translation texts, in languages L1 and L2, the word W_{L1} occurring in T_{L1} is said to be aligned to the word W_{L2} occurring in T_{L2} if the two words, in their contexts, represent reciprocal translations. In order to reduce the search space and to filter out significant information noise, the context is usually reduced to the level of sentence. Therefore, a parallel text $<T_{L1}\ T_{L2}>$ can be represented as a sequence of pairs of one or more sentences in language L1 ($S_{L1}^1\ S_{L1}^2 ... S_{L1}^k$) and one or more sentences in language L2 ($S_{L2}^1\ S_{L2}^2 \ldots S_{L2}^m$) so that the two ordered sets of sentences represent reciprocal translations. Such a pair is called a translation alignment unit (or translation unit). More often than not a translation unit is a 1 to 1 alignment, meaning that the content expressed in one sentence of L1 is fully contained in one sentence of L2. The n to m alignments are much rarer than the 1 to 1 alignments, especially in non-fiction texts. The word alignment of a bitext is an explicit representation of the pairs of words $<W_{L1}\ W_{L2}>$ (called translation equivalence pairs) co-occurring in the same translation units and representing mutual translations. The general word alignment problem includes the cases where words in one part of the bitext are not translated in the other part (these are called *null alignments*) and also the cases where multiple words in one part of the bitext are translated as one or more words in the other part (these are called expression alignments). The word alignment problem specification does not impose any restriction on the part of speech (POS) of the words making a translation equivalence pair, since cross-POS translations are rather frequent. However, for the aligned wordnet-based word sense disambiguation we discarded translation pairs which did not preserve the POS (and obviously null alignments). Removing duplicate pairs $<W_{L1}\ W_{L2}>$ one gets a translation lexicon for the given corpus.

A recent shared task evaluation of different word aligners, organized on the occasion of the NAACL Conference (www.cs.unt.edu/~rada/wpt) showed that word alignment may be solved quite reliably. Our winning system [13] produced relevant translation lexicons with an aggregated F-measure as high as 84.26%. Meanwhile, the word-aligner was further improved 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).

3 Wordnet-based Sense Disambiguation

Once the translation equivalents were extracted, then, for any translation pair $<W_{L1}\ W_{L2}>$ and two aligned wordnets, the algorithm performs the following operations:

[2] http://nl.ijs.si/ME/V3/

1. extract the interlingual (ILI) codes for the synsets that contain W^i_{L1} and W^j_{L2} respectively to yield two lists of ILI codes, $L^1_{ILI}(W^i_{L1})$ and $L^2_{ILI}(W^j_{L2})$
2. identify one ILI code common to the intersection $L^1_{ILI}(W^i_{L1}) \cap L^2_{ILI}(W^j_{L2})$ or a pair of ILI codes $ILI_1 \in L^1_{ILI}(W^i_{L1})$ and $ILI_2 \in L^2_{ILI}(W^j_{L2})$, so that ILI_1 and ILI_2 are the *most similar* ILI codes (defined below) among the candidate pairs $(L^1_{ILI}(W^i_{L1}) \otimes L^2_{ILI}(W^j_{L2})$ [$\otimes$ = Cartesian product].

The rationale for these operations derives from the common intuition which says that if the lexical item W^i_{L1} in the first language is found to be translated in the second language by W^j_{L2}, then it is reasonable to expect that at least one synset which the lemma of W^i_{L1} belongs to, and at least one synset which the lemma of W^j_{L2} belongs to, would be aligned to the same interlingual record or to two interlingual records semantically closely related.

Ideally step 2 above should identify one ILI concept lexicalized by W_{L1} in language *L1* and by W_{L2} in language *L2*. However, due to various reasons, the wordnets alignment might reveal not the same ILI concept, but two concepts which are semantically close enough to license the translation equivalence of W_{L1} and W_{L2}. This can be easily generalized to more than two languages. Our measure of interlingual concepts semantic similarity is based on PWN2.0 structure. We compute semantic-similarity[3] score by formula:

$ss(ILI_1, ILI_2) = 1/1+k$

where k is the number of links from ILI_1 to ILI_2 or from both ILI_1 and ILI_2 to the nearest common ancestor. The semantic similarity score is 1 when the two concepts are identical, 0.33 for two sister concepts, and 0.5 for mother/daughter, whole/part, or concepts related by a single link. Based on empirical studies, we decided to set the significance threshold of the semantic similarity score to 0.33. 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.

4 Word Sense Clustering Based on the Translation Lexicons

To perform the clustering, we derive for each target word *i* occurring *m* times in the corpus a set of *m* binary vectors *VECT*(TW^i). The number of cells in *VECT*(TW^i) is equal to the sum of distinct translations of the word *i* in all the other languages (called source languages) of the parallel corpus. The k^{th} VECT(TW^i) specifies which of the possible translations of TW^i were actually used in each language as an equivalent for the k^{th} occurrence of TW^i. All positions in the k^{th} *VECT*(TW^i) are set to 0 except at most one bit per source language identifying the word used (if any) as translation equivalent for the target word *i*.

For each target word *i* with *m* occurrences we fed the *m* vectors VECT(TW^i) into a hierarchical agglomerative clustering algorithm which produces clusters of similar vectors. Such a cluster would identify the occurrences of the target word which were used with the same meaning. The fundamental assumption of this algorithm is that if two or more instances of the same target word were translated the same in the source languages it is very likely that meaning is the same. The likelihood is increased as the number of source languages is larger and their types are more diversified. In our experiments we used four typologically distinct languages (Bulgarian, Czech, Greek, Romanian).

One big problem for the clustering algorithms in general and for agglomerative ones in particular is that the number of classes should be known in advance in order to obtain meaningful results. With respect to the word sense clustering this would mean knowing in advance for every word in a text how many of its possible senses are actually used in the given corpus. To overcome this lack of information, we use the results of the previous phase (word sense disambiguation based on the aligned wordnets) which generally successfully covers more than 75% of the text. For the words the occurrences were disambiguated by this phase we consider any other sense-unassigned occurrence was used with one of the previously seen senses, and thus we can provide the clustering algorithm with the number of classes. For all the words for which none of its occurrence was previously disambiguated (in the vast majority these words are happax legomena words) we automatically assign the first sense number in PWN2.0. The rationale for this heuristics is that in PWN2.0 senses are numbered according to the frequency (sense number one is the most frequent). This back-off mechanism is justified when the texts to be disambiguated are general texts, because PWN2.0 is a general semantic lexicon and the statistics on senses were drawn from a balanced corpus. For a specialized text, a more successful heuristics would be to take advantage of a prior classification of the text according to the IRST-

[3] Other approaches to similarity measures are described in [1].

domains and then to consider the most frequent sense with the same domain label as the one of the text. This is topic of future research and will build on our previous work [11] on document classification.

5 Generating the WSD-annotated corpus

The structure of the automatically generated WSD-annotated corpus is a simplified version of the XCES-ANA format[4] [5] with the additional attributes ***sn*** (sense number), ***oc*** (ontological category) and ***dom*** (domain) for the <w> tag. The values of these attributes have the following meanings:

- ***sn*** specifies the sense label for the current word as described in the wordnet of the respective language.
- ***oc*** represents the SUMO ontological concept(s) on which the wordnet sense of the current word is mapped on.
- ***dom*** identifies the IRST domain under which the wordnet sense of the current word is clustered.

The attributes ***sn***, ***oc*** and ***dom*** are specified only for words belonging to the content words (nouns, verbs - excluding auxiliaries, adjectives and adverbs). The use of all the three additional attributes is the default, but the user may specify one or two attributes to be generated in the WSD annotated parallel corpus. In Figure 1 below, we exemplify the default generated encoding (with the lexical token in bold characters and left-aligned).

```
<body>
    …
 <tu id="Ozz20">
    <seg lang="en">
       <s id="Oen.1.1.4.9">
          <w lemma="the" ana="Dd">
The</w>
          <w lemma="patrol" ana="Ncnp"
             sn="1"
             oc="SecurityUnit"
             dom="military">
patrols</w>
          <w lemma="do" ana="Vais">
did</w>
          <w lemma="not" ana="Rmp"
             sn="1"
             oc="not"
             dom="factotum">
not</w>
          <w lemma="matter" ana="Vmn
             sn="1"
             oc="SubjAssesmentAttribute"
             dom="factotum">
matter</w>
<c>,</c>
          <w lemma="however" ana="Rmp"
             sn="1"
             oc="SubjAssesmentAttribute|PastFn"
             dom="factotum">
however</w>
<c>.</c>
       </s>
    </seg>
    <seg lang="ro">
       <s id="Oro.1.2.5.9">
          <w lemma="şi" ana=Crssp>
Şi</w>
          <w lemma="totuşi" ana="Rgp"
             sn="1"
             oc="SubjAssesmentAttribute|PastFn"
             dom="factotum">
totuşi</w>
<c>,</c>
          <w lemma="patrulă" ana="Ncfpry"
             sn="1.1.x" oc="SecurityUnit"
             dom="military">
patrulele</w>
          <w lemma="nu" ana="Qz"
             sn="1.x" oc="not"
             dom="factotum">
nu</w>
          <w lemma="conta" ana="Vmii3p"
             sn="2.x"
             oc="SubjAssesmentAttribute"
             dom="factotum">
contau</w>
<c>.</c>
       </s>
    </seg>
    …
 </tu>
    …
</body>
```

Figure1: A sample of the WSD corpus encoding

6 The Experiment

The BalkaNet version of the "1984" corpus is encoded as a sequence of uniquely identified *translation units* (TU), each containing one sentence per language, so that they are reciprocal translations (see Figure 1). In order to evaluate both the performance of our WSD tool and to assess the accuracy of the interlingual linking of the BalkaNet wordnets we selected a bag of English target words (nouns and verbs) occurring in the corpus. The selection considered only polysemous words (at least two senses per part of speech) implemented (and ILI linked) in all BalkaNet wordnets. There resulted 211 words with 1644 occurrences in the English part of the parallel corpus.

Three experts sense-tagged (in terms of wordnet senses) all the occurrences of the target words and, the disagreements were negotiated until consensus

[4] http://www.cs.vassar.edu/XCES/

was obtained. The commonly agreed annotation represented the Gold Standard (GS) against which the WSD algorithm was evaluated. Once the values for the ***sn*** attributes have been established, the values for the ***oc*** and ***dom*** attributes are deterministically appended to the <w> tag annotation. The same targeted words were automatically disambiguated by our WSD tool algorithm (ALG) which was run both with and without the back-off clustering algorithm. For the basic wordnet-based WSD we used the Princeton Wordnet, the Romanian wordnet and the English-Romanian translation equivalence dictionary. For the back-off clustering we extracted a four language translation dictionary (EN-RO-CZ-BG) based on which we computed the initial clustering vectors for all occurrences of the target words.

Out of the 211 set of targeted words, with 1644 occurrences the system could not make a decision for 38 (18 %) words with 63 occurrences (3.83%). Most of these words were happax legomena (21) for which neither the wordnet-based step not the clustering back-off could do anything. Others, were not translated by the same part of speech, were wrongly translated by the human translator or not translated at all (28). Finally, four occurrences remained untagged due to the incompleteness of the Romanian synsets linked to the relevant concepts (that is the four translation equivalents had their relevant sense missing from the Romanian wordnet). Applying the simple heuristics (SH) that says that any unlabelled target occurrence receives its most frequent sense, 42 out of 63 of them got a correct sense-tag. The table below summarizes the results.

WSD System	Precision	Recall	F-measure
AWN	76.98%	76.18%	76.58%
AWN+C	79.48%	78.16%	78.81%
AWN+C+SH	78.74%	78.74%	78.74%
ST	72.99%	72.99%	72.99%

Table 1. WSD precision recall and F-measure for the algorithm based on aligned wordnets (AWN), for AWN with clustering (AWN+C) and for AWN+C and the simple heuristics (AWN+C+SH) and for the students' majority voting (ST)

Cross experiment evaluations of the WSD results are hard to compare when different granularity sense-inventory are used (PWN2.0: 115424 meanings vs. SUMO: 2066 categories, vs. Wordnet domains: 163). The Table 2 shows a great variation in terms of Precision and Recall when different granularity sense inventories are considered for the WSD problem. Therefore, it is important to make the right choice of the sense inventory to be used with respect to a given application.

Sense Inventory	Precision	Recall	State-of-the-Art
PWN 115424 categories	79.48%	78.74%	65%-68%
SUMO/MILO 2066 categories	87.16%	87.16%	80%-81%
IRST DOMAINS 163 categories	92.78%	92.78%	85%-88%

Table 2. Evaluation of the WSD task in terms of three different sense inventories

For instance, in case of a document classification problem, it is very likely that the IRST domain labels would suffice. The rationale is that IRST domains are directly derived from the Universal Decimal Classification as used by most libraries and librarians. The SUMO sense labeling will be definitely more useful in an ontology based intelligent system interacting through a natural language interface. Finally, the most refined sense inventory of PWN2.0 will be extremely useful in Natural Language Understanding Systems, which would require a deep processing. Also, such a fine inventory would be highly beneficial in lexicographic and lexicological studies.

7 Conclusion and further work

Considering the fine granularity of the PWN2.0 sense inventory, our disambiguation results using parallel resources are superior to the state of the art in *monolingual* WSD (with the same sense inventory). This is not surprising since the parallel texts contain implicit knowledge about the sense of an ambiguous word, which has been provided by human translators. The drawback of our approach is that it relies on the existence of parallel data, and aligned wordnets which in the vast majority of cases are not available. On the other hand, supervised monolingual WSD relies on the existence of large samples of training data, and our method can be applied to produce such data to bootstrap monolingual applications. Given that parallel resources are becoming increasingly available, in particular on the World Wide Web (see for instance www.balkantimes.com where the same news is published in 10 languages), and aligned wordnets are being produced for more and more languages, it should be possible to apply our and similar methods

to large amounts of parallel data in the not-too-distant future.

We plan to conduct experiments and comparisons on word alignment replacing our translation equivalence model with the one generated IBM model 5 as constructed by GIZA++ [9], available at http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html.

References

[1] A. Budanitsky and G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second meeting of the NAACL, Pittsburgh, June.*

[2] Erjavec T., Ide N., Tufiş, D., Automatic Sense Tagging Using Parallel Corpora, in *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, pp. 212-219, 2001

[3] Erjavec, T, MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora, in *Proceedings of LREC2004*, Lisbon, 2004

[4] Ide, N., Erjavec, T., Tufiş, D. Sense Discrimination with Parallel Corpora. In *Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, 56-60, 2002.

[5] Ide, N., Bonhomme, P., Romary, L., XCES: An XML-based Standard for Linguistic Corpora, in *Proceedings of LREC2000*, Athens, Greece, 825-30, 2000.

[6] Magnini B. Cavaglià G., Integrating Subject Field Codes into WordNet, in *Proceedings of LREC2000,* Athens, Greece, 2000.

[7] Niles, I., and Pease, A., Towards a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS-2001), Ogunquit, Maine, October 17-19, 2001.

[8] Peters, W., Vossen, P., Diez-Orzas, P., Adriaens, G., Cross-Linguistic Alignment of wordnets with an Inter-Lingual-Index, in P. Vossen (Ed.) *Special Issue on EuroWordNet, Computers and the Humanities*, 32, nos.2-3, 221-251, 1998.

[9] Och, F., J., Ney, H., Improved Statistical Alignment Models, *Proceedings of ACL2000*, Hong Kong, China, 440-447, 2000.

[10] Tufiş, D., Tiered Tagging and Combined Classifiers, in F. Jelinek, E. Nöth (eds) *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence* 1692, Springer, 28-33, 1999.

[11] Tufiş, D., Popescu, C., Roşu, R.: Automatic Classification of Documents by Random Sampling, in *Proceedings of The Romanian Academy* Series A, Volume 1, Number 2/2000, pp.117-127, 2000.

[12] Tufiş, D., A cheap and fast way to build useful translation lexicons. In *Proceedings of the 19th International Conference on Computational Linguistics*, COLING 2002, Taipei, 1030-1036, 2002.

[13] Tufiş, D., Barbu, A., M., Ion, R. 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*, Edmonton, 36-39, 2003.

[14] Tufiş, D., Ion, R., Ide, N., Word sense disambiguation as a wordnets validation method in Balkanet. *Proceedings of LREC'2004*, Lisbon, 741-744, 2004.

[15] Tufiş, D., Cristea, D., Stamou, S., BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In D. Tufiş (ed): *Special Issue on BalkaNet. Romanian Journal on Science and Technology of Information,* 7 (3-4): 9-44, 2004.

[16] Tufiş, D., Ion, R. Interlingual wordnets validation and word-sense disambiguation. In *Proceedings of the Natural Language Understanding and Cognitive Systems* NLUCS, Porto, 97-105, 2004.

[17] Tufiş, D, Ion, R., Ide, N., 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*, COLING2004, Geneva, 1312-1318, 2004.