Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets

Dan TUFIŞ
Institute for Artificial Intelligence
13, "13 Septembrie"
Bucharest, 050711
Romania
tufis@racai.ro

Radu ION
Institute for Artificial Intelligence
13, "13 Septembrie"
Bucharest, 050711
Romania
radu@racai.ro

Nancy IDE
Department of Computer Science, Vassar College
Poughkeepsie,
NY 12604-0520
USA
ide@cs.vassar.edu

Abstract

The paper presents a method for word sense disambiguation based on parallel corpora. The method exploits recent advances in word alignment and word clustering based on automatic extraction of translation equivalents and being supported by available aligned wordnets for the languages in the corpus. The wordnets are aligned to the Princeton Wordnet, according to the principles established by EuroWordNet. The evaluation of the WSD system, implementing the method described herein showed very encouraging results. The same system used in a validation mode, can be used to check and spot alignment errors in multilingually aligned wordnets as BalkaNet and EuroWordNet.

1 Introduction

Word Sense Disambiguation (WSD) is well-known as one of the more difficult problems in the field of natural language processing, as noted in (Gale et al, 1992), (Kilgarriff, 1997), (Ide and Véronis, 1998), and others. The difficulties stem from several sources, including the lack of means to formalize the properties of context that characterize the use of an ambiguous word in a given sense, lack of a standard (and possibly exhaustive) sense inventory, and the subjectivity of the human evaluation of such algorithms. To address the last problem, (Gale et al, 1992) argue for upper and lower bounds of precision when comparing automatically assigned sense labels with those assigned by human judges. The lower bound should not drop below the baseline usage of the algorithm (in which every word that is disambiguated is assigned the most frequent sense) whereas the upper bound should not be "too restrictive" when the word in question is hard to disambiguate even for human judges (a measure of this difficulty is the computation of the agreement rates between human annotators).

Identification and formalization of the determining contextual parameters for a word used in a given sense is the focus of WSD work that treats texts in a monolingual setting—that is, a setting where translations of the texts in other languages either do not exist or are not considered. This focus is based on the assumption that for a given word w and two of its contexts C_1 and C_2, if $C_1 \equiv C_2$ (are perfectly equivalent), then w is used with the same sense in C_1 and C_2. A formalized definition of context for a given sense would then enable a WSD system to accurately assign sense labels to occurrences of w in unseen texts. Attempts to characterize context for a given sense of a word have addressed a variety of factors:

- *Context length*: what is the size of the window of text that should be considered to determine context? Should it consist of only a few words, or include much larger portions of text?
- *Context content*: should all context words be considered, or only selected words (e.g., only words in a certain part of speech or a certain grammatical relations to the target word)? Should they be weighted based on distance from the target or treated as a "bag of words"?
- *Context formalization*: how can context information be represented to enable definitions of an inter-context equivalence function? Is there a single representation appropriate for all words, or does it vary according to, for example, the word's part of speech?

The use of multi-lingual parallel texts provides a very different approach to the problem of context identification and characterization. "Context" now becomes the word(s) by which the target word (i.e., the word to be disambiguated) is translated in one or more other languages. The assumption here is that different senses of a word are likely to be lexicalized differently in different languages; therefore, the translation can be used to identify the correct sense of a word. Effectively, the translation captures the context as the translator conceived it.

The use of parallel translations for sense disambiguation brings up a different set of issues, primarily because the assumption that different senses are lexicalized differently in different languages is true only to an extent. For instance, it is well known that many ambiguities are preserved across languages (for example, the French *intérêt* and the English *interest*), especially languages that are relatively closely related. This raises new questions: how many languages, and of which types (e.g., closely related languages, languages from different language families), provide adequate information for this purpose? How do we measure the degree to which different lexicalizations provide evidence for a distinct sense[1]?

We have addressed these questions in experiments involving sense clustering based on translation equivalents extracted from parallel corpora (Ide et al., 2001), (Ide et al., 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 English that is not bound to any particular sense inventory. Our experiments confirm that the accuracy of word sense clustering based on translation equivalents is heavily dependent on the number and diversity of the languages in the parallel corpus and the language register of the parallel text. For example, using six source languages from three language families (Romance, Slavic and Finno-Ugric), sense clustering of English words was approximately 75% accurate; when fewer languages and/or languages from less diverse families are used, accuracy drops dramatically. This drop is obviously a result of the decreased chances that two or more senses of an ambiguous word in one language will be lexicalized differently in another when fewer languages, and languages that are more closely related, are considered.

To enhance our results, we have explored the use of additional resources, in particular, the aligned wordnets in BalkaNet. BalkaNet is a European project that is developing monolingual wordnets for five Balkan languages (Bulgarian, Greek, Romanian Serbian, and Turkish) and improving the Czech wordnet developed in the EuroWordNet project. The wordnets are aligned to the Princeton Wordnet, following the principles established by the EuroWordNet consortium. The underlying hypothesis in this experiment exploits the common intuition that reciprocal translations in parallel texts should have the same (or closely related) interlingual meanings (in terms of BalkaNet, ILI record-projections or simply ILI codes). However, this hypothesis is reasonable if the monolingual wordnets are reliable and correctly linked to the interlingual index (ILI). Quality assurance of the wordnets is a primary concern in the BalkaNet project, and to this end, the consortium developed several methods and tools for validation, described in various papers authored by BalkaNet consortium members (see Proceedings of the Global WordNet Conference, Brno, 2004).

We previously implemented a language-independent disambiguation program, called WSDtool, which has been extended to serve as a multilingual wordnet checker and specialized editor for error-correction. In (Tufis, *et al*., 2004) it was demonstrated that the tool detected several interlingual alignment errors that had escaped human analysis. In this paper, we describe a disambiguation experiment that exploits the ILI information in the corrected wordnets

2 The Basic Methodology

Our methodology consists of the following basic steps:

1. given a bitext T_{L1L2} in languages L_1 and L_2 for which there are aligned wordnets, extract all pairs of lexical items that are reciprocal translations: $\{<W^i_{L1}\ W^j_{L2}>^+\}$
2. for each lexical alignment $<W^i_{L1}\ W^j_{L2}>$, extract the 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 $L^2_{ILI}(W^j_{L2})$
3. 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 accuracy of step 1 is essential for the success of the validation method. The word alignment problem includes cases of *null alignment,* where words in one part of the bitext are not translated in the other part; and cases of *expression alignment,* where multiple words in one part of the bitext are translated as one or more words in the other part. Word alignment algorithms typically do not take into account the part of speech (POS) of the words comprising a translation equivalence pair, since cross-POS translations are rather frequent. However, for the aligned wordnet-based word sense disambiguation we discard both translation pairs which do not preserve the POS and null alignments. Multiword expressions included in a wordnet are dealt with by the underlying tokenizer.

[1] See (Ide, 1999), for an extended discussion

Therefore, we consider only one-to-one, POS-preserving alignments.

If for any translation equivalence pair <W_{L1} W_{L2}> the following conditions hold true:

- wordnet WN_{L1} contains *literal*(W_{L1}) and wordnet WN_{L2} contains *literal*(W_{L2}) where the *literal*(W) function transforms the occurrence form of W to its lemma form, and
- all possible senses of *literal*(W_{L1}) are present in WN_{L1} and all possible senses of *literal*(W_{L2}) are present in WN_{L2}, and
- wordnets WN_{L1} and WN_{L2} are linked through an ILI-like mechanism

then, a bilingual WSD algorithm should ideally output one ILI code that stands for the same concept lexicalized by W_{L1} in language *L1* and by W_{L2} in language *L2*. This can be easily generalized to more than two languages.

The second condition above, requiring that all senses of both words are included in both wordnets is unrealistic, and must be relaxed. (Kilgarriff, 1997) states that:

> *"... a (WSD) task-independent set of word senses for a language is not a coherent concept. Word senses are simply undefined unless there is some underlying rationale for clustering, some context which classifies some distinctions as worth making and others as not worth making."*

None of the BalkaNet wordnets is lexically dense (see (Tufiş et al., 2004a)) meaning that although the literals in a translation pair could be present in the wordnets of interest, not all their senses (as glossed in a reference explanatory dictionary) are included.

For this experiment, we selected a set of fairly frequent English literals for which all senses (i.e., all of their synsets) are represented in the BalkaNet wordnets, thus ensuring that at least one synset containing the translation and one synset in PWN with the same ILI-code should exist. This experiment involves the Romanian-English language pair. We also treat PWN2.0 as a BalkaNet wordnet, such that ILI is regarded as a bag of identifiers (code) representing the interlingual concepts. In the XML encoding of the BalkaNet wordnets (including PWN2.0) every synset has a unique ID, the value of which is one of the labels in ILI. Thus although our experiment treats only Romanian and English, the method is identical for any language pair.

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 1 may be solved quite reliably. Our system (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 so that current performance on the same data is 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).

Step 2 is where the aligned wordnets come to work. The correctness of the interlingual alignment is essential to find a pair of ILI codes that disambiguate translation equivalents.

Our measure of ILI similarity is based on the principle of *hierarchy preservation* (Tufis & Cristea, 2002), which asserts that the hierarchies induced by semantic relations in two aligned wordnets should be topologically similar. According to this conjecture, *relatedness* (rel) of two ILI records R_1 and R_2 is a measure of *semantic-similarity* (ss) between two synsets Syn_1 and Syn_2 in PWN2.0 that correspond to R_1 and R_2. We compute semantic-similarity by

$$ss(Syn_1, Syn_2) = 1/1+k$$

where k is the number of links from Syn_1 to Syn_2 or from both Syn_1 and Syn_2 to the nearest common ancestor. The semantic similarity is 1 when the two synsets are identical (or have the same ILI code), .33 for two sister synsets, and 0.5 for mother/daughter, whole/part, or synsets related by a single link.

Two ILI records R_1 and R_2 are considered closely related if: $rel(R_1, R_2) = ss(Syn_1, Syn_2) \geq t$

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, L_1, $L_2 \ldots L_k$), where for the purposes of disambiguation T is the target language and L_1, $L_2 \ldots L_k$ 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 (for more details on the encoding see: http://nl.ijs.si/ME/V2/msd/html/).

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 (eq_{ij} represents the translation equivalent in the i^{th} source language of the j^{th} occurrence of the word in the target language). If the target word is not translated in language L_i, eq_{ij} is represented by the null string. The table is generated as a result of step 1, as described in the previous section.

	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 1. The translation equivalents matrix (EQ matrix)

Step 2 transforms the matrix in Table 1 to a matrix with the same dimensions (Table 2) called VSA (Validation and Sense Assignment):

	Occ #1	Occ #2	…	Occ #n
L_1	VSA_{11}	VSA_{12}	..	VSA_{1n}
L_2	VSA_{21}	VSA_{22}		VSA_{22}
..	…	…	..	…
L_k	VSA_{k1}	VSA_{k2}	..	VSA_{kn}

Table 2. The VSA matrix

with $VSA_{ij} = L^{EN}_{ILI}(W_{EN}) \cap L^{i}_{ILI}(W^{j}_{Li})$, where $L^{EN}_{ILI}(W_{EN})$ represent the ILI-codes of all synsets in which the target word W_{EN} occurs, and $L^{i}_{ILI}(W^{j}_{Li})$ 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 L_i for the j^{th} occurrence of W_{EN}, VSA_{ij} 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 VSA_{ij} contains a single ILI code, the target word occurrence and its translation equivalent are assigned the same sense.

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 $L^{EN}_{ILI}(W_{EN}) \otimes L^{i}_{ILI}(W^{j}_{Li})$ 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 $L^{EN}_{ILI}(W_{EN}) \otimes L^{i}_{ILI}(W^{j}_{Li})$ 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 i^{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.

2.1 Back-off mechanism

In the previous section it was noted that when no solution is provided by the ILI method, we may get the information from a VSA corresponding to the same occurrence of the target word but in a different language. However, this demands that aligned wordnets are available for all languages in the parallel corpus, and 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). 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.The advantages of such a combined approach are:

- it eliminates reliance only on high-quality, k+1 aligned wordnets. Indeed, having k+1 languages in our corpus, we need only apply the WSD method to the aligned wordnets for the target and one source language and alignment lexicons from the target language to every other language in the corpus. The bilingual setting (target language – source language) would ensure the applicability of the WSD procedure and the clustering heuristic would apply a uniform sense labeling among translation equivalents belonging to the same cluster.
- it can reinforce or modify the sense assignment for every translation equivalence pair that fall into the cases b) and c), and will be able to assign a sense for all translation pairs falling into case d), which the previous algorithm could not do; all non-disambiguated members

of one cluster will be disambiguated according to the majority sense of the already disambiguated members of the cluster;

The clustering algorithm we used is derived from (Stolke, 1996). In what follows, an updated description of back-off clustering algorithm is given.

First, a few notations are in order:

1. TWL=$\{TW^i\}_{1 \le i \le n}$, the Target Word List;
2. TW^i_k, the *k*-th occurrence of TW^i;
3. $DEL(L_p,TW^i)=\{W^j \mid <TW^i, W^j>$ is a translation equivalence pair$\}$, the Dictionary Entry List. This is the *ordered* list of all the translation equivalents in the source language L_p of the target word TW^i. These translation equivalents were automatically extracted from the parallel corpus using a hypotheses testing algorithm which is described at length in (Tufiş and Barbu, 2002);
4. $|DEL(L_p,TW^i)|$ = the number of elements in $DEL(L_p,TW^i)$;
5. $TEQ(L_p,TW^i_k)$ = the Translation EQuivalent in language L_p for the *k*-th occurrence of TW^i, $TEQ(L_p,TW^i_k) \in DEL(L_p,TW^i)$;
6. $DEL_h(L_p,TW^i)$ = the *h*-th element of $DEL(L_p,TW^i)$;
7. $LVECT(L_p,TW^i_k)$= a binary vector of $|DEL(L_p,TW^i)|$ positions; all the binary positions are 0 except at most one bit at position h wich is 1 if $TEQ(L_p,TW^i_k)=DEL_h(L_p,TW^i)$. This binary vector specifies for the language L_p which of the possible translations of TW^i was actually used as a translation equivalent for the k^{th} occurrence of TW^i.
8. $VECT(TW^i) = CON_{p=1,S}\ (LVECT(L_p,TW^i_k))$, with CON a vector concatenation operator and *S* the number of source languages in the parallel corpus.

Having a set of *m* binary vectors $VECT(TW^i)$ for one target word occurring *m* times in the corpus, we use a Hierarchical Clustering Algorithm based on Stolcke's Cluster2.9 to classify similar vectors in into sense classes.

The algorithm progresses as follows:

- **Input**: define *m* classes, each containing one $VECT(TW^i_k)$ binary vector $(1 \le k \le m)$ and for each class compute the centroid; initially the centroid of the class k is the vector $VECT(TW^i_k)$;
- **Processing phase**: compute the minimum distance among the centroids of any pairs of classes and cluster together the classes with the minimal distance; the distance we use is a Euclidean distance in a *n*-dimensional space (here v_1 and v_2 are the centroids of the classes between which the distance is computed):

$$D = \sqrt{\sum_{i=1}^{n} (v_1(i) - v_2(i))^2} \qquad (3)$$

The centroid v_r of the new class is a weighted mean of the centroids of the two clustered classes; the cell values of the centroid vector are computed as shown in (4) where $size(v_1)$ and $size(v_2)$ represent the numbers of elements in the two clustered classes respectively:

$$v_r(i) = \frac{v_1(i)size(v_1) + v_2(i)size(v_2)}{size(v_1) + size(v_2)} \qquad (4)$$

At each processing step the number of classes decreases by 1 and, obviously, $size(v_r)=size(v_1)+size(v_2)$.

- **Exit condition**: without any restriction, the algorithm stops when everything has been clustered into a single class. Tracing the clustering operations produces a binary tree with the initial *m* vectors $VECT(TW^i_k)$ as leafs; an *interior cut* in the clustering tree will produce a specific number (say X) of sub-trees, the roots of which stand for X classes each containing the vectors of their leaves. An interior cut is called a *pertinent cut* if X is equal to the number of senses TW^i has been used throughout the entire corpus. One should note that in a clustering tree many pertinent cuts could be possible. The pertinent cut which corresponds to the correct sense clustering of the *m* occurrences of TW^i is called a *perfect cut*. However, assuming TW^i has Y possible senses, unfortunately, one cannot predict how many of them will be used in an arbitrary text. Therefore, a pertinent cut in a clustering tree cannot be deterministically computed. Instead of deriving the clustering tree and trying to guess a perfect cut, we stop the clustering algorithm when there have been created Z clusters, where, idealistically, Z should be the number of senses in which the *m* occurrences of TW^i have been used. The Z number is specific to each word and depends on the type and size of the texts in which the respective word appears, so it cannot be a-priori computed. To overcome this indeterminism we used as an exit condition for the clustering algorithm (thus a way of computing Z) a distance heuristics. When the minimal distance between the existing classes increases *too much*, then the algorithm should stop. This

fuzzy statement was turned into the exit condition as shown in (5):

$$\frac{dist(k+1) - dist(k)}{dist(k+1)} > \alpha \qquad (5)$$

where $dist(k)$ is the minimal distance between two clusters at the *k*-th iteration step and α is an empirical numerical threshold. After numerous experiments we set α to 0.12. Although the threshold is a parameter for the clustering algorithm, irrespective of the target words, the number of classes that the clustering algorithm generates (Z value) is still dependent on the particular target word and the corpus in which it appears.

The combination of the aligned wordnets based WSD and the clustering algorithm can be extended so that it is possible to drop (5) as the clustering exit condition. One possible way to state the clustering exit condition is to prohibit joining classes that contain occurrences already sense-labeled unless the sense-labels are identical. The common sense for all unlabeled occurrences in a cluster will be imported from the sense-labeled occurrences in the same cluster. However, this approach is very sensitive to the accuracy of the aligned wordnets based WSD, since if two occurrences that should have the same label are incorrectly labelled differently, there is no chance they will be clustered together. In our approach, the final sense labeling, based on majority voting, favors the clustering algorithm; if the clustering algorithm is wrong, a correct sense label could be overridden and changed to an incorrect one.

3 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 ILI linked) 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). A number of 13 students, enrolled in the Computational Linguistics Master program, were asked to manually sense-tag the occurrences of the target words occurring in a set of assigned sentences. An extraction script ensured that the same sentence was in at least three student-sets. Out of the students' hand disambiguated occurrences a simple majority sense assignment 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 (about 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. A number of 34 occurrences were not disambiguated because of wordnet alignment errors. In this experiment we didn't use the back-off mechanism because we were mainly interested in the Romanian wordnet accuracy (interlingual alignment correctness, synset completeness). The back-off mechanism would have obliterated occurrences that had been sense-tagged by classification and not by wordnet alignment. 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
73.22 %	78.68 %	67.13%	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 WSDtool 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 78-79%. 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.

4 Conclusion

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 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. If we note that there is a considerably large number of literals with a single sense in PWN (119528 out of 145627 which means approximately 82%), we see that the WSD method proposed here can almost have a full coverage if we extend it by saying that every translation pair for which there is a single sense in its English part (as extracted from PWN) receives that sense. The resulting resources could provide substantial training data for monolingual WSD.

This WSD algorithm can be applied only on parallel corpora. We can see that the Internet becomes each day a richer depositary of documents published in several languages (see for instance http://www.balkantimes.com where the same news are 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 to face the monolingual WSD problem much better equipped in terms of training data.

5 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

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

N. Ide, T. Erjavec, D. Tufiş. 2001. Automatic Sense Tagging Using Parallel Corpora. In "*Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*", pages 212-219, Tokyo.

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

A. Stolcke.1996. http:// ley.edu/ftp/global/pub/ai/ stolcke/software/cluster-2.9.tar.Z.

D. Tufiş, D. Cristea. 2002. Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet. In "*Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation*", pages 3--41, Las Palmas.

D. Tufiş, A. M. Barbu. 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.

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

D. Tufiş, R. Ion. 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*", pages 13-26, Bucharest.

D. Tufiş, A.M. Barbu, R. Ion. 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*", pages 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*", pages 332-340, Brno.

Tufiş, D., Cristea, D., Stamou, S. (2004a). 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