ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 183–200

Multilingual Word Sense Disambiguation Using Aligned Wordnets

Radu ION, Dan TUFIŞ

Research Institute for Artificial Intelligence, Romanian Academy
E-mail: {radu, tufis}@racai.ro

Abstract. Word Sense Disambiguation (WSD from now on) represents an established task within Natural Language Processing community, aiming at finding the right sense of a word occurring in a free running text through the use of a computer algorithm. Currently, most of the WSD approaches consider only monolingual texts, and, as such, they rely mainly on the discriminatory power of the words appearing in the same context with the targeted words (the words to be disambiguated). We propose a more precise WSD method, based on parallel texts, and relying on interlingually aligned wordnets. The method exploits recent advances in translation equivalents extractions and word alignment.

1. Introduction

Word Sense Disambiguation (WSD) is well-known as one of the most difficult problems in the field of natural language processing, as noted in [3], [5], [6], 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, [3] 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. To put it differently, 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 of the same word 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 in 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 [8], [9]. [17] 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.

[1]See [7], for an extended discussion.

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.

To enhance our results, we have explored the use of additional resources, in particular, the aligned wordnets in BalkaNet (Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish). The wordnets are aligned to the version 2.0 of Princeton WordNet [2], PWN 2.0 henceforth, 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).

The paper's organization is as follows: in section 2 a mapping methodology overview is presented along the general lines of Romanian wordnet development and its linking to the PWN 2.0, section 3 defines the WSD problem within the current bi- and multi-lingual setting, section 4 gives a technical description of our WSD procedure, section 5 discusses improvements to the basic WSD procedure, section 6 presents an evaluation of the WSD experiments, section 7 discusses the implementation and its adequacy for semantic validation of interlingually aligned wordnets developed in the BalkaNet project and finally, section 8 draws some conclusions and discusses further work.

2. Matching the Concepts between Wordnets

One of the main aims of the BalkaNet project was to ensure as much cross-lingual coverage as possible. This assumes not only the development of appropriate synsets corresponding to a commonly agreed set of concepts but also ensuring the correctness of the monolingual synsets' interlingual linking to the ILI. The WSD disambiguation method described here relies on the existence of a set of wordnets which are cross-lingually aligned via an Inter-lingual Index. When the BalkaNet project started, ILI was "inherited" from EuroWordNet (EWN henceforth) [19] as an unstructured collection of language neutral records, with the property that each such record (called a concept) was linguistically realized as a synset in at least one language represented in EWN. The ILI was built from PWN 1.5 with later additions of concepts lexicalized in some other languages. Each concept had an English gloss attached, irrespective the language it came from, and a few of them, called Base Concepts, were associated with a top ontology description [10]. As the BalkaNet project developed, it became apparent that a more effective way to keep up with the PWN developments was to add a commonly agreed structure to the ILI thus enabling more efficient implementations

of cross-lingual applications. The easiest way to do it was to adopt PWN itself as an interlingual index and to keep in the monolingual wordnets whatever structural departures from ILI. As in EWN, the BalkaNet ILI may contain concepts motivated by other languages than English. Further details and motivations on this issue are given in the overall presentation of the project, included into this volume [12].

One strong requirement is that every synset in a monolingual wordnet is linked to only one ILI concept and no two different synsets of a monolingual wordnet point to the same ILI concept. The individual approaches in the development of the BalkaNet wordnets and their alignment to ILI are largely described in various papers of this volume and therefore we will not go into further details. The interlingual alignment raised problems to all teams involved in the project but the difficulties were to a large extent dependent on the development methodologies adopted in the case of each monolingual wordnet. As far as the Romanian wordnet is concerned, the main difficulties in its construction and in the ILI linking are described in details elsewhere in this volume [13], [1].

While checking the correctness of the individual wordnets was to a large extent in charge of each development team, the cross-lingual validation was a special concern of the consortium for which a common multilingual corpus-based approach has been agreed. The rationale for using parallel texts, translated by professionals in all the languages of the consortium, was to evaluate the cross-lingual coverage of our wordnets against real running texts and see to what degree the monolingual synsets reflect real languages usages. In the context of BalkaNet project, each monolingual wordnet (say Romanian) linked to the ILI (PWN 2.0) represent a bilingual lexical resource which can be objectively evaluated against a parallel corpus, provided that a reliable word alignment method is available. Such an evaluation exercise turns into a WSD task with correct wordnets alignment resulting in a simultaneous sense disambiguation in both languages, with a common sense inventory at the fine granularity level of PWN 2.0. The next section gives a brief formulation of the WSD problem in the context of cross-lingually aligned wordnets.

3. Bilingual and Multilingual WSD: Problem Statement

Word alignment is a hard NLP problem which can be simply stated as follows: given $\langle T_{L1}\ T_{L2}\rangle$ 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. The pair $\langle W_{L1}\ W_{L2}\rangle$ is called a translation equivalence pair. 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 would discard translation pairs which do not preserve the POS (and obviously null alignments). Multiword expression

supposed to be found in a wordnet are dealt with as single lexical items and therefore we will consider only one-to-one POS-preserving alignments.

If for any translation equivalence pair $\langle W_{L1}\ W_{L2}\rangle$ the following conditions hold true:

- the wordnet WN_{L1} contains $literal(W_{L1})$ and the wordnet WN_{L2} contains $literal(W_{L2})$ where the $literal(W)$ function transforms the occurrence form of W to its lemma form,
- 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
- the 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 thus obtaining the multilingual WSD problem statement.

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."

Moreover, none of the BalkaNet wordnet is lexically dense (see [12], [13] in this volume) 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 implemented.

Considering all these, 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. This way we ensured that no matter what the translation was for a target English word, there *should exist* at least one synset containing the translation and one synset in PWN with the same ILI-code. In what follows, our method will be exemplified considering the Romanian-English pair of languages. We also treat PWN 2.0 as a BalkaNet wordnet, such that ILI is regarded as a bag of identifiers (codes) representing the interlingual concepts. In the XML encoding of the BalkaNet wordnets (including PWN 2.0) every synset has a unique ID, the value of which is one of the labels in ILI (see [4] in this volume for the XML encoding schema). Thus although the methodology is exemplified for Romanian and English, it remains the same for any language combination irrespective of whether English is one of them or not.

4. WSD as Sets of ILI Codes Intersection

The methodology for 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: $\{\langle W^i_{L1}\ W^j_{L2}\rangle\ ^+\}$

B) for each lexical alignment of interest, $\langle W^i_{L1}\ W^j_{L2}\rangle$, one extracts, for each language, the ILI codes for the synsets that contain $literal(W^i_{L1})$ and $literal(W^j_{L2})$ respectively; thus, one gets two lists of ILI codes, $L^1_{ILI}(W^i_{L1})$ and $L^2_{ILI}(W^j_{L2})$, one for each language. The WSD of the lexical items under consideration comes to identify one ILI code common to the intersection $L^1_{ILI}(W^i_{L1}) \cap L^2_{ILI}(W^j_{L2})$ or a pair of ILI codes $\text{ILI}_1 \in L^1_{ILI}(W^i_{L1})$ and $\text{ILI}_2 \in L^2_{ILI}(W^j_{L2})$ so that ILI_1 and ILI_2 are the codes of the most similar ILI concepts (below we elaborate on this issue) among the candidate pairs $(L^1_{ILI}(W^i_{L1}) \otimes L^2_{ILI}(W^j_{L2})$ with $\otimes$ 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 [15] produced relevant lexicons for wordnets evaluation 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).

The **B)** step is where the aligned wordnets come to work. The correctness of the interlingual alignment is essential in finding a pair of ILI codes that disambiguate the translation equivalents.

However, since we considered here (as in the EuroWordNet) the ILI as an unstructured set of labels denoting interlingual concepts, we need to clarify what "codes of the most similar ILI concepts" means. In the context of this research, we assume that the *hierarchy preservation principle* [16] is sound. Under this assumption, we take the *similarity* of two ILI codes $R1$ and $R2$ as a measure for the *semantic-similarity* between the synsets *Syn1* and *Syn2* in PWN 2.0 that correspond to $R1$ and $R2$. We used a very simple definition of the semantic similarity between two synsets:

$$\textit{semantic-similarity}(\textit{Syn1, Syn2}) \ = \ \frac{1}{1+N}, \tag{1}$$

where N 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, equivalently said, they 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 similar if

$$\textit{similarity}(R1, R2) = \textit{semantic-similarity}(\textit{Syn1}, \textit{Syn2}) \geq t, \tag{2}$$

where t 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).

We should note at this point that *similarity* is meant as a language independent score which is approximated by an English specific score. This is justified, irrespective of ILI being structured or not, because the general model for all the wordnets was the PWN. Yet, since in BalkaNet ILI is PNW 2.0 (thus, structured) the two measures are identical and, when the WSD task is considered among a pair of languages that includes English, the distinction we made seems a useless complication. However, if we consider the WSD task on a Czech-Romanian bitext, for instance, it is very likely that the topologies of the Czech and the Romanian wordnets differ. Therefore the *semantic-similarity* score would have different values, depending whether it was computed between the Czech synsets that correspond to R1 and R2 or between the Romanian synsets that correspond to the same ILI codes.

The PWN-based *semantic-similarity* mediates among different (but similar) wordnet topologies. The similarity of the wordnets topologies is a direct consequence of the *hierarchy preservation principle* mentioned earlier.

Having a parallel corpus containing texts in $k + 1$ languages (T, L_1, L_2, ..., L_k) and having monolingual wordnets for all of them, interlinked via an ILI-like concept repository, let us call the T language as the target language and the languages L_1, L_2, ..., 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/`)[2]:

```
 <tu id="Ozz.113"> <seg lang="en">
   <s id="Oen.1.1.24.2">
   <w lemma="Winston"ana="Np">Winston</w>
   <w lemma="be" ana="Vais3s">was</w>
        ... </s>
</seg>
<seg lang="ro">
   <s id="Oro.1.2.23.2">
   <w lemma="Winston" ana="Np">Winston</w>
   <w lemma="fi" ana="Vmii3s">era</w>
        ... </s>
</seg>
<seg lang="cs">
   <s id="Ocs.1.1.24.2">
   <w lemma="Winston" ana="Np">Winston</w>
   <w lemma="se" ana="Px---d--ypn--n">si</w>
        ... </s>
</seg>
...
</tu>
```

Fig. 1. A partial translation unit from the parallel corpus.

[2]At `http://nl.ijs.si/ME/mteV3-2004-03-19/` one can find a newer version of these encoding specifications.

For each source language and for all occurrences of a specific word in the target language T (called the target word, the occurrences of which are to be disambiguated throughout the corpus) 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 target word).

Table 1. The translation equivalents matrix (EQ matrix)

	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}

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 at the beginning of this section.

The step **B)** of the basic methodology transforms the matrix shown in Table 1 to a matrix with the same dimensions (Table 2) called VSA (Validation and Sense Assignment):

Table 2. The VSA matrix

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

with $VSA_{ij} = L^{EN}_{ILI}(W_{EN}) \cap L^{i}_{ILI}(W^{j}_{Li})$, where $L^{EN}_{ILI}(W_{EN})$ is the set of the ILI-codes of all PWN synsets in which the target *literal*(W_{EN}) occurs, and $L^{i}_{ILI}(W^{j}_{Li})$ is the set of the 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 *VSA*s, the algorithm cannot determine the sense number of 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 in any of the source languages, and as such the lack of information from one source language could be compensated by looking at the other source languages.

a) When the cell VSA_{ij} contains a single ILI code, then this is the common interlingual concept realized in the two considered languages by the j-th occurrence of the target word and its translation equivalent. Knowing the concept, by following the interlingual relations in the two wordnets, one uniquely identifies the synsets and thus the word senses for both words. For instance, let us consider the English-Romanian translation equivalence pair ⟨*toe deget*⟩ for

which a corresponding *VSA* contains only the ILI-code *ENG20-0528265-n.* This code uniquely identifies the English synset (toe:1) in PWN and to the synset (deget:1.1.2) in the Romanian wordnet both with the same gloss meaning: (one of the digits of the foot). Thus the disambiguation of this translation pair is ⟨toe(1) deget(1.1.2)⟩.

b) When the cell VSA_{ij} contains two or more ILI codes, this exemplifies what we call *cross-lingual ambiguity*, i.e. the two words in the translation equivalence relation can be used to linguistically realize a common set of two or more inter-lingual concepts. For instance, at least two senses of the English word *movement* corresponding to the concepts of motion (a 2^{nd}OrderEntity-Dynamic) and social group (1^{st}OrderEntity-Composition-Group) are identical to the senses carried by the Romanian word *mişcare*. Therefore at least 2 ILI codes will be in a VSA cell corresponding to the translation equivalence pair ⟨*movement mişcare*⟩. In such a case, out of all candidates, the concept corresponding to the most frequent sense of the target word (as seen in the English part of current bitext) is selected. If this heuristics cannot make the difference, the choice is made in favour of the concept corresponding to the PWN 2.0 synset containing the target word with the smallest sense number.

c) When the cell VSA_{ij} 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 $L^{EN}_{ILI}(W_{EN}) \otimes L^{i}_{ILI}(W^{j}_{Li})$ which shows the highest similarity. In case of ties, the heuristics discussed before are applied.

d) 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, as mentioned before, it is extremely rare that there is no translation equivalent for an occurrence of the target word in any of the source languages.

5. An Improvement to the WSD Algorithm

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. In [17] we described a clustering algorithm [11] based on translation equivalents and there we used the same parallel corpus as in our current experiment.

The back-off method consists of applying the clustering method after the wordnet-based method has been applied. Thus each cluster containing non-disambiguated occurrences 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 the reliance only on high quality, $k+1$ source wordnets. Indeed, having $k+1$ languages in our corpus, we need only to apply the WSD method, as described before, for the target and one source language and use the 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 falls into the cases b) and c) discussed at the end of the previous section, and will be able to assign a sense for all translation pairs falling into the 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;

Before going into the details of the sense clustering, let us introduce a few notations:

1. $\mathrm{TWL} = \{\mathrm{TW}^i\}_{1 \le i \le n}$, the Target Word List;
2. TW^i_k, the k-th occurrence of TW^i;
3. $\mathrm{DEL}(\mathrm{L}_p,\mathrm{TW}^i) = \{\mathrm{W}^j \mid \langle \mathrm{TW}^i, \mathrm{W}^j \rangle$ 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 [18];
4. $|\mathrm{DEL}(\mathrm{L}_p,\mathrm{TW}^i)|$ = the number of elements in $\mathrm{DEL}(\mathrm{L}_p,\mathrm{TW}^i)$;
5. $\mathrm{TEQ}(\mathrm{L}_p,\mathrm{TW}^i_k)$ = the Translation EQuivalent in language L_p for the k^{th} occurrence of TW^i, $\mathrm{TEQ}(\mathrm{L}_p,\mathrm{TW}^i_k) \in \mathrm{DEL}(\mathrm{L}_p,\mathrm{TW}^i)$;
6. $\mathrm{DEL}_h(\mathrm{L}_p,\mathrm{TW}^i)$ = the h-th element of $\mathrm{DEL}(\mathrm{L}_p,\mathrm{TW}^i)$;
7. $\mathrm{LVECT}(\mathrm{L}_p,\mathrm{TW}^i_k)$ = a binary vector of $|\mathrm{DEL}(\mathrm{L}_p,\mathrm{TW}^i)|$ positions; all the binary positions are 0 except for at most one bit at position h wich is 1 if $\mathrm{TEQ}(\mathrm{L}_p,\mathrm{TW}^i_k) = \mathrm{DEL}_h(\mathrm{L}_p,\mathrm{TW}^i)$. 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. $\mathrm{VECT}(\mathrm{TW}^i) = \mathrm{CON}_{p=1,S}(\mathrm{LVECT}(\mathrm{L}_p,\mathrm{TW}^i_k))$, with CON a vector concatenation operator and S the number of source languages in the parallel corpus.

The sense clustering algorithm is the following:

- **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):

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

  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:

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

  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 $VECT(TW_k^i)$ as leaves; 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 (a perfect one even less) 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 a distance heuristics as an exit condition for the clustering algorithm (thus a way of computing Z). When the minimal distance between the existing classes increases *too much*, then the algorithm should stop. This fuzzy statement is turned into the exit condition as shown in (5):

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

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 (as a back-off mechanism) can be extended so that 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 wordnet-based WSD since if two occurrences were wrongly labeled as different they would not have a chance to be clustered together. In our approach, the final sense labeling, based on a majority voting, gives credit to the clustering algorithm so, if this is wrong, some initial good sense labeling could be overridden and turned into wrong sense-labeling.

6. Experimental Results

In order to evaluate both the performance of the WSD algorithm 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 target words were extracted from the parallel corpus *1984* so that all their senses (at least two per POS) defined in PWN 2.0 were also implemented (and interlingually aligned) in all BalkaNet wordnets. There resulted 211 words with 1 810 occurrences in the English part of the parallel corpus. We manually assigned senses to all these 1 810 occurrences of the target words, building the Gold Standard annotation (GS). A number of 13 students, enrolled in the Computational Linguistics Master program at the University "Al. I. Cuza" from Iaşi, were asked to manually sense-tag the occurrences of the target words occurring in a set of assigned sentences. 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 targeted words, a simple majority sense was computed (MAJ). Finally, the same targeted words were automatically disambiguated by the WSD algorithm (ALG). From the entire set of target words, the system could not sense-disambiguate 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. The error rate of our last version of the word aligner (for non-null alignments) is about 11.5% and is largely due to English words occurring only once, or English words that are translated each time differently so that the corresponding

translation pairs are hapax legomena[3]. 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; we spotted 34 such errors, very difficult to find by eye inspection of the wordnet). 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 clustering mechanism discussed in the previous section). 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.

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

GS=MAJ	GS=ALG	MAJ=ALG	GS=MAJ=ALG
72.99%	74.88%	62.99%	60.4%

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 that the agreement of one of the students was significantly better than GS = ALG score: 78.71%).

We found this result extremely encouraging as it shows that the tedious handmade WSD in building word-sense disambiguated corpora (presumably done by an expensive expert) can be avoided.

7. WSDTool

The WSD algorithm described in section 4 was incorporated into a Java application called WSDTool which also ensures editing facilities to spot and correct either incorrect/incomplete synsets or their interlingual alignment. The validation mode of WSDTool is relevant for the semantic validation of the BalkaNet interlingually linked wordnets.

WSDTool offers an easy-to-use interface with a graphic visualization of the various wordnets. In this section we will exemplify its use in the validation regime with English as target language and Romanian as source language. We will focus the presentation on cases where the automatic WSD cannot be completed or performed at all for

[3]However, most of the time hapax legomena pairs which are cognates are correctly found.

the reasons discussed previously (similar cross-lingual ambiguity, incomplete synsets, wrong interlingual alignments). The interface (see Figure 2) allows loading a list of target words (the *T-Words* headed window) for which the associated VSA matrixes will be computed and displayed (the *VSA Matrix* headed window).

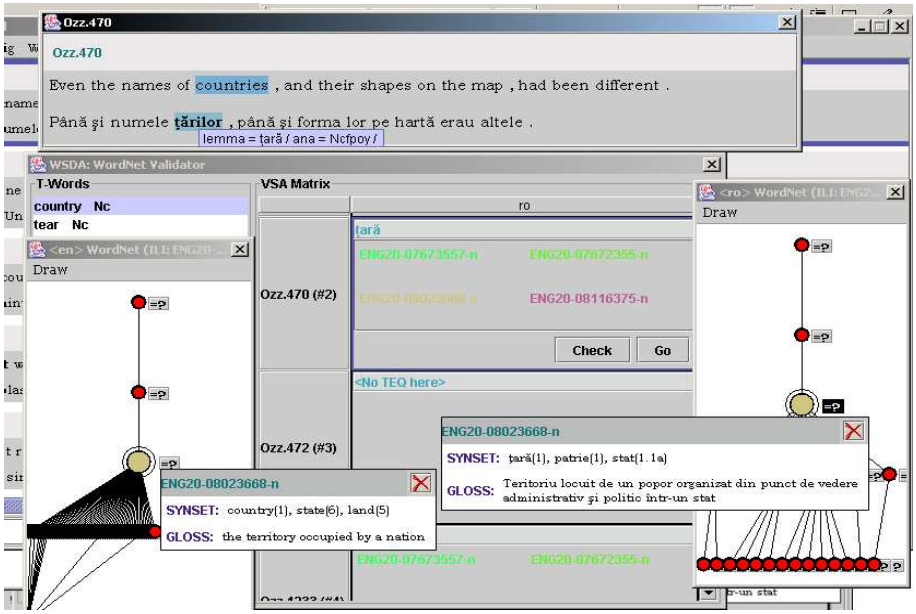

Fig. 2. WSDTool, case 1.

With one source language, the VSA matrix is a vector the lines of which are labeled with an occurrence identifier of the form Ozz.N(#M) where Ozz.N represents the ID of the translation unit in which the target word appears and M represents the occurrence number of the target word. For instance the identifier Ozz.470(#2) indexing one of the lines of the VSA vector associated to the target word **country** refers to the second occurrence of this word which appeared in the translation unit Ozz.470 (see Figure 2). The cell of the VSA vector, labeled with the Ozz.N(#M) occurrence identifier, displays the translation equivalent for the M^{th} occurrence of the target word found in the translation unit Ozz.N as well as the result of the intersection between the set of ILI-codes that correspond to all senses (included in the target wordnet) of the target word and the set of ILI codes that correspond to all senses (included in the source wordnet) of the corresponding translation equivalent. This VSA vector cell also contains two buttons: "Go" and "Check". Clicking on "Go", would point the corpus view manager to the translation unit Ozz.N and double-clicking this translation unit, would cause it to be displayed in a larger "editable" window in which every word can be inspected for its attributes (POS and lemma). "Check" displays a validation window like the one in Figure 5.

Depending on the intersection result displayed in a VSA cell there are three main cases of interest, illustrated and commented below:

1. the cell contains a set of ILI codes; this means that the target word and its respective translation equivalent belong, with different senses, to different synsets which are interlingually aligned to the corresponding concepts. This could be a case of similar cross-lingual ambiguity or an error in the source synsets corresponding to the current translation equivalent: either the word should not belong to all synsets, or one or more synsets are wrongly interlingually linked. The user is offered editing facilities to correct the spotted error (if any).

2. the cell contains pairs of ILI codes; each pair is tailed by a real number denoting the similarity measure between the members of the pair; the similarity measure was calculated as described above (see section 4, equation (1)). As we presented earlier a VSA cell contains pairs of ILI codes only when the words in the current translation pair have no senses belonging to synsets that are linked to the same concepts. Frequently this happens because of the human translation (the translator used a more generic or more specific word than the one which would have been the proper translation equivalent). This case could also appear due to an objective lexical gap or due to an alignment error. The user is now required to choose the pair which corresponds best to the contextual senses of the words in the translation pair – see Figure 3). If such a pair does not exist, the necessary corrective editing should be performed;

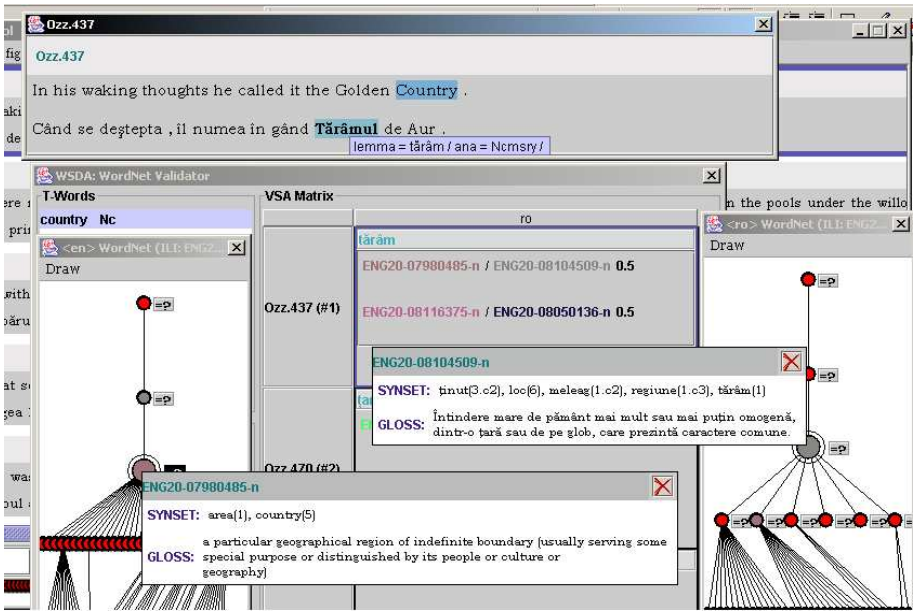

Fig. 3. WSDTool, case 2.

3. the cell is empty; this is a potential interlingual linking error or an incomplete synset (see figure 4). If (w_{EN}, w^i_{RO}) is a correct translation pair, then one of the following must hold: the relevant w^i_{RO} synset is wrongly mapped, or the

sense of the i^{th} occurrence of w_{EN} is not yet implemented for the corresponding translation equivalent literal w^i_{RO} (see Figure 5) or the literal w^i_{RO} does not belong to the relevant RO synset. If this is the case, the user is asked to add it to the proper synset (this way, synset expanding can be achieved in a focused way).

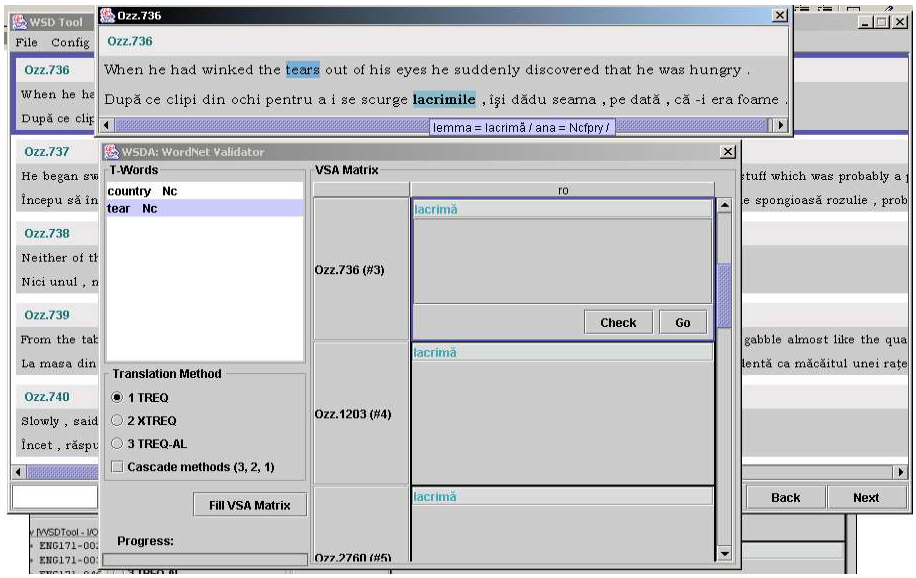

Fig. 4. WSDTool, case 3a.

In this example, the third occurrence of the target word **'tear'**, occurring in the translation unit Ozz.736, was correctly translated by **'lacrimă'**; but the corresponding cell is empty. The reason is that **'lacrimă(1)'** was wrongly aligned on another PWN synset (namely `lachrymal_secretion(1)`) and the corresponding correct PWN synset `tear(1)` was doomed not to be lexicalized in Romanian. This "mistake" is due to our Romanian Explanatory Dictionary which provides a general definition for the part-whole pair. The snapshot in Figure 5 shows the check window where the user realizes that the relevant sense of **'lacrimă'** is wrongly aligned in the Romanian wordnet.

8. Conclusions

Our disambiguation results, at the PWN 2.0 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 (119 528 out of 145 627 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.

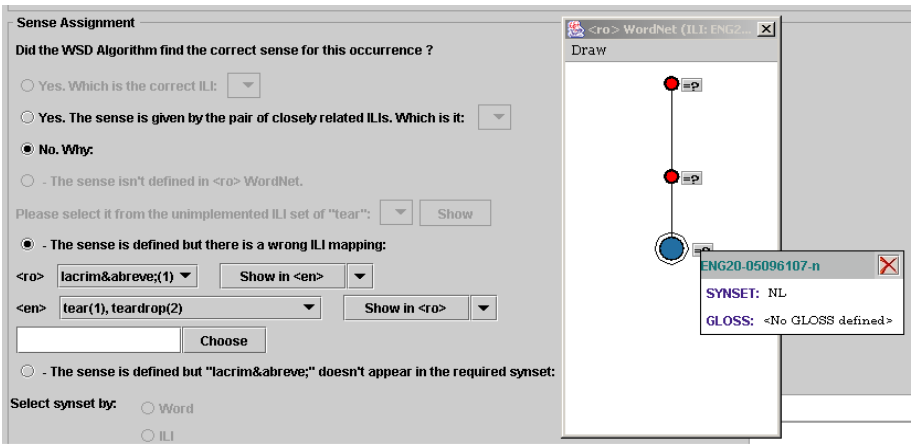

Fig. 5. WSDTool, case 3b.

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

References

[1] CRISTEA, D., MIHĂILĂ, C., FORĂSCU, C., TRANDABAT, D., HUSARCIUC, M., HAJA, G., POSTOLACHE, O., *Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets*, in this volume, 2004.

[2] FELLBAUM, C. (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[3] GALE, W., WARD CHURCH, K., YAROWSKY, D., *Estimating upper and lower bounds on the performance of wordsense disambiguation programs*, in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 249–256, 1992.

[4] HORÁK, A., SMRŽ, P., *New features of wordnet editor VisDic*, in this volume, 2004.

[5] KILGARRIFF, A., *I don't believe in word senses*, Computers and the Humanities, **31** (2), 91–113, 1997.

[6] IDE, N., VÉRONIS, J., *Word Sense Disambiguation: The State of the Art*, Computational Linguistics, **24**:1, 1–40, 1998.

[7] IDE, N., *Parallel translations as sense discriminators*, in *SIGLEX99: Standardizing Lexical Resources*, ACL99 Workshop, College Park, Maryland, 52–61, 1999.

[8] IDE, N., ERJAVEC, T., TUFIŞ, D., *Automatic Sense Tagging Using Parallel Corpora*, in *Proceedings of the 6^{th} Natural Language Processing Pacific Rim Symposium*, 212–219, Tokyo, 2001.

[9] 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*, ACL2002, July 2002, Philadelphia, 56–60.

[10] RODRIGUEZ, H., CLIMENT, S., VOSSEN, P., BLOKSMA, L., PETERS, W., ALONGE, A., BERTAGNA, F., ROVENTINI, A., *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*, Computers and the Humanities, **32**(2–3), 117–152, 1998.

[11] STOLCKE, A., *CLUSTER 2.9.*, `http://www.icsi.berkeley.edu/ftp/global/pub/ai/stolcke/software/cluster-2.9.tar.Z`, 1996.

[12] TUFIŞ, D., CRISTEA, D., STAMOU, S., *BalkaNet: Aims, Methods, Results and Perspectives: A General Overview*, in this volume, 2004.

[13] TUFIŞ, D., BARBU, E., BARBU-MITITELU, V., ION, R., BOZIANU, L., *The Romanian Wordnet*, in this volume, 2004.

[14] TUFIŞ, D., ION, R., BARBU, E., BARBU-MITITELU, V., *Cross-Lingual Validation of Wordnets*, in *Proceedings of the 2^{nd} International Wordnet Conference*, 332–340, Brno, 2004.

[15] 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, Canada, 36–39, 2003, (`www.cs.unt.edu/~rada/wpt/index.html#proceedings/`).

[16] TUFIŞ, D., CRISTEA, D., *Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet*, in *Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation*, Las Palmas, Spain, May, 35–41, 2002.

[17] TUFIŞ, D., ION, R., *Word Sense Clustering Based on Translation Equivalence in Parallel Texts; A Case Study in Romanian*, in Corneliu Burileanu (ed.), *Speech and Dialogue Systems*, Romanian Academy Publishing House, Bucharest, 2003, 8–22.

[18] TUFIŞ, D., BARBU, A.M., *Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing*, International Journal on Speech Technology, **5**, no. 3, Kluwer Pbls, 2002.

[19] VOSSEN, P., PETERS, W., GONZALO, J., *Towards a Universal Index of Meaning*, in *Proceedings of ACL/SIGLEX'99.*