Interlingual wordnets validation and word-sense disambiguation

Dan Tufiş[1,2], Radu Ion[1]

[1]Institute for Artificial Intelligence, 13, Calea 13 Septembrie, 050711, Bucharest 5, Romania
[2]Faculty of Informatics, University "A.I. Cuza", 16, Gral. Berthelot, Iaşi, 6600, Romania

Abstract. Understanding natural language assumes one way or another, being able to select the appropriate meaning for each word in a text. 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 a few of the senses one word would have listed in any general purpose dictionary. Moreover, when multiple senses are considered for a lexical item, the granularity of these senses is very coarse so that discriminating them is much simpler than in the general case. 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 WordNet 2.0 (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 via an interlingual index (ILI), for word-sense disambiguation in parallel texts. The words of interest, irrespective of the language in the multilingual documents are uniformly disambiguated by using the same sense-inventory labels.

1 Introduction

Semantic lexicons are one of the most valuable resources for a plethora of natural language applications. Incorporating Wordnet or its monolingual followers in modern NLP-based systems becomes a general trend motivated by numerous reports showing significant improvements in the overall performances of these systems. Multilingual wordnets, such as EuroWordNet and the ongoing BalkaNet, which adopted the Princeton Wordnet [1] as a conceptual interlingua, represent one step further with great promises in the domain of multilingual processing. We describe a multilingual environment for word-sense disambiguation in parallel texts, relying on several monolingual wordnets developed within the BalkaNet European project. The BalkaNet wordnets are aligned via PWN2.0 which is used as an interlingual index (ILI). A general presentation of the BalkaNet project is given in [2]. The detailed presentation of the Romanian wordnet, part of the BalkaNet multilingual lexical ontology, is given in [3, 4]. The EuroWordNet is largely described in [5].

Tufiş D. and Ion R. (2004).
Interlingual wordnets validation and word-sense disambiguation.
In *Proceedings of the 1st International Workshop on Natural Language Understanding and Cognitive Science*, pages 97-105
DOI: 10.5220/0002677900970105

Terminologically, in a multilingual wordnet of the type considered here, one may distinguish meanings from concepts:

- the meanings are lexicalized in different languages by synsets (synonymy lists, each lemma being indexed by the sense number that justifies the synonymy relation); the synsets/meanings of the monolingual wordnets are structured similarly, using both standard relations from the set defined in PWN2.0 and language specific relations (especially to deal with idiosyncratic lexical relations among words in each languages); language specific synsets are linked, by equivalence relations, to the concepts in the interlingual index;
- the concepts are "language independent" representations of the similar meanings expressed in different languages; they are anchor points that allows one to navigate from a synset in one language to the synsets that lexicalize the same (or a very close) meaning in all the other languages; currently the concepts in the BalkaNet multilingual wordnet are in a one-to-one mapping to the PWN2.0 but several region specific concepts (more often then not they are lexicalized across Balkan languages by cognates), will be added in the interlingual index.

The BalkaNet project aims at building, along the lines of the EuroWordNet lexical ontology, wordnets for five new Balkan languages (Bulgarian, Greek, Serbian, Romanian and Turkish) and at improving the Czech wordnet developed in the EuroWordNet project. The BalkaNet consortium adopted a concerted strategy for building the monolingual wordnets so they would maximize the cross-lingual coverage. To this end, a set of common ILI concepts corresponding to a conceptually dense subset of PWN2.0 was selected and implemented in each language.

The methodology and the system which implements the multilingual environment for word-sense disambiguation presuppose the correctness of the monolingual wordnets and their *accurate interlingual linking*. If this is not the case, the same system can be interactively used for identifying missing senses for the targeted words, for pinpointing conceptual alignment errors between the senses of words in different languages and for correcting whatever errors were found. Both the autonomous and the interactive regimes of the system result in uniformly sense-tagging of the words of interest, irrespective of the language in the multilingual documents. The uniform sense labels are ILI codes. For instance the ILI-code 04209815-n identifies the interlingual concept expressed in English by any synonym of the word *table* (sense 2 in PWN2.0) and which is lexicalized in the BalkaNet language wordnets by the words (and their respective synonyms) *маса* (Bulgarian sense 1), τραπέζι (Greek sense 1), *sto* (Serbian sense 1), *masă* (Romanian sense 13) *masa* (Turkish sense 1) or *stůl* (Czech sense 1).

2 Assumptions and the Basic Methodology

One fundamental assumption in the study of language is its compositional semantics. Compositionality is a feature of language by virtue of which the meaning of a sentence is a function of the meanings of its constituent parts (going down to the level of the constituent words). With this tarskian approach to meaning, our methodology assumes that the meaning building blocks (lexical items – single or multiple word

units) in each language of a parallel text could be automatically paired (at least some of them) and as such, these lexical items should be aligned to closely related concepts at the ILI level. In other words, if the lexical item W^i_{L1} in the first language is found to be translated in the second language by W^j_{L2}, common intuition says that 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. However, in both EuroWordNet and BalkaNet the interlingual index is not structured, so we need to clarify what "closely related ILI recods" means. We define the *relatedness* of two ILI records R_1 and R_2 as the *semantic similarity* between the synsets Syn_1 and Syn_2 of PWN2.0 that correspond to R_1 and R_2. A semantic similarity function $SYM(Syn_1, Syn_2)$ could be defined in many ways [6]. We used a very simple and effective one: $SYM(Syn_1, Syn_2)=(1+N)^{-1}$ where N is the number of oriented links traversed from one synset to the other or from the two synsets up to the closest common ancestor. One should note that every synset is linked (EQ-SYN) to exactly one ILI record and that no two different synsets of a given wordnet have the same ILI code assigned to them. In the context of this research, we assume that the *hierarchy preservation* principle [4] holds true.

As a test-bed, we use the wordnets developed within the BalkaNet European project and the "*Nineteen Eighty-Four*" parallel corpus [7] which currently includes four relevant languages for BalkaNet (with the prospects of extending the corpus to all the BalkaNet languages). The methodology for semantic validation assumes the following basic steps:

A) given a bitext T_{L1L2} in languages L1 and L2 for which there are aligned wordnets, one extracts the pairs of lexical items that are reciprocal translations: $\{<W^i_{L1}\ W^j_{L2}>^+\}$
B) for each lexical alignment of interest, $<W^i_{L1}\ W^j_{L2}>$, one extracts the synsets in each language that contain the lexical items of the current pair and respectively their ILI projections. There will result two lists of ILI labels, one for each language, L^1_{ILI} and L^2_{ILI}. Based on the content evaluation of these two lists, several lines of reasoning might be followed highlighting various problems related to: the implementation of one or the other of the two wordnets, the alignment to the ILI; different sense granularity among wordnets; lexical gaps; wrong translation in the bitext, etc.

The first processing step is crucial and its accuracy is essential for the success of the validation method. A recent shared task evaluation (http://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. The best performing word alignment system [8] produced bilingual translation lexicons, relevant for wordnets evaluation, with an aggregated F-measure as high as 84.26%.

3 Interlingual Validation Based on Parallel Corpus Evidence

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 structure, let us call T the target language and 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/):

Table 1. A partial translation unit from the parallel corpus

```
<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>
```

We will refer to the wordnet for the target language as T-wordnet and to the one for the language L_i as the i-wordnet. We use the following notations:
T_word = a target word, say w_{TL};
T_word_j = the j-th occurrence of T_word;
eq_{ij} = the translation equivalent (TE) for T_word_i in the source language L_j, say w_{SLj}; the pair (w_{TL}, w_{SL}) is called a translation pair (for the languages considered);
EQ = the n*k matrix containing translation equivalents of the T_word (n occurrences, k languages):

Table 2. The translation equivalents matrix (EQ matrix)

	L_1	L_2	…	L_k
Occ #1	eq_{11}	eq_{12}	…	eq_{1k}
Occ #2	eq_{21}	eq_{22}	…	eq_{2k}
…	…	…	…	…
Occ #n	eq_{n1}	eq_{n2}	…	eq_{nk}

TU_j = the translation unit containing T_word_j;
EQ_i = a vector, containing the TEs of T_word in language L_i: (eq_{1i} eq_{2i} …eq_{ni})

If T_word_j is not translated in the language L_j then eq_{ij} is represented by the null string. Every non-null element eq_{ij} of the EQ matrix is subsequently replaced with the set of all ILI codes that correspond to the senses of the word eq_{ij} as described in the wordnet of the j-language. Thus we obtain the matrix EQ_ILI which is the same as EQ matrix except that it has a set of ILI codes for every cell. If some cells in the EQ matrix contain empty strings, then the corresponding cells in EQ_ILI will obviously contain empty sets. For T_word the set of ILI codes is T_ILI = (ILI_{T1} ILI_{T2} … ILI_{Tq}).

The next step is to define our target data structure. Let us consider a new matrix, called VSA (Validation and Sense Assignment):

Table 3. The VSA matrix

	L_1	L_2	…	L_k
Occ #1	VSA_{11}	VSA_{12}	…	VSA_{1k}
Occ #2	VSA_{21}	VSA_{22}		VSA_{2k}
…	…	…	…	…
Occ #n	VSA_{n1}	VSA_{n2}	…	VSA_{nk}

with VSA(i,j) = T_ILI ∩ EQ_ILI(i,j), if EQ_ILI(i,j) is non-empty and ⊥ (undefined) otherwise.

The j^{th} column of the VSA matrix provides valuable corpus-based information for the evaluation of the interlingual linking of the the j-wordnet and T-wordnet.

Ideally, computing for each line i the set SA_i (sense assignment) as the intersection VSA(i,1)∩VSA(i,2)…∩VSA(i,k) one should get at a single ILI code: $SA_j=(ILI_{T\alpha})$, that is the i^{th} occurrence of the target word was used in all source languages with the same meaning, represented interlingually by $ILI_{T\alpha}$. If this happened for any T_word, then the WSD problem (at least with the parallel corpora) would not exist. But this does not happen, and there are various reasons for it: the wordnets are partial and (even the PWN) are not perfect, the human translators make mistakes, there are lexical gaps between different languages, the automatic extraction of translation equivalents is far from error-free, etc.

Yet, for cross-lingual validation of interlinked wordnets the analysis of VSAs may offer wordnet developers extremely useful hints on senses and/or synsets missing in their wordnets, wrong ILI mappings of synsets, wrong human translation in the parallel corpus and mistakes in translation equivalents extraction. Once the wordnets have been validated and corrected accordingly, the WSD (in parallel corpora) should be very simple. There are two ways of exploiting VSAs for validation:

Vertical validation (VV): the development team of i-wordnet (native speakers of the language L_i with very good command of the target language) will validate their own i-wordnet with respect to the T-wordnet, that is from all VSA matrixes (one for each target word) they would pay attention only to the i^{th} column (the VSA(L_i) vector).

Horizontal validation (HV): for each VSA all SAs will be computed. Empty SAs could be an indication of ILI mapping errors still surviving in one or more wordnets (or could be explained by lexical gaps, wrong translations etc) and as such, the suspicious wordnet(s) might be re-validated in a focused way. The case of an SA containing more than a single ILI identifier could be explained by the possibility of having in all i-languages words with similar ambiguity.

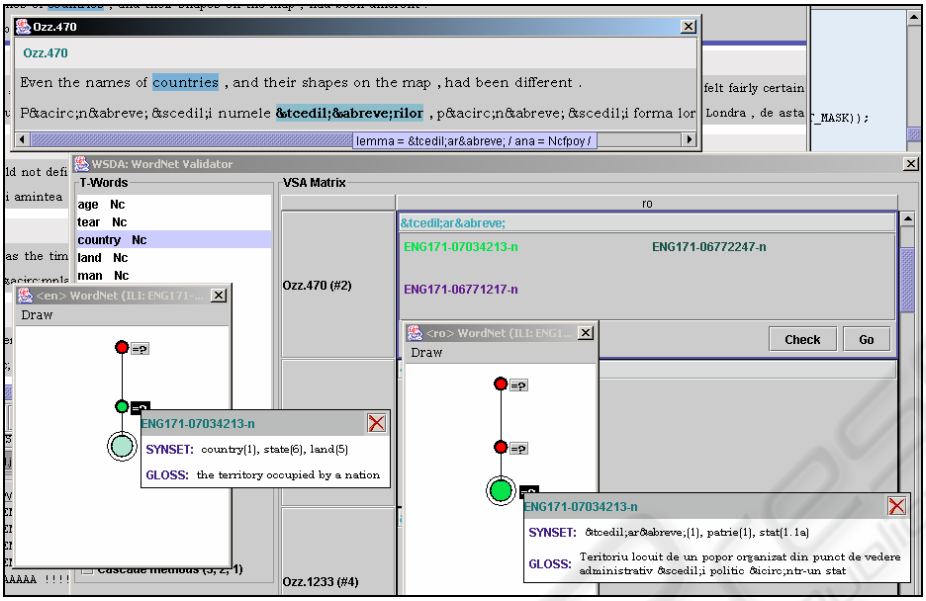

Figure 1. WSD Tool validation interface (English-Romanian pair of languages)

Our system called WSD Tool implements the methodology described above and offers an easy-to-use interface for the interactive task of semantic validation. It incorporates the statistical translation equivalents extraction system (TREQ&TREQ-AL, described in [9, 10]) as well as a graphic visualization of the two wordnets used in the validation process. When it runs in the validation (interactive) regime, the WSD Tool interface displays (see Figure 1):

- a browsable list of target words (T-Words);
- when a target word is clicked, all the translation units in which it appears are shown in a new scrollable window with the occurrences of T-word and their corresponding translation equivalents eq_{ij} being highlighted;
- clicking the eq_{ij} translation equivalent in the i^{th} translation unit simultaneously opens three smaller windows: the first one displays the content of the VSA(i,j) cell while the second and third ones contain browsable graphical representations of the hierarchies in the T-wordnet and j-wordnet of the synsets that are linked to the ILI codes in the VSA(i,j) set. Clicking the nodes of these graphs will display the appropriate entries from the respective wordnet.

4 Evaluation of the Automatic Word Sense Disambiguation

The evaluation of the WSD Tool accuracy in word-sense disambiguation requires proper command of languages present in the parallel corpus. Therefore we restrict ourselves on the English-Romanian bitext and use data from the vertical validation with EN as target language and RO as source language. We should mention that vertical validations for other source languages in BalkaNet multilingual wordnet are

planned by each consortium team, so that a horizontal validation and a harmonized sense-tagging of all the languages in the "1984" parallel corpus is expected soon.

For the purpose of this evaluation we selected a bag of English nouns and verbs occurring in the original text of Orwell with the following restrictions: a) all their senses listed in PWN2.0 corresponded to ILI records that were implemented in the Romanian wordnet (this way we were sure that irrespective of the sense in which such a word was used in the English text, its meaning was present in the Romanian wordnet) and b) each selected English word had at least two senses. Without the second restriction, the initial bag of English words contained 530 lemmas but, after removing the trivial cases (restriction b) the target bag of words contained only 211 lemmas (122 nouns and 89 verbs) with the number of senses ranging from two to eight.

Table 5. The WSD evaluation in the validation regime

target words	occ.	occ. not translated	translated occ.	occ. fully disambiguated	occ. partially disambiguated	wrong TEs	Wordnet errors
211	1756	355	1401	681 (48.60%)	514 (36.68%)	174 (12.41%)	36 (2.56%)

These 211 word types, altogether, occurred 1756 times. Out of the total number of the target English occurrences, 355 were not translated in the Romanian part of the bitext so that they were discarded from this evaluation. The precision of the disambiguation procedure for the 1401 translated occurrences of the targeted words is summarized in the table 5.

Almost half of the occurrences of the target English words (681) were fully disambiguated (the corresponding cell in the VSA matrix contained a single ILI identifier). For 514 occurrences (36.86%) of the target words the disambiguation was partial meaning that the corresponding cells of the VSA matrix contained at least two pairs of ILI identifiers, each of them being associated with a similarity score. In the validation mode, we selected the correct disambiguation. In 398 cases the correct pair was the one with the highest similarity score. In 108 cases there were two pairs with the best score and the correct disambiguation was among them. The heuristics according to which our system resolves the draws (picking the pair with the smallest sum of sense numbers) gave the correct result 104 times. In 3 cases the correct sense was present but not the best scored and in 5 cases the correct sense was not in the list (these cases revealed wrong Wordnet alignments).

The last two rubrics in Table 5 show errors in the data used by WSD Tool.

Wrong TEs are mistakes done by either TREQ-AL, our word aligner and translation equivalence extractor, or the preprocessing phases (tokenization, tagging). The TREQ-AL error rate (12.41%) in this experiment is consistent with the error rate (for the dictionary extraction) previously reported (13.32%) in the word-alignment competition at the *NAACL 2003 Workshop on Building and Using Parallel Texts* (Romanian-English Shared Task) [8]. The slightly better figure in this experiment is due to the fact that here we considered only nouns and verbs, while in [8] the evaluation was for all parts of speech. In our current approach, the WSD error rate is bound to the TREQ-AL error rate so, unless the word aligner is further improved, it cannot go beyond to 12-13%.

The wordnet errors rubric contains the number of errors directly ascribable to the Romanian wordnet construction and its linking to the ILI. We identified several cases of incomplete Romanian synsets (28) and a few cases of interlingual linking mistakes (8). The table 6 summarizes the discussion above, with WSD Tool ran in the automatic regime.

Table 6. The WSD evaluation in the automatic regime

target words	occ.	occ. Not translated	translated occ.	occ. correctly disambiguated	occ. wrongly disambiguated
211	1756	355	1401	1183 (84.44%)	218 (15.56%)

For computing the recall of the disambiguation procedure we considered all the target word occurrences (translated and not translated in Romanian). The word sense disambiguation recall in English is 67.36%. However, in a multi-languages parallel corpus, the recall of WSD for the target language could be significantly improved considering other source languages. It is very likely that occurrences of the target words not translated in one language could be translated in other languages, and thus, by the same procedure, they get a sense-tag from another pair of languages. Also, one could try an agglomerative sense clustering [10] for the target words the occurrences of which were not all sense-tagged. Most of the untagged occurrences will be clustered together with tagged occurrences and again, one could get a strong clue on the appropriate semantic tags for the untagged words.

5 Conclusions

This preliminary experiment shows that using translation equivalents extracted from a test-bed parallel corpus may precisely pinpoint various problems in the wordnets structuring and interlingual linking. Since our wordnet is essentially based on human expertise and on language resources of very good quality (printed explanatory and synonyms dictionaries, turned into machine readable dictionaries) the percentage of errors due to the synsets linking or due to incomplete data in the reference language resources (missing senses for a literal or literals missing from a given synset) are reasonably low. However, the detected wordnet errors (very hard to detect by simply inspecting the synsets of the wordnet under construction) showed that this approach is not only an effective way to check out ongoing work, but also one way to continuously update a monolingual dictionary in accordance with the actual use of languages in multilingual environments.

The WSD Tool system is implemented in Java and is language independent. Vertical validations for all languages in the BalkaNet are planned for the immediate future which will enable us to perform a horizontal validation with at least four source languages. The evaluation of the word-sense disambiguation exercise shows a very high accuracy. The word-sense disambiguation based on PWN2.0 sense inventory appears to be much more accurate in a parallel corpus than in a monolingual one (see for instance the results reported in SensEval conferences). Actually this is not surprising, because a parallel corpus embeds translators' expertise which, once

revealed (by the translation equivalents extraction program) is an extremely powerful source of knowledge for semantic disambiguation.

References

1. Fellbaum, Ch. (Ed.) (1998) WordNet: An Electronic Lexical Database, MIT Press
2. Stamou, S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufiş, D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002): BalkaNet A Multilingual Semantic Network for the Balkan Languages, in Proceedings of the 1st *International Wordnet Conference*, Mysore
3. Tufiş, D., Cristea, D. (2002): Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet, In Proceedings of *LREC2002 Workshop on Wordnet Structures and Standardisation*, Las Palmas, Spain, May, 35-41
4. Tufiş, D., Cristea, D.: Probleme metodologice în crearea Wordnet-ului românesc şi teste de consistenţă pentru BalkaNet, în Tufiş, D., F. Gh. Filip (eds.) *Limba Română în Societatea Informaţională - Societatea Cunoaşterii*, Editura Expert, Academia Română, (2002) 139-166.
5. Vossen, P. (Ed.) (1999): EuroWordNet: a multilingual database with lexical semantic networks for European Languages, Kluwer Academic Publishers, Dordrecht
6. Budanitsky, A., Hirst, G. (2001): Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the *Workshop on WordNet and Other Lexical Resources*, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, June.
7. Erjavec T., Ide, N. (1998) "The Multext-East corpus". In *Proceedings LREC'1998*, Granada, Spain, pp. 971-974.
8. Tufiş D., Barbu A.M., Ion R. (2003): A word-alignment system with limited language resources, Proceedings of the *NAACL 2003 Workshop on Building and Using Parallel Texts*; Romanian-English Shared Task, Edmonton, Canada, 36-39 (also at: http://www.cs.unt.edu/~rada/wpt/index.html#proceedings/).
9. Tufiş, D. Barbu, A.M. (2002): „Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing", in *International Journal of Speech Technology*. Kluwer Academic Publs, no.5, pp. 199-209.
10. Dan Tufiş, Ana Maria Barbu, Radu Ion: "Extracting Multilingual Lexicons from Parallel Corpora", 38 pages (to appear in *Computers and the Humanities*, 2004)
11. Nancy Ide, Tomaz Erjavec, Dan Tufiş: „Sense Discrimination with Parallel Corpora" in Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, July Philadelphia 2002, pp. 56-60