Word Senses: The Stepping Stones in Semantic-Based Natural Language Processing

Dan Tufiş
Institute for Artificial Intelligence, 13, Calea 13 Septembrie, 050711,
Bucharest 5, Romania
Faculty of Informatics, University "A.I. Cuza", 16, Gral. Berthelot, Iaşi,
6600, Romania
tufis@racai.ro

Abstract. Most of the successful commercial applications in language processing (text and/or speech) dispense of any explicit concern on semantics, with the usual motivations stemming from the computational high costs required by dealing with semantics in case of large volumes of data. With recent advances in corpus linguistics and statistical-based methods in NLP, revealing useful semantic features of linguistic data is becoming cheaper and cheaper and the accuracy of this process is steadily improving. Lately, there seems to be a growing acceptance of the idea that multilingual lexical ontologies might be the key towards aligning different views on the semantic atomic units to be used in characterizing the general meaning of various and multilingual documents. Depending on the granularity at which semantic distinctions are necessary, the accuracy of the basic semantic processing (such as word sense disambiguation) can be very high with relatively low complexity computing. The paper substantiates this statement by presenting a statistical/based system for word alignment (WA) and word sense disambiguation (WSD) in parallel corpora.

1 Introduction

Most difficult problems in natural language processing stem from the inherent ambiguous nature of the human languages. Ambiguity is present at all levels of traditional structuring of a language system (phonology, morphology, lexicon, syntax, semantics) and not dealing with it at the proper level, exponentially increases the complexity of the problem solving. Currently, the state of the art taggers (combining various models, strategies and processing tiers) ensure no less than 97-

Please use the following format when citing this chapter:
Tufis, Dan, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 575–582

98% accuracy in the process of morpho-lexical full disambiguation. For such taggers a 2-best tagging[1] is practically 100% accurate.

One further step is the word sense disambiguation (WSD) process. In the fregean compositional semantics, the meaning of a complex expression is supposed to be derivable from the meanings of its parts, and the way in which those parts are combined. Depending on the representation formalisms for the word-meaning representation, various calculi may be considered for computing the meaning of a complex expression from the atomic representations of the word senses. Obviously, one should be able, before hand, to decide for each word in a text which of its possible meanings is the contextually right one.

Therefore, it is a generally accepted idea that the WSD task is highly instrumental (if not indispensable) in semantic processing of natural language documents.

Considering the word senses, one calls upon an informal concept, namely the *context* of a word. The definition of *NLP context* is rarely made independent of an intended application and thus, it is very hard to find a generally acceptable formalization for this concept. The context of a targeted word is its vicinity. This vicinity can be a sequence (ordered or not) of orthographic words in a limited window (not very well linguistically motivated), a typographical unit (sentence, paragraph, section, chapter, or even the entire document), a sequence of linguistically interpreted (morphology, syntax, semantics) atoms making a coherent unit (phrase, sentence, chain of sentences). In the case of multilingual environments or, more precisely, parallel corpora (as is the case in this paper), the context is defined as the pair of the sentences which are mutual translations.

The WSD problem can be stated as being able to associate to an ambiguous word (w) in a text or discourse, the sense (s_k) which is distinguishable from other senses (s_1, …, s_{k-1}, s_{k+1}, …, s_n) prescribed for that word by a reference semantic lexicon. One such semantic lexicon (actually a lexical ontology) is Princeton WordNet [1] version 2.0[2] (henceforth PWN). PWN is a very fine-grained semantic lexicon currently containing 203,147 sense distinctions, clustered in 115,424 equivalence classes (synsets). Out of the 145,627 distinct words, 119,528 have only one single sense. However, the remaining 26,099 words are those that one would frequently meet in a regular text and their ambiguity ranges from two senses up to 36. Several authors considered that sense granularity in PWN is too fine-grained for the computer use, arguing that even for a human (native speaker of English) the sense differences of some words are very hard to be reliably (and systematically) distinguished. There are several attempts to group the senses of the words in PWN in coarser grained senses – *hyper-senses* – so that clear-cut distinction among them is always possible for humans and (especially) computers. We will refer in this paper to two hyper-sense inventories used in the BalkaNet project [2]. A comprehensive review of the WSD state-of the art at the end of 90's can be found in [3]. Stevenson and Wilks [4] review several WSD systems that combined various knowledge

[1] In k-best tagging, instead of assigning each word exactly one tag (the most probable in the given context), it is allowed to have occasionally at most k-best tags attached to a word and if the correct tag is among the k-best tags, the annotation is considered to be correct.

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

sources to improve the disambiguation accuracy and address the issue of different granularities of the sense inventories. SENSEVAL[3] series of evaluation competitions on WSD is a very good source on learning how WSD evolved in the last 6-7 years and where is it nowadays.

We describe a multilingual environment, containing several monolingual wordnets, aligned to PWN used as an interlingual index (ILI). The word-sense disambiguation method combines word alignment technologies, and interlingual equivalence relations in multilingual wordnets [5]. Irrespective of the languages in the multilingual documents, the words of interest are disambiguated by using the same sense-inventory labels. The aligned wordnets were constructed in the context of the European project BalkaNet. The consortium developed monolingual wordnets for five Balkan languages (Bulgarian, Greek, Romanian Serbian, and Turkish) and extended the Czech wordnet initially developed in the EuroWordNet project [5]. The wordnets are aligned to PWN, taken as an interlingual index, following the principles established by the EuroWordNet consortium. The version of the PWN used as ILI is an enhanced XML version where each synset is mapped onto one or more SUMO [6] conceptual categories and is classified under one of the IRST domains [7]. In the present version of the BalkaNet ILI there are used 2066 SUMO distinct categories and 163 domain labels. Therefore, for our WSD experiments we had at our disposal three sense-inventories, with very different granularities: PWN senses, SUMO categories and IRST Domains.

2 Word Alignment

The word alignment is the first step (the hardest) in our approach for the identification of word senses. In order to reduce the search space and to filter out significant information noise, the context is reduced to the level of sentence. Therefore, a parallel text <T_{L1} T_{L2}> is represented as a sequence of pairs of one or more sentences in language L1 (S_{L1}^{1} S_{L1}^{2}...S_{L1}^{k}) and one or more sentences in language L2 (S_{L2}^{1} S_{L2}^{2}...S_{L2}^{m}) so that the two ordered sets of sentences represent reciprocal translations. Such a pair is called a translation alignment unit (or translation unit). The word alignment of a bitext is an explicit representation of the pairs of words <w_{L1} w_{L2}> (called translation equivalence pairs) co-occurring in the same translation units and representing mutual translations. The general word alignment problem includes the cases where words in one part of the bitext are not translated in the other part (these are called *null alignments*) and 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 input format is obtained from two raw texts that represent reciprocal translations. If not already sentence aligned, the two texts are aligned by a sentence aligner, similar to Moore's aligner [8] but which unlike it, is able to recover the non-one-to-one sentence alignments. The texts in each language are then tokenized, tagged and lemmatized. Frequently, the translation equivalents have the same part-of

[3] http://www.cs.unt.edu/~rada/senseval

speech, but relying on such a restriction would seriously affect the alignment recall. However, when the translation equivalents have different parts of speech, this difference is not arbitrary. *POS affinities,* $\{p(POS_m^{RO}|POS_n^{EN})\}$ and $p(POS_n^{EN}|POS_m^{RO})\}$, are easy to estimate and we use them to filter out improbable translation equivalents pairs.

The next pre-processing step is represented by the sentence chunking in both languages. The chunks are recognized by a set of regular expressions defined over the tagsets and they correspond to (non-recursive) noun phrases, adjectival phrases, prepositional phrases and verb complexes (analytical realization of tense, aspect mood and diathesis and phrasal verbs). The texts are further processed by a statistical dependency linking parser. Finally, the bitext is assembled as an XML document (XCES[4] compliant format), which is the standard input for most of our tools.

The proper word alignment process is achieved by a statistics-based module, named *COWAL* [9]. The alignment model considers a link between two candidate words as an object that is described by a feature-values structure which we call the *reification* of the link. The program starts building the most probable links *(anchor links)*: cognates, numbers, dates, and translation pairs with high translation probabilities. Then, it iteratively aligns content words (open class categories) in the immediate vicinity of the anchor links. The links to be added at any later step are supported or restricted by the links created in the previous iterations. The aligner has different weights and different significance thresholds on each feature and iteration. Each of the iterations can be configured to align different categories of tokens (named entities, dates and numbers, content words, functional words, punctuation) in decreasing order of statistical evidence.

A link between two tokens is characterized by a set of features with values in the [0,1] interval. The score of a candidate link (LS) between a source token *i* and a target token *j* is computed by a linear function of several features scores:

$$\boldsymbol{LS(i,j) = \sum_{i=1}^{n} \lambda_i * ScoreFeat_i \; ; \sum_{i=1}^{n} \lambda_i = 1}$$

Although far from being perfect, the accuracy of word alignments and of the translation lexicons extracted from parallel corpora is rapidly improving. In the shared task evaluations of different word aligners, organized on the occasion of the 2003 Conference of the North American Association for Computational Linguistics and the 2005 Conference of the Association for Computational Linguistics, our winning systems[5] TREQ-AL[10] and COWAL produced wordnet-relevant translation lexicons[6] with an F-measure as high as 84.26% and respectively 89.92%.

The major features used by the COWAL aligner are briefly discussed below.

Translation equivalence. The word aligner invokes GIZA++ [11] to build translation probability lists for either lemmas or the occurrence forms of the bitext. The considered token for the translation model build by GIZA++ is the respective

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

[5] We participated only in the Romanian-English track (some other pairs of languages were French-English (in 2003) and Hindi-English and Inuktitut-English (in 2005).

[6] wordnet-relevant dictionaries are restricted only to translation pairs of the same major POS (nouns, verbs, adjectives and adverbs).

lexical item (lemma or wordform) trailed by its POS tag (eg. plane_N, plane_V plane_A). In this way we avoid data sparseness and filter noisy data. A further way of removing the noise created by GIZA++ is to filter out all the translation pairs below a LL-threshold. We made various experiments and empirically set the value of this threshold to 6. All the probability losses by this filtering were redistributed proportionally to their initial probabilities to the surviving translation equivalence candidates.

Translation equivalence entropy score. The translation equivalence entropy score is a favouring parameter for the words which have few high probability translations. Since this feature is definitely sensitive to the order of the lexical items, we compute an average value for the link: αES(A)+βES(B). Currently we use α=β=0.5, but it might be interesting to see, depending on different language pairs, how the performance of the aligner would be affected by a different settings of these parameters.

$$ES(W) = 1 - \frac{-\sum_{i=1}^{N} p(W,TR_i) * \log p(W,TR_i)}{\log N}$$

Part-of-speech affinity. In faithful translations the translated words tend to be translated by words of the same part-of-speech. When this is not the case, the different POSes, are not arbitrary. The part of speech affinity, P(cat(A)|cat(B)), can be easily computed from a gold standard alignment. Obviously, this is a directional feature, so an averaging operation is necessary in order to ascribe this feature to a link: PA=αP(cat(A)|cat(B)) + βP(cat(B)|cat(A)). Again, we used α=β=0.5 but different values of these weights might be worthwhile investigating.

Cognates. The similarity measure, COGN(T_S, T_T), is implemented as a Levenstein metric. Using the COGN test as a filtering device is a heuristic based on the *cognate conjecture* which says that when the two tokens of a translation pair are orthographically similar, they are very likely to have similar meanings (i.e. they are cognates).

Obliqueness. Each token in both sides of a bi-text is characterized by a position index, computed as the ratio between the relative position in the sentence and the length of the sentence. The absolute value of the difference between tokens' position indexes, subtracted from 1 gives the link's "obliqueness".

$$OBL(SW_i, TW_j) = 1 - \left| \frac{i}{length(Sent_S)} - \frac{j}{length(Sent_T)} \right|$$

Locality. Locality is a feature that estimates the degree to which the links are sticking together. There are three features to account for locality: (i) *weak locality*, (ii) *chunk-based locality* and (iii) *dependency-based* locality.

The value of the *weak locality* feature is derived from the already existing alignments in a window of N tokens centred on the focused token. The window size is variable, proportional to the sentence length. If in the window there exist k linked tokens and the indexes of their links are $<i_1\ j_1>$, $\ldots<i_k\ j_k>$ then the locality feature of the new link $<i_{k+1}, j_{k+1}>$ is defined by the equation below:

$$LOC = 1 - \min(1, \frac{1}{k}\sum_{m=1}^{k}\frac{|i_{k+1} - i_m|}{|j_{k+1} - j_m|})$$

In the case of *chunk-based locality* the window span is given by the indexes of the first and last tokens of the chunk.

Dependency-based locality uses the set of the dependency links of the tokens in a candidate link for the computation of the feature value. In this case, the LOC feature of a candidate link <i_{k+1}, j_{k+1}> is set to 1 or 0 according to the following rule:

if between i_{k+1} and i_α there is a (source language) dependency and if between j_{k+1} and j_β there is also a (target language) dependency then LOC is 1 if i_α and j_β are aligned, and 0 otherwise. Please note that in case $j_{k+1} \equiv j_\beta$ a trivial dependency (identity) is considered and the LOC attribute of the link <i_{k+1}, j_{k+1}> is set to always to 1 (thus enabling a many to one word alignment).

Collocation. Bi-gram lists (only content words) were built from each monolingual part of the training corpus, using the log-likelihood score (threshold of 10) and minimal occurrence frequency (3) for candidates filtering.

We used the bi-grams list to annotate the chains of lexical dependencies among the contents words. Then, the value of the collocation feature is computed similar to the dependency-based locality feature. The algorithm searches for the links of the lexical dependencies around the candidate link.

3 Wordnet-based Sense Disambiguation

Once the translation equivalents identified, it is reasonable to expect that the words of a translation pair <w^i_{L1}, w^j_{L2}> share at least one conceptual meaning stored in an interlingual sense inventory. As we mentioned in the introduction, the most generally used sense inventory is represented by the set of unique identifiers of the synsets in Princeton Wordnet. When interlingually aligned wordnets are available (as is our case), obtaining the sense labels for the words in a translation pair is straightforward: one has to identify for w^i_{L1} the synset S^i_{L1} and for w^j_{L2} the synset S^j_{L2} so that S^i_{L1} and S^j_{L2} are projected over the same interlingual concept. The index of this common interlingual concept (ILI) is the sense label of the two words w^i_{L1} and w^j_{L2}. However, it is possible that no common interlingual projection will be found for the synsets to which w^i_{L1} and w^j_{L2} belong. In this case, the senses of the two words will be given by the indexes of the most similar interlingual concepts corresponding to the synsets of the two words. Our measure of interlingual concepts semantic similarity is based on PWN structure. We compute the semantic-similarity[7] score by the formula $SYM(ILI_1, ILI_2) = \frac{1}{1+k}$ where k is the number of links from ILI_1 to ILI_2 or from both ILI_1 and ILI_2 to the nearest common ancestor.

[7] For a detailed discussion and an in-depth analysis of several other measures see [12]

4 Evaluation

The BalkaNet version of the "1984" corpus is encoded as a sequence of uniquely identified translation units. For the evaluation purposes, we selected a set of frequent English words (123 nouns and 88 verbs) the meanings of which were also encoded in the Romanian wordnet. The selection considered only polysemous words (at least two senses per part of speech) since the POS-ambiguous words are irrelevant as this distinction is solved with high accuracy (more than 99%) by our tiered-tagger [13]. All the occurrences of the target words were disambiguated by three independent experts who negotiated the disagreements and thus created a gold-standard annotation for the evaluation of precision and recall of the WSD algorithm. The table below summarizes the results.

Table 1. WSD precision, recall and F-measure

Precision	Recall	F-measure
76.12%	76.12%	76.12%

With the PWN senses identified (synset unique identifiers), sense labeling with either SUMO and/or IRST domains inventories is trivial: as we said in section 1, the synset unique identifiers of PWN are already mapped (clustered) onto these two sense inventories. The Table 2 shows a great variation in terms of Precision, Recall and F-measure when different granularity sense inventories are considered for the WSD problem. Therefore, it is important to make the right choice on the sense inventory to be used with respect to a given application. In case of a document classification problem, it is very likely that the IRST domain labels (or a similar granularity sense inventory) would suffice. The rationale is that IRST domains are directly derived from the Universal Decimal Classification as used by most libraries and librarians. The SUMO sense labeling will be definitely more useful in an ontology based intelligent system interacting through a natural language interface. Finally, the most refined sense inventory of PWN will be extremely useful in Natural Language Understanding Systems, which would require a deep processing. Such a fine inventory would be highly beneficial in lexicographic and lexicological studies.

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

Sense Inventory	Precision	Recall	F-measure
PWN 115424 categories	76.12%	76.12%	76.12%
SUMO 2066 categories	82.64%	82.64%	82.64%
DOMAINS 163 categories	91.90%	91.90%	91.90%

Similar findings on sense granularity for the WSD task are discussed in [4] where for some coarser grained inventories even higher precisions are reported. However, we are not aware of better results in WSD exercises where the PWN sense inventory was used. The major explanation for this is that unlike the majority work in WSD that is based on monolingual environments, we use for the definition of sense contexts the cross-lingual translations of the occurrences of the target words. The way one word in context is translated into one or more other languages is a very accurate and highly discriminative knowledge source for the decision-making.

5. Conclusions

The results in Table 2 show that although we used the same WSD algorithm on the same text, the performance scores (precision, recall, f-measure) significantly varied, with more than 15% difference between the best (DOMAINS) and the worst (PWN) f-measures. This is not surprising, but it shows that it is extremely difficult to objectively compare and rate WSD systems working with different sense inventories.

The potential drawback of this approach is that it relies on the existence of parallel data and at least two aligned wordnets that might not be available yet. Nevertheless, parallel resources are becoming increasingly available, in particular on the World Wide Web, and aligned wordnets are being produced for more and more languages (currently there are more than 40 ongoing wordnets projects for 37 languages). In the near future it should be possible to apply our and similar methods to large amounts of parallel data and a wide spectrum of languages.

References

1. Fellbaum, Ch. (ed.) WordNet: An Electronic Lexical Database, MIT Press (1998).
2. Tufiş, D. (ed): Special Issue on BalkaNet. Romanian Journal on Science and Technology of Information, Vol. 7 no. 3-4 (2004) 9-44.
3. Ide, N., Veronis, J., Introduction to the special issue on word sense disambiguation. The state of the art. Computational Linguistics, Vol. 27, no. 3, (2001) 1-40.
4. Stevenson, M., Wilks, Y., The interaction of Knowledge Sources in Word Sense Disambiguation. Computational Linguistics, Vol. 24, no. 1, (1998) 321-350.
5. Vossen P. (ed.) A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht, 1998
6. Niles, I., and Pease, A., Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine, (2001) 17-19.
7. Magnini B. Cavaglià G., Integrating Subject Field Codes into WordNet. In Proceedings of LREC2000, Athens, Greece (2000) 1413-1418.
8. Moore, R. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora in Machine Translation: From Research to Real Users. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany: 135-244.
9. Tufiş, D., Ion, R. Ceauşu, Al., Stefănescu, D.: Combined Aligners. In *Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*. June, 2005, *Ann Arbor, Michigan, June,* Association for Computational Linguistics, pp. 107-110.
10. 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 (2003) 36-39.
11. Och, F., J., Ney, H., Improved Statistical Alignment Models, *Proceedings of ACL2000*, Hong Kong, China, 440-447, 2000.
12. Budanitsky, A., Hirst, G., Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Proceedings of the Workshop on WordNet and Other Lexical Resources, NAACL, Pittsburgh, June, (2001) 29-34.
13. Tufiş, D., Tiered Tagging and Combined Classifiers, in F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence, Vol. 1692. Springer-Verlag, Berlin Heidelberg New-York (1999) 28-33.