The Role of WSD for Multilingual Natural Language Applications

Andrés Montoyo, Rafael Romero, Sonia Vázquez, Carmen Calle and Susana Soler

Research Group of Language Processing and Information Systems
Department of Software and Computing Systems
University of Alicante, Alicante, Spain
{montoyo, romero, susana}@dlsi.ua.es

Abstract. Nowadays, the need of advanced free text filtering in multilingual environment is increasing. Therefore, when searching for specific keywords in multilingual information space, it is desirable to eliminate occurrences where the word or words of each language are used in an inappropriate sense. This task could be exploited in internet browsers, and resource discovery systems, relational databases containing free text fields, electronic document management systems, data warehouse and data mining systems, etc. In order to resolve this problem in this paper we present a Word Sense Disambiguation interface, which it returns the words senses in different languages and it could be employed for multilingual natural language applications. This interface resolve lexical ambiguity of nouns and verbs in some European languages (English, Spanish) input texts, using the taxonomy of the EuroWordNet lexical knowledge database, and returning a multilingual output of the words senses (English, Spanish, Catalan and Basque). In addition to the relations in WordNet 1.5, EuroWordNet includes cross-language and cross-category relations, which are directly useful for multilingual Word Sense Disambiguation. This interface has been implemented using programming language C++ and providing a visual framework.

1 Introduction and Motivation

The development and convergence of computing, telecommunications and multilingual information systems has already led to a revolution in the way that we work, communicate with each other in different countries and different languages, buy goods and use services, and even in the way we entertain and educate ourselves. The revolution continues and one of its results is that large volumes of multilingual information will increasingly be held in a form which is more natural for users than the data presentation formats typical of computer systems of the past. Natural Language Processing (NLP) is crucial in solving these problems and language technologies will make an indispensable contribution to the success of the information systems.

Designing a system for NLP requires abundant knowledge on language structure, morphology, syntax, semantics and pragmatic nuances. Morphological knowledge

provides the tools for building words, while syntactic knowledge combines words to form sentences. Semantic knowledge provides the meaning of a given word, and pragmatic knowledge helps us to interpret the complete sentence in its true context. All of these different linguistic knowledge forms, however, have a common associated problem, their many ambiguities, which is difficult to resolve. One of the main objectives in designing any NLP system, therefore, is the resolution of ambiguity. Furthermore, each type of ambiguity, whether it be structural, lexical, quantifying, contextual or referential, requires its specific resolution procedure.

This paper is motivated by two reasons. First, we concentrate on the design and implementation of an interface to resolve the lexical ambiguity that arises when a given word has several different meanings. This specific task is commonly referred to as Word Sense Disambiguation (WSD) [1]. In general terms, WSD involves assigning a definition to a given word, in either a text or a discourse, that endows it with a meaning that distinguishes it from all of the other possible meanings that the word might have in other contexts.

Second, the WSD interface we propose in this paper, will allow a tool to disambiguate the content words of different languages by matching the context in which they appear with information from an external knowledge source (knowledge-driven WSD). This tool will also provide support across different languages converting the words senses from one language into other languages. It will also could be used as a basic resource for supporting other multilingual natural language applications, such as Machine Translation (MT), Question Answering, Information Retrieval (IR), hypertext navigation, etc.

In order to accomplish these tasks, we use the following two different WSD methods for disambiguating nouns and verbs of texts: Specification Marks Method [6, 7] for nouns and Semantic Similarity [11] for verbs. Both methods use EuroWordNet [13] as it combines the features of both dictionaries and thesauruses, and also includes other links among words by means of several semantic relations, (Hyponymy, hypernymy, meronymy, etc). In other words, EuroWordNet provides definitions for the different senses that a given word might have (as a dictionary does) and defines groups of synonymous words by means of "Synsets", which represent distinct lexical concepts, and organises them into a conceptual hierarchy (as a thesaurus does). EuroWordNet also contains multilingual relations from each monolingual wordnet (Spanish, Catalan, etc) to English (WordNet 1.5 [5]).

The organization of this paper is as follows: after this introduction, in Section 2 we describe the architecture employed in developing the multilingual natural language interface. In Section 3, we describe the two WSD methods used and its application. In Section 4, we describe the WSD interface which allows the disambiguation of words from one language into other languages. And finally, discussions, conclusions and an outline of further lines of research are shown.

2 Architecture employed

In this section we describe, in detail, the architecture employed in developing the WSD interface for multilingual natural language applications. Figure 1 illustrates this architecture.

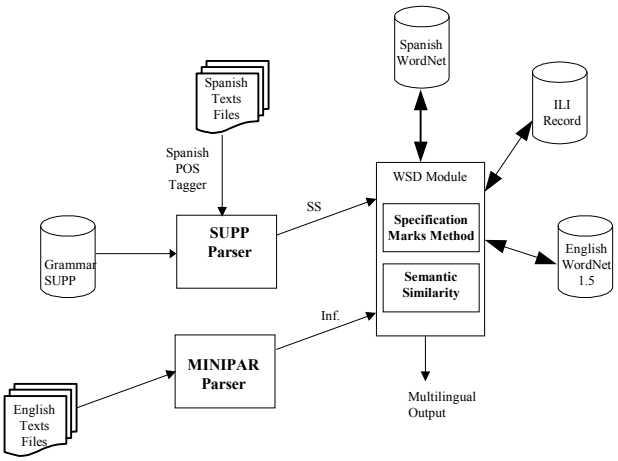

Figure 1. Architecture

The input text (Spanish or English) that is to be disambiguated come from different files and are passed through a preprocessing stage.

If the input text is Spanish, the first step in preprocessing consist of using a Spanish part-of-speech (POS) tagger [10] to automatically assign lexical tags to the text. Next, a Spanish SUPP Partial Parser [3] is used to extract constituents like noun phrase (NP), prepositional phrase (PP) and verbal chunks. This parser also includes some heuristics to provide functional categories, such as subject and object. A grammar (SUG) for Spanish that recognises every syntactic constituent (NP, PP, verbal chunks) is first defined. Our process, however, only uses the NP to disambiguate the nouns of a sentence and the verbal chunk with its functional category to disambiguate the verbs. This grammar is automatically translated into Prolog clauses. The translator will provide a Prolog program that can parse sentences. The program will return structure (SS) for each parsed sentence. This SS stores the syntactic, morphological and semantic information of the NP and verbal chunks constituents. After a sentence has been parsed, its SS will be the input for the WSD module.

If the input text is English, the first step in preprocessing consist of using MINIPAR parser [2] to extract the same constituents as in Spanish process. After a sentence has been parsed, this information will be the input for the WSD module.

The WSD module is composed of two different WSD methods: Specification Marks Method for disambiguating nouns and Semantic Similarity for disambiguating verbs. This module will consult the EuroWordNet[1] knowledge database for nouns and

[1] EuroWordNet is a multilingual database resembling WordNet that stores semantic relations between words in four different languages of the European Community: Dutch, Italian,

verbs that appear in the context, returning all of their possible senses. The WSD module will then select the correct method for disambiguating the nouns or verbs and after these methods will be applied a multilingual output will be returned, in which the words have the correct sense assigned.

We should like to emphasise that this resolution skill allows us to produce modular NLP systems in which the grammatical rules, the knowledge base (EuroWordNet), the parsing scheme and the WSD module are all quite independent of one other.

3 Word Sense Disambiguation Modules

The Word sense disambiguation module employed in this paper is composed of two different WSD methods. This two method are Specification Marks Method for disambiguating nouns and Semantic Similarity Method for disambiguating verbs. We should like to clarify, these two methods employ EuroWordnet for disambiguating the words senses. EuroWordNet lexical database consists on a WordNet-like database for each of the language covered, linked through a common interlingual index. Then, if we map a word in Spanish against its corresponding synset in the Spanish WordNet, we can obtain its Interlingual Index (ILI) record number. This number is the language-neutral semantic representation of the Spanish word. Mapping all the words into their corresponding synsets will provide us with the Spanish and English words senses. In this section, we describe each one of this method.

3.1 Specification Marks Method

WSD with Specification Marks is a method for the automatic resolution of lexical ambiguity of groups of words, whose different possible senses are related. The disambiguation is resolved with the use of the EuroWordNet lexical knowledge base. The method requires the knowledge of how many of the words are grouped around a specification mark, which is similar to a semantic class in the monolingual WordNet taxonomy. The word-sense in the sub-hierarchy that contains the greatest number of words for the corresponding specification mark will be chosen for the sense-disambiguating of a noun in a given group of words.

Specification Marks Method consists basically of the automatic sense disambiguating of nouns that appear within the context of a sentence and whose different possible senses are related. Its context is the group of words that co-occur with it in the sentence and their relationship to the noun to be disambiguated. The input for the WSD algorithm will be the group of words $w=\{w_1, w_2, ..., w_n\}$. Each word w_i is sought in monolingual WordNet, each one has an associated set $s_i=\{s_{i1}, s_{i2}, ..., s_{in}\}$ of possible senses. Furthermore, each sense has a set of concepts in the IS-A taxonomy (hypernym/hyponym relations). First, the concept that is common to all the senses of all the words that form the context is sought. We call this concept the Initial

---

Spanish and English. It contains multilingual relations from each individual wordnet to English (WordNet 1.5). Such relations will form an Interlingual Index (ILI).

Specification Mark (ISM), and if it does not immediately resolve the ambiguity of the word, we descend from one level to another through monolingual WordNet´s hierarchy, assigning new Specification Marks. The number of concepts that contain the sub-hierarchy will then be counted for each Specification Mark. The sense that corresponds to the Specification Mark with highest number of words will then be chosen as the sense disambiguation of the noun in question, within its given context.

We should like to point out that after having evaluated the method, we subsequently discovered that it could be improved with a set of heuristics, providing even better results in disambiguation. The set of heuristics are Heuristic of Hypernym, Heuristic of Definition, Heuristic of Common Specification Mark, Heuristic of Gloss Hypernym, Heuristic of Hyponym and Heuristic of Gloss Hyponym. Detailed explanation of the method can be found in [6, 7, 8], while its application to NLP tasks are addressed in [9].

3.2 Semantic Similarity Method

The method of WSD proposed in this paper is based on knowledge and consists basically of sense-disambiguating of the verb that appear in an Spanish or English sentence.

A simple sentence or question can usually be briefly described by an action and an object [4]. For example the main idea from the sentence "*He eats bananas"* can be described by the action-object pair *"eat-banana"*. Our method determine which senses of these two words are more similar between themselves.

For this task we use the concept of semantic similarity [12] between nouns based on EuroWordNet hierarchy. In EuroWordNet, the gloss of a verb synset provides a noun-context for that verb, i.e. the possible nouns occurring in the context of that particular verb [4]. The glosses are used here in the same way a corpus is used.

Our method takes into consideration the verb-noun pair extracted from the sentence. This verb-noun pair is the input for the algorithm. The output will be the sense tagged verb-noun pair, so we assign the sense of the verb. The algorithm is described as follows:

Step 1. Determine all the possible senses for the verb and the noun by using WordNet. Let us denote them by $<v_1, v_2, \ldots, v_k>$ and $<n_1, n_2, \ldots, n_m>$

Step 2. For each sense of verb v_h and all senses of noun $<n_1, n_2, \ldots, n_m>$:

2.1. Extract all the glosses from the sub-hierarchy including v_h. The sub-hierarchy including a verb v_h is determined as follows: consider the hypernym h_h of the verb v_h and consider the hierarchy having h_h as top [4].

2.2. Determine the nouns from these glosses. These constitute the noun-context of the verb. Determine all the possible senses for all these nouns. Let us denote them by $<x_1, x_2, \ldots, x_n>$.

2.3. Then we obtain the similarity matrix (Sm) using the semantic similarity, where each element is defined as follows:

$$Sm(i, j) = sim (x_i, n_j)$$

For determining the semantic similarity (sim(x_i, n_j)) between each sense of the nouns extracted from the gloss of verb and each sense of the input noun, we use the formula followed:

$$\text{sim}\ (x_i, n_j) = 1 - \text{sd}\ (x_i, n_j)^2$$

$$\underline{\text{sd}}\ (x_i, n_j) = \frac{1}{2} \cdot \left(\frac{D1 - D}{D1} + \frac{D2 - D}{D2}\right)$$

where sim (x_i, n_j) is the semantic similarity between two concepts defined by their WordNet synsets x_i and n_j; sd (x_i, n_j) is the semantic distance for nouns. D1 is the depth of synset x_i, D2 is the depth of synset n_j, and D is the depth of their nearest common ancestor in the monolingual WordNet hierarchy.

2.4. Determine the total similarity between the sense *h* of verb (v_h) and all the senses of input noun <n_1, n_2, …, n_m>. For each n_j:

$$\text{Ts}(h, j) = \sum_{i=1}^{n} \text{sim}\ (x_i, n_j)$$

where n is the number of nouns extracted from the gloss of the sense *h* of the verb.

Step 3

To resume all similarity matrixes (Sm) obtained in step2 for each sense of verb, we make now the total similarity matrix (Tsm) composed by total similarity (Ts) for each sense of verb and each sense of noun. Each element of this matrix is defined as follows:

$$\text{Tsm}\ (i, j) = \text{Ts}\ (i, j)$$

Step 4

The most similar sense combination scores the highest value in the total similarity matrix (Tsm). So the output of the algorithm is the pair verb-noun (v_i-n_j) that contains this value in the matrix. Therefore the sense of the verb is chosen and given as the solution.

4 Multilingual Natural Language Interface

In order to resolve the lexical ambiguity in Spanish and English texts it is necessary the creation of an interface to disambiguate the words senses that appear in the context of a sentence. This interface is made up of a set of computer programs that do all the processing to assign ultimately the correct sense to the Spanish and English words.

First, users introduce a Spanish or English sentence in the WSD interface. This sentence is sent to the parser for starting the analysis process. After, a process checks the information returned by the parser and endows it the appropriate structure for their handling for the WSD module and insert it in a file. This file is the input data to the WSD process that carries out the disambiguation of the text based in Specification Marks and Semantic similarity methods and using the lexical database EuroWordNet. Finally, when the WSD process concludes another process formats the information

disambiguated and this information is sent to the interface in order to show it to the user.

This interface is illustrate in the figure 2. The user interface offers the operations followed:

Run WSD process. The command button WSD allows one to run the lexical ambiguity methods (Specification Marks and Semantic Similarity) from the input sentence and the words senses is returned in Spanish and English languages.

Clear Process. The user clicks on this command button to delete the information that appears in both text windows.

Sometimes one or more words cannot be disambiguated, then this kind of words is shown in both texts window below command buttons preceded by the symbol asterisk (*). In this case it is shown all the possible senses of the word.

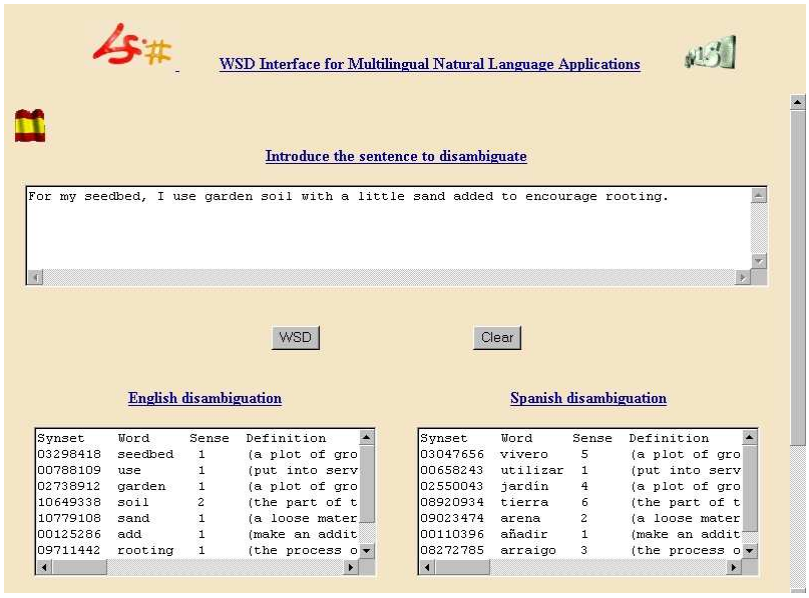

Figure 2: WSD Interface

5 Discussions and Conclusions

This paper presents a WSD interface, which it could be very useful for multilingual natural language applications, because it returns the words senses in different languages. This interface use the taxonomy of EuroWordNet lexical knowledge database to resolve lexical ambiguity of nouns and verbs in Spanish and English input texts. WSD module uses two different WSD methods for disambiguating the nouns

and verbs of the texts: Specification Marks Method for nouns and Semantic Similarity for verbs, respectively. The University of Alicante system presented at Senseval-2 workshop joins these two methods in the WSD task. Specification Marks for nouns, and Semantic Similarity for verbs had been used in order to process the test data of English lexical sample task and Spanish lexical sample task. The system obtains a successful score when comparing with the evaluation results of other unsupervised systems.

A relevant consequence of the application of this interface is that provide support across different languages converting the words senses from one language into other languages. Besides, it will also could be used as a basic resource for supporting other natural language applications, such as Machine Translation (MT), Question Answering, Information Retrieval (IR). For example, an interesting possibility is to use this multilingual interface to improve cross languages information retrieval systems.

References

1. Ide N. and Véronis J. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics* **24** (1): 1-40.
2 . Lin, D. 1998. Dependency-based Evaluation of MINIPAR In *Workshop on the Evaluation of Parsing Systems*, Granada, Spain, May, 1998.
3. Ferrández A., Palomar M., Moreno L. 1997. Slot Unification Grammar. *In Proceedings of Joint Conference On Declarative Programming* (APPIA-GULP-PRODE'97). Grado, Italy. June.
4. Mihalcea R. and Moldovan D. 1999. A Method for word sense disambiguation of unrestricted text. *Proc. 37th Annual Meeting of the ACL* 152-158*,* Maryland, Usa.
5. Miller G. A., Beckwith R., Fellbaum C., Gross D., and Miller K. J. 1990. WordNet: An on-line lexical database*. International Journal of Lexicography* **3**(4): 235-244.
6. Montoyo, A. and Palomar M. 2000. WSD Algorithm applied to a NLP System. *In Proceedings 5th International Conference on Application of Natural Language to Information Systems (NLDB´2000)*. Versailles, (France).
7. Montoyo, A. and Palomar M. 2000. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. *In Proceedings 11th International Workshop on Database and Expert Systems Applications (DEXA 2000)*, pages 103-108. Greenwich, (London).
8. Montoyo, A. and Palomar, M. 2001. Specification Marks for Word Sense Disambiguation: New Development. *2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001)*. México D.F. (México).
9. Palomar M., Saiz-Noeda M., Muñoz, R., Suárez, A., Martínez-Barco, P., and Montoyo, A. 2001. PHORA: NLP System for Spanish. *In Proceedings 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001)*. México D.F. (México).
10. Pla F. 2000. Etiquetado Léxico y Análisis Sintáctico Superficial basado en Modelos Estadísticos. *PhD Thesis*. Departamento de Sistemas Informáticos y Computación. Universidad de Politécnica de Valencia.
11. Soler, S. and Montoyo, A. 2002. A Proposal for WSD using Semantic Similarity. *In Proceedings Third International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*. México D.F. (México).

12. Stetina J., Kurohashi S. and Nagao M. (1998) General word sense disambiguation method based on full sentencial context. *In Usage of WordNet in Natural Language Processing*. COLING-ACL Workshop, Montreal, Canada.
13. Vossen, P. 1996. EuroWordNet: building a multilingual wordnet database with semantic relations between words. *Technical and Financial Annex, EC funded project LE #4003*.