Semantic information in anaphora resolution

R. Muñoz, A. Montoyo and M. Saiz-Noeda

Grupo de investigación del Procesamiento del Lenguaje y Sistemas de Información.
Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. Spain
{rafael,montoyo,max}@dsli.ua.es

Abstract. Building a NLP system requires the adding of linguistic phenomena resolution. One of the most relevant tasks regarding to these phenomena is the coreference resolution. Coreference is defined as a semantic phenomenon and, therefore, apart from the morphological and syntactic information that is highly useful, adding semantic sources improves the capabilities of such a NLP system. In this paper, a complete NLP system is proposed. This system counts on a WSD module that provide semantic information needed for the coreference resolution. This coreference resolution will deal with both pronouns and definite descriptions (DD), two of the most important parts of the anaphora resolution research area. The WSD module is a variant of the Specification Marks Method [7] where for each word in a text a domain label is selected instead of a sense label.

1 Introduction

Coreference resolution consists of establishing a relation between an anaphoric expression and an antecedent. Different kinds of anaphoric expressions can be located in the text, such as pronouns, DDs, etc. Moreover, different information sources are needed in order to guarantee an adequate resolution. The majority of the anaphora resolvers, extensively cited at the literature, only use morphological and syntactic information [3, 6, 10]. In this paper, we focus on the resolution of pronouns and DDs using morphological, syntactic and semantic information.

The need of semantic information adds a new problem to be solved, Word Sense Disambiguation (WSD), which is one of the most important task for any natural language processing system. Therefore, we propose a way to deal with this problem starting with the hypothesis that many sense distinctions are not relevant for anaphora resolution [14, 5]. Moreover, we want to investigate how the polysemy reduction caused by domain clustering can help to improve the anaphora resolution. In this paper we propose to use a variant of the Specification Marks Method [7] where for each word in a text a domain label is selected instead of a sense label.

2 Preprocessing and resources

In this section we describe the tools and resources employed in developing a new method, based on semantic information, to anaphora resolution in unrestricted

Spanish texts. The Spanish text goes through a preprocessing stage. The first step in preprocessing consists of using a part-of-speech (POS) tagger to automatically assign morphological information (POS tags). Next, it also performs a surface syntactic parsing using dependency links that show the head-modifier relations between words. This kind of information is used for extracting NPs constituent parts, and these NPs are the input for a Word Sense Disambiguation module. This module returns all the head nouns with a domain sense assigned from all the head nouns that appear in the context of a sentence. Figure 1 shows WSD process and the resources used by this module:

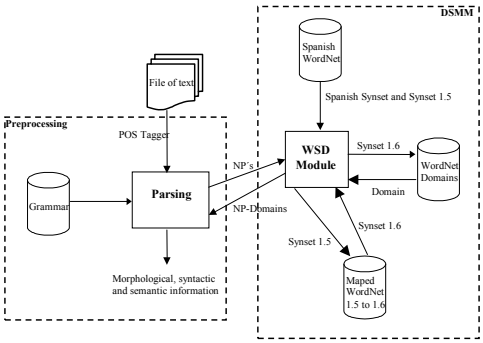

Fig. 1. *Process and resources used by WSD module*

- Spanish WordNet is a generic database with 30,000 senses. The Spanish WordNet will be linked through the English WordNet 1.5, so each English synonym will be associated with its equivalent in Spanish.
- WordNet 1.5 mapped to WordNet 1.6 is a complete mapping of the nominal, verbal, adjetival and adverbial parts of WordNet 1.5 onto WordNet 1.6 [1]
- WordNet Domain [4] is an extension of WordNet 1.6 where synsets are clustered by means of domain labels.

3 Domain Specification Marks Method

The WSD method used in this paper consists of a variant of the Specification Marks Method, which we named Domain Specification Marks Method (DSMM), where for each head noun in a text a domain label is selected instead of a sense label. The Specification Marks Method (SMM) is applied 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 Spanish WordNet lexical knowledge base. This 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 WordNet taxonomy. The word sense in the subhierarchy that contains

the greatest number of words for the corresponding Specification Mark will be chosen for the sense disambiguation of a noun in a given group of words. It has been shown in [8] that the SMM works successfully with groups of words that are semantically related. Therefore, a relevant consequence of the application of this method with domain labels is the reduction of the word polysemy (i.e. the number of domains for a word is generally lower than the number of senses for the word). That is, domain labels (i.e. Health, Sport, etc) provide a way to establish semantic relations among word senses, grouping them into clusters. Detailed explanation of the Specification Marks Method can be found in [7]. Next, we describe the steps to obtain the domain label of WordNet Domain from the word sense obtained by SMM:

1. Starting from the Spanish word sense already disambiguated by the SMM, we should obtain the corresponding synset in WordNet 1.5. For this task, the Spanish WordNet has been used to disambiguate the Spanish word sense. It allows us to calculate the intersections among the Spanish synsets and the English synsets version 1.5. For example, the output of the SMM applied to the word "planta $\rightarrow$ *plant*" is the Spanish Synset "Planta#2". As the two WordNets are linked (i.e. they share synset offsets), therefore the intersection determines the synset of WordNet 1.5, which is "Plant#2".
2. WordNet 1.5 is mapped with the WordNet 1.6, therefore the synsets obtained in step 1 are searched in this resource. Then, the synset 1.6 corresponding to the previous synset 1.5 is obtained. For example, the synset 1.5 "plant#2" is mapped to the synset 1.6 "Plant#2".
3. Finally, the synset 1.6 obtained in step 2 is searched for in the WordNet Domain, where the synsets have been annotated with one or more domain labels. For example, the synset 1.6 "00008864" belonging to the sense "plant#2" is searched for in the WordNet Domain giving the label "botany".

4 Anaphora Resolution

Anaphora is one of the most frequent linguistic phenomena used that should be solved in order to establish the coherence in a text. Different kinds of anaphoric expressions can be found in the text, such as pronouns, DDs and adjectives. Each type of anaphoric expression needs a specific way of resolution due to their different features. Pronouns and DDs are the most usual grammatical expressions used to refer to a person, and object, an event, etc. All NPs used to describe the same specific concept, the same entity of real world, will be related or closed in some way.

We illustrate in figure 2 the proposed architecture for resolving the anaphora. The following section shows both method to solve pronouns and DD using the previously described domain labels.

4.1 Pronoun Resolution

Pronominal anaphora, unlike others, requires a special treatment from the semantical point of view. In general, pronouns do not provide semantic information

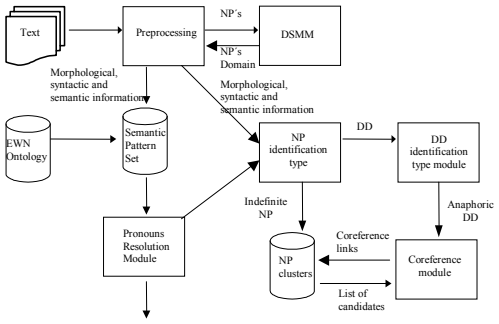

Fig. 2. *Architecture for resolving the anaphora*

at all. This fact forces the use of the semantic information provided by the verb accompanied by the anaphor. Through the verb, it is possible to establish connections between the pronoun and its possible antecedent due to the antecedent must be semantically compatible with the verb of the anaphor which refers to and in its same syntactic role. According to this, the application of semantic knowledge in pronoun resolution requires not only morphological and syntactic analysis but also semantic features related to candidate NPs and verbs. Traditional approaches based on limited knowledge have used morphological agreement and syntactic restrictions in order to reject incompatible candidates. Furthermore, some approaches have included also the semantic information defining compatibility relations between nouns (subjects and complements) and verbs through collocation patterns manually stated in order to be applied in the resolution process.

The automatic obtention of this patterns from the corpus would allow the application of the pronoun resolution method in any domain. In this way, the pronoun resolution module included in this system will deal with personal, demonstrative, reflexive and omitted pronouns. Next section will show the detailed aspects of this pronoun resolution algorithm.

Algorithm for pronoun resolution The algorithm go through different steps:

- Pronoun identification: the text has been previously morphologically tagged and syntactically parsed. This syntactic parsing includes also information about syntactic roles of the NPs and pronouns. In this step, the pronoun is detected through its morphological label. This step also includes a set of rules for the identification of omitted pronouns.
- Candidate list construction: depending on the type of pronoun, the algorithm selects the *solution searching space*, that is, the minimum portion of text where the correct antecedent should be found. According to this, the reflexive pronouns will find the solution in the same clause, while personal and demostrative pronouns will find it no more than two sentences before. All the

NPs contained in this solution searching space are added to the candidate list with all their morpho-syntactic information.
- Semantic pattern set building: while processing the corpus, all the noun in a subject, direct object or indirect object role are included with their corresponding verbs into a set of semantic patterns. Patterns are formed not only by the textual word but by the semantic domain concepts obtained from the WSD module.
- Constraint application: morphological agreement, and syntactic restrictions are applied in order to reject incompatible candidates. Classical morpho-syntactic rules are enriched with semantic information provided from Spanish WordNet regarding to *anymacy* or *group* features. The application of these enrichment is detailed in next section.
- Preference application: morphological, syntactic, structural, statistical and semantic information contributes their own features to a weighting system to decide the candidate most probable to be the correct antecedent. Next two sections will explain the semantic constraint and preference rules.

Semantic Constraints Semantic constraints, as mentioned above, are applied to morphological agreement in order to avoid the rejection of potentially correct candidates. This occurs in group names where their referent pronoun can be singular or plural (i.e. in Spanish, just as in English, although the NP 'the police' is singular, it is possible to refer to it with the pronouns 'it' or 'they', singular and plural respectively). Furthermore, semantic constraints are also useful for reject candidates that agree morphologically but that has incompatible features from the syntactical-semantic point of view. Therefore, third-person masculine and feminine personal pronoun in a subject role cannot refer to a noun-phrase without animate feature. In other languages such as English, as commented above, some pronouns provide semantic characteristics such as '+person' (he/she) or '-person' (it). The method proposed in [11] takes advantage of these features to considerably improve the anaphora resolution success rate. The semantical or ontological features used in this constraints are extracted from the set of concepts contained in the Top Ontology of EuroWordNet [13].

Semantic Preferences If the candidate list contains more than one NPs after the constraint application, the preferences will decide the most probable one to be the correct antecedent. Apart from structural preferences (the nearer or the more repeated the candidate is, the more probable it is the antecedent), morphological preferences (candidates with the same number as the anaphor are preferred) and syntactic preferences (the candidates with the same syntactic role as the anaphor are preferred) there are a group of preferences that are extracted from semantic features of nouns and verbs:

- NPs that are not of time, direction, quantity or abstract are preferred. The study of the corpus reveal that this kind of nouns are not the solution of the pronoun in almost 100% of the times. This semantic features are also extracted from the Top Ontology of EuroWordNet'.

- NPs semantically more compatible with the verb of the pronoun are preferred. For this preference, the set of learned semantic patterns is used. The semantic pattern set contains the verb and the NP in its own syntactic role (subject, direct or indirect complement). For the NP in the pattern, there is available its domain label and its ontological features according to the EuroWordNet Top Ontology. Both are possible thanks to the WSD module enriched with the DSMM. This way, the compatibility is provided in two levels: 1) It will be considered more compatible the candidate that are represented by the ontological concept most relevant (in terms of frequency) in the semantic pattern set with the syntactic role and the verb of the pronoun. 2) Furthermore, it will be considered more compatible the candidate that shares domain label with the verb of the pronoun. For determining this compatibility, the WSD module studies the nouns contained in the gloss of the verb and deduces its domain label.

All these semantic sources contribute their own weight to the pronoun resolution module for the selection of the antecedent.

4.2 Definite Description Resolution

As in pronoun resolution, DD treatment requieres the use of several information sources, among which semantic information plays an important role. For DD treatment, the semantic information is used to identify some non-anaphoric definite description, to build the list of candidates and choose the adequate antecedent from the candidate list. As in pronoun resolution, the verb of the sentence helps to choose the antecedent for a specific kind of DD: thematic role. According to [9], different kinds of DDs can be found in the text. Thematic role is a DD which antecedent is related to the verb of the sentence (the seller, to sell). Solving this kind of DD, the verb of the sentence contained the antecedent plays an important role. In order to establish this relation a lexical resource as EuroWordNet is used (involved agent and role agent relationships). Traditional approaches based on knowledge, extensively cited in the literature as [12, 2], use morphologic and syntactic information. Although Vieira and Poesio's algorithm also uses semantic information extracted from WordNet, the evaluation carried out was manually made and the scores achieved were not so successful.

DD treatment presents different features to take in account to develop an effective resolution algorithm. Three factors kept in mind in DD resolution: accessibility space, non-anaphoric identification and resolution of anaphoric DDs. Our algorithm uses three different information sourcea (morphological, syntactic and semantic) to solve the three main problems of DD resolution. The accessibility space for pronouns is only a limited number of sentences. However, the accessibility space for DDs is greater. For this reason, the number of potential candidates can be very high for very large texts. So, if the coreference algorithm compares the DD with all candidates and the number of them is high then the algorithm becomes slow. Unlike other authors that reduce the number of previous sentences to be considered as the anaphoric accessibility space, our

algorithm proposes the use of domain labels to group the NPs. This grouping is used to identify some non-anaphoric DDs (remaining non-anaphoric DDs will be classified by coreference algorithm) and to build the list of candidates for each DD. A DD looks for their antecedent along the previous NPs with the same domain label. This fact makes possible the use of a full anaphoric space made up of all previous sentences with the reduction of comparisons. The coreference algorithm provides an antecedent of a DD or if no candidate is found classifies it as non-anaphoric. The coreference algorithm is a system based on weighted heuristics. These heuristics treat the relation between heads and modifiers of both NPs (candidate and DD). Moreover, DDs can establish different kinds of relations with their antecedent. A definite description can refer to the full antecedent (identity coreference) or a part of the antecedent (part-of, set-member, set-subset). Our algorithm resolves the identity and part-of coreferences. If no candidate is selected as antecedent then the DD is re-classified as non-anaphoric. If more than one candidate is proposed then the closest candidate is selected.

Algorithm for DD resolution DD algorithm go through different steps:

- DD identification: the algorithm goes through the text (previously tagged with POS, syntactic and semantic information) extracting all NPs. Once the NP is found, its type (definite or indefinite) is identified according to the first constituent of NP.
- NPs grouping: the head of a NP contains the domain label provided by the module of Domain Specification Marks Method. This domain label is used to cluster all NPs (definite and indefinite). As the parser also tags the pronouns as NPs, the antecedent resolution of the pronoun module is also used.
- Identification of non-anaphoric DDs: previous step also helps to identify some non-anaphoric definite descriptions. If the noun phase (DD) processed cannot be grouped with previous NPs (if it is the first with a specific domain label) then the DD is classified as non-anaphoric.
- Candidate list construction: the list of candidates for DD is made up of all previous NPs with the same domain label.
- Antecedent selection: a set of heuristics using morphological and semantic information are applied in order to choose the more adequate antecedent from the list of candidates. These heuristics treat the relation between head nouns and between modifiers of anaphoric expression and candidates.

5 Experimental work

The experimental work carried out was focused on two different task: word sense disambiguation and anaphora resolution. Regarding to WSD task, SMM has been applied for the automatic resolution of lexical ambiguity of groups of words giving success rates of 70%. With reference to anaphora resolution, semantic information adding has allowed the raising of the success rates and the treatment of other types of DD, such as bridging reference. In average, success rate for anaphora resolution has been improved from 76% to the 85% [10, 11, 9].

6 Conclusions

We have presented an anaphora resolution algorithm based on semantic information to solve pronouns and DD. In addition to classical semantic information provided by the word sense number in WordNet, domain labels are used to cluster NPs. This clustering helps us to reduce the number of candidates. Experimental work shows that the use of WSD based on this domain labels improves the values of anaphora resolution task.

References

1. J. Daudé, L. Padró, and G. Rigau. A Complete WN1.5 to WN1.6 Mapping. In *Proceedings of the NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customisations.*, pages 83–88, 2001.
2. M. Kameyama. Recognizing Referential Links: An Information Extraction Perspective. In Mitkov, R. and Boguraev, B., editor, *Proceedings of ACL/EACL*, pages 46–53, 1997.
3. S. Lappin and H.J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.
4. B. Magnini and G. Cavaglia. Integrating subject field codes into WordNet. In *Proceedings of the LREC-2000*, 2000.
5. B. Magnini and C. Strapparava. Experiments in Word Domain Disambiguation for Parallel Texts. In *Proceedings of the ACL Workshop on Word Senses and Multilinguality*, 2000.
6. R. Mitkov. Robust pronoun resolution with limited knowledge. In ACL, editor, *Proceedings of the COLING-ACL'98*, pages 869–875, Montreal, Canada, 1998.
7. A. Montoyo and M. Palomar. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. In *Proceedings of the DEXA-2000*, pages 103–107. IEEE Computer Society, September 2000.
8. A. Montoyo, M. Palomar, and G. Rigau. WordNet Enrichment with Classification Systems. In *Proceedings of the NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customisations.*, pages 101–106, 2001.
9. R. Muñoz, M. Palomar, and A. Ferrández. Processing of Spanish Definite Descriptions. In O. Cairo et al., editor, *Proceeding of MICAI*, volume 1793 of *LNAI*, pages 526–537, 2000.
10. M. Palomar, A. Ferrández, L. Moreno, P. Martínez-Barco, J. Peral, M. Saiz-Noeda, and R. Muñoz. An Algorithm for Anaphora Resolution in Spanish Texts. *Computational Linguistics*, 27(4):545–567, 2001.
11. M. Saiz-Noeda, J. Peral, and A. Suárez. Semantic Compatibility Techniques for Anaphora Resolution. In *Proceedings of ACIDCA'2000*, Tunisia, 2000.
12. R. Vieira and M. Poesio. An Empiricall Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4):539–593, 2000.
13. P. Vossen, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, and W. Peters. The EuroWordNet Base Concepts and Top Ontology. Technical report, University of Amsterdam, EuroWordNet, LE2-4003 TR-11, 1998.
14. Y. Wilks and M. Stevenson. Word sense disambiguation using optimised combination of knowledge sources. In *Proceedings of COLING-ACL'98*, 1998.