Definite Description Resolution enrichment with WordNet Domain Labels

R. Muñoz, A. Montoyo, M. Palomar

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

Abstract. This paper presents a new method, based on semantic information, to resolve Definite Descriptions in unrestricted Spanish text. The method is performed in two consecutive steps. First, a lexical knowledge word sense disambiguation process where words in a text are tagged with a domain label in place of a sense label. Second, an algorithm to identify and to resolve the Spanish definite description taking advantage of domain labels. In addition, this paper presents an experimental work that will show the benefits that a Word Sense Disambiguation method produces when it is used in Definite Description (DD) resolution process. Moreover, this experimental work proves that using WordNet Domain in unsupervised WSD method improves DD resolution.

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, definite descriptions, adverbs, etc. In this paper, we focus on the treatment and resolution of definite descriptions[1].

Previos work such as [1, 11, 12] showed that most of definite descriptions in the text are non-anaphoric. The treatment of DD has been made up of two different tasks. The first one, is focused on identifying the type of DD (anaphoric or non-anaphoric). And, the second task is focused on providing the antecedent of the anaphoric DD. Definite descriptions whose antecedents are full sentences or full paragraphs are treated like non-anaphoric DDs. In this work, we only establish the coreference of DDs whose antecedents are any kind of noun phrases (indefinite, definite, entity). Previous identification of non-anaphoric DD is useful to only to apply the coreference resolution algorithm to anaphoric DDs. According to Frege [4], the identification of DD type cannot be carried out using structural information alone without comparison with previous candidates. Frege states that the reference property of a DD depends on semantic characteristics. A DD can only refer to a semantically compatible NP.

[1] We only considered as DD the noun phrases headed by a definite article (el, la, los, las → *the*) or a demonstrative (este, esta, estos, estas → *this*, *these*).

The use of semantic information is associated to Word Sense Disambiguation (WSD). In relation to the WSD task several authors [14, 7] have stated that for many applications the fine-grained sense distinctions provided by WordNet are not necessary. Therefore, we propose a way to deal with this problem starting with the hypothesis that many sense distinctions are not relevant for a DD resolution. Moreover, we want to investigate how the polysemy reduction caused by domain clustering can help to improve the DDs resolution. Because, a single domain label may group together more than one word sense, resulting in a reduction of the polysemy. Therefore, in this paper we propose to use a variant of the Specification Marks Method (SMM) [8] where for each word in a text a domain label is selected instead of a sense label.

2 Preprocessing and resources

The Spanish text that is to be treated came from different files and is passed through a preprocessing stage. The first step in preprocessing consists of using a POS-tagger to automatically assign morphological information (POS tags). Next, it also performs a surface syntactic parsing of the text 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 WSD 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. This process is illustrate in Figure 1.

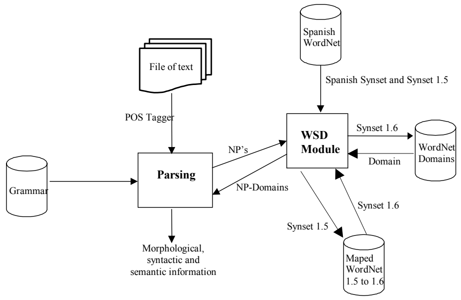

Fig. 1. *Process and resources used by WSD module*

The Figure 1 shows that the WSD module used the following resources:

- Spanish WN is a generic database with 30,000 senses. The Spanish WN will be linked through the English WN 1.5, so each English synonym will be associated with its equivalent in Spanish.
- WN 1.5 mapped to WN 1.6 is a complete mapping of the nominal, verbal, adjetival and adverbial parts of WN 1.5 onto WN 1.6 [3]

- WordNet Domain [6] is an extension of WN 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 SMM, 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 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. In the work [10] it has been shown 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 then into clusters. Detailed explanation of the SMM can be found in [9].

Next, we describe the way to obtain the domain label of WordNet Domain from word sense obtained by SMM. That is, SMM initially obtains the Spanish word sense and from this information has to apply the three following steps.

1. Starting from the Spanish word sense disambiguated by the SMM, we should obtain the corresponding synset in WN 1.5. For this task, we use the Spanish WN 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 → *plant*" is the Spanish Synset "08517914" (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 "00008894" (Plant#2).
2. WN 1.5 is mapped with the WN 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 "00008894" belonging to the sense "plant#2" is mapped to the synset 1.6 "00008864".
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 WN Domain giving the label "botany".

4 Coreference Resolution of Definite Description

Coreference resolution for DD presents different characteristics as pronouns. Three main differences can be pointed out: accessibility space, previous identification of non-anaphoric and different kinds of coreference (identity, part-of, set-member, set-subset). The accessibility space for pronouns is only a limited number of sentences. However, the accessibility space for DD represents a much greater number when encompassing the full text. For this reason, the number of potential candidates can be high for larger texts. If the coreference algorithm compares the DD to 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 DD (remaining non-anaphoric will be classified by coreference algorithm) and to built the list of candidates for each DD. A DD looks for their antecedent among 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 and the reduction of comparisons. The coreference algorithm provides an antecedent of DD or it classifies the DD as non-anaphoric, if no candidate is found. The coreference algorithm is a system based on weighted heuristics. These heuristics study the relation between heads and modifiers of both NP (candidate and DD). Moreover, DD can establish different kinds of relations to their antecedent. DD can refer to the full antecedent (identity coreference) or a part of the antecedent (part-of, set-member, set-subset). Our algorithm resolve the identity and part-of coreference. The following section shows the algorithm in detail.

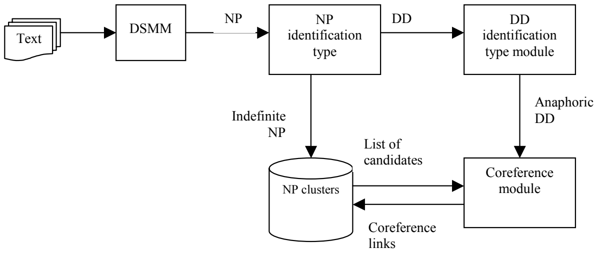

Fig. 2. *Full system*

4.1 Algorithm

The algorithm is focused on solving two tasks: non-anaphoric identification and coreference resolution. The algorithm takes advantage of DSMM (domain specification mark method) to solve both tasks. Two different modules are distinguished in the algorithm. The first module, Identification module, establishes the type of DD (anaphoric or non-anaphoric DD). A process of clustering is developed using the domain label proposed by DSMM. This module uses the Frege's idea of 'a word can only refer to a semantically compatible word'. Because of, this cluster is used in order to classify a DD between anaphoric and non-anaphoric. The second module, Coreference resolution module, is only applied to anaphoric DD. This module is based on a weight-heuristic system to choose the antecedent or to re-classify the DD as non-anaphoric if no antecedent is found.

Identification module The main goal of this module is to classify DDs between an anaphoric and non-anaphoric DD. For this reason, a previous task of identification of the NP type is made. The NP identification type is made by studying the first premodifier of NP. If the first modifier is a definite article or a demonstrative then the NP is classified as a DD. Otherwise, the NP is classified as an indefinite NP.

Every NP (DD and indefinite NP) is stored next to previous NPs with the same domain label. In addition, a virtual cluster is linked (label as v_link) to the NP (indefinite and non-anaphoric) made up of synonym, hyperonym, hyponym, meronym and holonym. All the words belonging to the virtual cluster do not previously appear in the text.

Moreover, the following process is only applied for DDs. If the DD is the first NP related to a domain label then the DD is classified as non-anaphoric. Otherwise, the coreference resolution mechanism is applied.

Coreference resolution module The goal of this module is to identify the antecedent of a DD or re-classify the DD as non-anaphoric if no antecedent is found. The algorithm needs as input the DD and a list of candidates. The list of candidates used for this coreference resolution module is made up all NPs with the same domain labels excluding words from virtual cluster. This virtual cluster is only used as a repository of words that are semantically related to the head noun of NP. The following steps are carried out: 1) The algorithm selects from the list of candidates those that have the same head noun as the anaphoric expression (DD). 2) If no candidate is selected then it goes through the virtual clusters that are related to the NP with the same domain label. The algorithm looks for the head noun of the anaphoric expression (DD). If it is found then the NP with the same domain label is selected as a candidate. 3) A weighting-heuristic algorithm is applied to choose the antecedent from the list of candidates or if the candidate list is empty then the DD is classified as non-anaphoric. The following heuristics are used:

- Identity coreference. The algorithm looks for previous noun phrases with the same head noun or a previous NP whose head noun is related using a synonym, hyperonym or hyponym relation and no incompatible modifiers. If one is found then both are linked using a identity coreference link (ic_link). Otherwise, the resolution process treats the anaphoric expression as a part-of coreference.

Hi1.- Same head. If a candidate has the same head noun as the DD then a value of 50 is added to the salience value (the red car, the car).

Hi2.- Synonym head. If the head noun of a candidate is a synonym of the head noun of the DD then a value of 45 is added to the salience value (the red car, the auto).

Hi3.- Hyper/hyponym head. If the head noun of a candidate is a hyperonym or hyponym of the head noun of the DD then a value of 35 is added to the salience value (the red car, the taxi).

Hi4.- Same modifier. A value of 10 is added to the salience value for each modifier that appears in both NP (candidate and DD) (the red car, the red auto).

Hi5.- Synonym modifier. A value of 9 is added to the salience value for each synonym modifier (the slow car, the lazy car)

Hi6.- Hyper/hyponym modifier. A value of 8 is added to the salience value for each hyper/hyponym modifier (the wood furniture, the mahogany furniture)

Hi7.- Antonym modifier. A value of -1000 is added to the salience value for each antonym modifier (the left ear, the right ear)

- Part-of coreference. Looking for a previous NP whose head noun is related using a meronym or holonym. If one is founded both are linked using a part-of coreference link (pc_link). The algorithm looks for the head noun at the virtual clusters linked by the same label.

  Hp1.- Holo/meronym head. If the head noun of a candidate is a holo/meronym of the DD head noun then a value of 25 is added to the salience value (car, engine).

  Hp2.- Head as modifier. If the head noun of DD is a modifier of candidate then a value of 10 is added to the salience value (the car, the car engine).

  Hp3.- Synonym as modifier. If the head noun of DD is a synonym of a modifier of a candidate then a value of 9 is added to the salience value (the car, the auto engine).

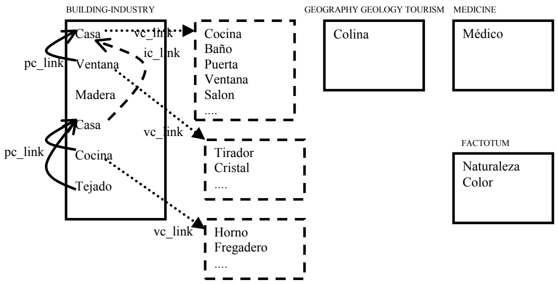

Fig. 3. *NP clustering using WN Domain tag*

If no candidate is selected as antecedent in identity coreference and part-of coreference then the DD is re-classified as non-anaphoric. And, if more than one candidate is proposed then the closest criteria is applied. Figure 3 shows the NP grouping after processing the following sentences: La casa de la colina era de un médico. Las ventanas eran de madera maciza. La casa estaba en plena naturaleza. La cocina era muy amplia y el tejado era de color rojizo.

5 Experimental work and results

Corpus	**Total**	**n-anaph DD**	**anaph DD**	
			IC	PC
Training	560	340	164	56
Test	742	451	217	74
Total	1302	791	381	130

Table 1. DD distribution

The experimentation data was taken from different HTML pages. In table 1 a distribution of DD in the corpora is shown. We distinguish anaphoric from non-anaphoric DD (n-anaph DD). Moreover, anaphoric DDs (anaph DD) are also divided into identity coreference (IC) and part-of coreference (PC). The test corpus was used to evaluate the identification of non-anaphoric DD (previous and full) and the coreference resolution (identity and part-of). Moreover, two experiments have been carried out. Obviously, the goal of the experimentation process is to evaluate the DD treatment. But, experiments were carried out to establish the influence of WSD module. The first experiment 1 evaluates the full algorithm carrying on errors produced by WSD module. And, the second experiment evaluates the algorithm supervising the errors from WSD module.

Exp.	**Previous**			**Full**		
	C	E	S%	C	E	S%
exp. 1	130	0	100	405	46	89.8
exp. 2	141	0	100	421	30	93.3

Table 2. Identification of non-anaphoric values using test corpus

5.1 Experiments for non-anaphoric identification

Table 2 shows the values obtained in each experiment for the identification of non-anaphoric DD. In the first experiment, 130 non-anaphoric DD were correctly classified (C) obtaining a success rate (S) of a 100%. This is due to the fact that the algorithm can only classify as non-anaphoric those DDs that cannot be compared with anyother because they have the first word as their domain label. The 321 remaining non-anaphoric DD were treated by the coreference algorithm. If this coreference algorithm did not find an antecedent then the DD was re-classified as non-anaphoric. The full task of non-anaphoric identification (adding previous identification and coreference identification) obtained a success rate around a 90%. In the second experiment, the algorithm obtained a small improvement in both stages (previous and full). For previous identification, 141 non-anaphoric DD were identified. And, the 310 remaining were treated by coreference algorithm. The full process achieved a success rate around 93%.

5.2 Experiments for coreference resolution

The evaluation of coreference algorithm involves the evaluation of two different kind of coreference: identity and part-of. Others kinds of coreference such as set-member or set-subset are not solved by treating them as non-anaphoric DD. Moreover, identity coreference can be divided into two types: direct anaphora and bridging references[2]. According to this definition, part-of coreference is also a type of bridging reference. Table 3 shows the values obtained in each experiment for the coreference resolution. In the first experiment, the algorithm achieved a success rate of 76% for identity coreference and a success rate of 58.1% for part-of coreference. In the second experiment, both coreferences (identity and part-of) increased their values. Identity coreference achieved a success rate of 80.1% and part-of coreference achieved a success rate of 62.1%. The values achieved for identity coreference can be divided into two different types: direct anaphora and bridging reference. The algorithm achieved a 83% of success rate for direct anaphora and a 64% of success rate for identity bridging reference[3]

Exp.	**Identity coref.**			**Part-of coref.**		
	C	E	S%	C	E	S%
exp. 1	165	52	76	43	31	58.1
exp. 2	174	43	80.1	46	28	62.2

Table 3. Coreference values using test corpus

5.3 Comparative results

The comparison of different approaches should be carried out using the same features. The main problem we found in this work was carrying out the compar-

[2] DD with different head noun as their antecedent were called bridging references by Clark [2]

[3] We use this term to refer to DD with different head noun as their antecedent and establishing an identity coreference.

ison between two different languages (Spanish and English), the use of specific tools (partial or full parser, ontologies, lexical resources, etc). For this reason, we decided to carry out an indirect comparison with approaches extensively cited in the literature and a direct comparison with a baseline algorithm.

A baseline algorithm was developed for this experiment. A simple algorithm for DD resolution is taken as a baseline algorithm. This algorithm looks for each DD as the candidates, with the same head noun as the anaphoric expression (DD) choosing the closest. If no candidate is selected then the DD is classified as non-anaphoric. The values achieved for baseline algorithm are the same in experiments 1 and 2 because this algorithm does not use semantic information. The success rate calculated for non-anaphoric identification was around 63% for baseline algorithm and around a 90% for our algorithm without supervising the errors produced by DSMM (exp. 1) and a 93% supervising the DSMM' errors (exp. 2). The comparison made for coreference resolution shows the values achieved in two type of coreference: identity (IC) and part-of (PC) for both algorithm. The success rate calculated for identity coreference was around 56% for baseline algorithm and around a 76% for our algorithm without supervising the errors produced by DSMM (exp. 1) and a 80% supervising the DSMM's errors (exp. 2). Moreover, identity coreference can be divided into two types: direct anaphora and identity bridging reference. The identity bridging reference resolution needs to use semantic information, for this reason the value achieved by baseline algorithm is null. The direct anaphora resolution is solved by both algorithm (baseline and our algorithm) achieving a success rate of 70% for baseline and 83% for our algorithm. The success rate calculated for part-of coreference was 0% for baseline algorithm because it does not use semantic information and around 58% for our algorithm without supervising the errors produced by DSMM (exp. 1) and a 62% supervising the DSMM's errors (exp. 2).

We selected for indirect comparative evaluation two approaches extensively cited in the literature. For non-anaphoric identification, we used Vieira & Poesio 'algorithm [13] and Bean & Rillof [1]. And, for coreference resolution, we used Vieira & Poesio 'algorithm [13] and Kameyama [5]. For non-anaphoric identification, our algorithm achieved better score (93%) than Bean & Rillof algorithm (86%) and Vieira & Poesio (72%). For coreference resolution, our algorithm achieved similar values for direct anaphora as Vieira & Poesio, around a 83% and for bridging reference our algorithm (65%) is better than Poesio & Vieira (28%). The bridging reference values of our algorithm included identity bridging reference and part-of coreference due to Vieira & Poesio work does not separately show these values. Moreover, Kameyama work show an overall value for coreference resolution task, 59%.

6 Conclusions

We have introduced a DD algorithm based on semantic information to identify non-anaphoric DD and to solve anaphoric DD. In addition to typical semantic information (synonym, hyperonym, etc.), domain labels are used to cluster NPs.

This clustering helps us to establish a mechanism for previous non-anaphoric identification and to reduce the number of candidates. Experimental work shows that the use of WSD improves the values of DD resolution tasks. Our algorithm resolves two different types of coreference, identity and part-of, achieving better values than others work developed for English.

References

1. D. L. Bean and E. Riloff. Corpus-based Identification of Non-Anaphoric Noun Phrases. In *Proceedings of the 37th ACL*, pages 373–380, 1999.
2. H. H. Clark. Bridging. In P. Johnson-Laird and P Wason, editors, *Thinking: readings in cognitive science*, pages 411–420. Cambridge UP, 1977.
3. 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.
4. G. Frege. Sobre sentido y referencia. In Luis Ml. Valdés Villanueva, editor, *La búsqueda del significado: Lecturas de filosofía del lenguaje*. 1892.
5. M. Kameyama. Recognizing Referential Links: An Information Extraction Perspective. In Mitkov, R. and Boguraev, B., editor, *Proceedings of ACL/EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, 1997.
6. B. Magnini and G. Cavaglia. Integrating subject field codes into WordNet. In *Proceedings of the LREC-2000*, 2000.
7. 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.
8. A. Montoyo and M. Palomar. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. In *Proceedings of the DEXA-2000, 11th International Workshop on Database and Expert Systems Applications*, pages 103–107. IEEE Computer Society, September 2000.
9. A. Montoyo and M. Palomar. Specification Marks for Word Sense Disambiguation: New Development. In Gelbukh, editor, *Proceedings of the CICLing-2001*, LNCS, pages 182–191, 2001.
10. 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.
11. 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.
12. M. Poesio and R. Vieira. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, 24:183–216, 1998.
13. R. Vieira and M. Poesio. An Empiricall Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4):539–593, 2000.
14. Y. Wilks and M. Stevenson. Word sense disambiguation using optimised combination of knowledge sources. In *Proceedings of COLING-ACL'98*, 1998.