Semantic Approach to Bridging Reference Resolution

R. Muñoz, M. Saiz-Noeda, A. Suárez and M. Palomar
Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas de Informáticos. Universidad de Alicante
Apartado 99. 03080 Alicante, Spain
{rafael,max,armando,mpalomar}@dlsi.ua.es

Abstract This paper presents an algorithm to solve Spanish bridging references. It uses Spanish WordNet for applying semantic criteria and some heuristic rules to choose the correct antecedent. The Spanish WordNet semantic resource provides information about synonymy, hyperonymy/hyponymy, thematic role (role agent) and antonymy relations. The algorithm has been developed in Prolog, but we have also developed some C++ libraries in order to obtain all semantic information from Spanish WordNet. This algorithm achieves an average precision of 60.9% and a recall of 78% in bridging co-reference resolution.

1 Introduction

We commonly use different expressions to refer to a person, an object, an event, a place or a process. These expressions usually used in this way are pronouns and definite descriptions. Among them, definite descriptions are the most difficult to treat because they do not always refer to an antecedent but they can introduce a new entity in the discourse. Moreover, a definite description can establish different kinds of relations between antecedent and anaphoric expression (definite description). These relations can be *identity*, *parts of*, *set-subset* and *set-member*. To these relations we can add another between the head nouns of both antecedent and anaphoric expression. So, we distinguish between definite descriptions with the same or different head noun as their antecedent, the latter being called *bridging reference* in Clark (1977). This paper is focused on co-reference resolution produced by these bridging references.

2 Relevant related work on definite descriptions

Most automated approaches to definite description resolution are focused on English texts. Systems for noun phrase co-reference resolution can be characterized as learning and knowledge-based approaches.

2.1 Learning approaches

All previous attempts to view co-reference as a learning problem treat its resolution as a classification task. The algorithms classify a pair of noun phrases as co-referent or not. Learning approaches are divided into two categories: supervised and unsupervised algorithms.

On the one hand, the main problem of supervised approaches is the need of a large amount of training data annotated with co-reference resolution information. Two of the most representative algorithms are MLR (Aone & Bennet, 1995) and RESOLVE (McCarthy & Lehnert, 1995). Both algorithms apply the C4.5 decision tree induction algorithm (Quinlan, 1992) and a mechanism to coordinate the collection of pairwise co-reference decisions. This mechanism forces the algorithm to satisfy the transitive property (if NP_i co-refers to NP_j and NP_j co-refers to NP_k then NP_i must co-refer to NP_j). On the other hand, unsupervised approaches do not require training data. Some of the most representative work in this field has been developed in Cardie & Wagstaff (1999).

2.2 Knowledge-based approaches

Most knowledge-based approaches use a series of linguistic constraints and discourse information. Below, we explain some of them.

2.2.1 Vieira & Poesio algorithm

An algorithm focused on the resolution of definite descriptions is described in Viera and Poesio (1998) and Vieira (1999). This system solves references between definite descriptions and antecedents with either the same head noun or a semantic relationship (bridging references). It is based on several tests in order to classify the definite description as non-anaphoric expression or to provide the correct antecedent. The algorithm executes four tests for identifying new discourse descriptions before trying to find an antecedent. If these tests fail, the system will look for an antecedent with same head as the anaphoric expression (direct anaphora). Finally, the system applies several heuristic rules in order to look for semantic relations (*synonymy, hyponymy* and *meronymy*) between both head nouns (indirect anaphora). This algorithm achieved 62% recall and 83% precision in solving direct anaphora (same head noun). Bridging descriptions (indirect anaphora) were evaluated by hand, and 61 relations of 204 in the corpus were achieved.

2.2.2 LaSIE-II System

The CO mechanism used by LaSIE-II System (1998) takes a set of entities newly added to the discourse model and compares each one with the set of instances already in the discourse model. Proper names, pronouns and common nouns are handled separately for object co-reference, first attempting intra-sentential co-reference for each set, and then inter-sentential co-reference. Each pair of new-old instances, if compatible, has a similarity score calculated for it, based on the distance between the instances' parent classes in the concept hierarchy, and the number of shared properties. The highest scoring pair, for each new instance, is merged in the discourse model, deleting the instance with the least specific class in the ontology, and combining the properties of both instances. This mechanism achieved a recall of 56.1% and a precision of 68.8%.

2.2.3 Muoz & Palomar algorithm

The algorithms presented above are focused on solving references in English texts. These algorithms solve different types of definite descriptions from several taxonomies such as those in Cristopherson (1939), Hawkins (1978), Prince (1981) and Poesio & Vieira (1998), carried out in the English language. The algorithm developed in Muñoz & Palomar (2000) is focused exclusively on definite description with the same head as their antecedent in Spanish texts. In order to solve this kind of definite description, syntactic and semantic information is needed. This algorithm processes the texts sentence by sentence adding every noun phrase into a list of antecedent (LA). The steps of the algorithm are the following:

- *The detection of the definite description in the text.*

- *Search for candidate antecedents with the same head noun.*the system extracts from the list of antecedents whose with the same head noun as definite description making up the list of candidates (LC).

- *Choice of the correct antecedent.* If there is more than one antecedent in the list of candidates (LC), then heuristic rules are applied. Otherwise, if there is only one antecedent, then it is considered to be the solution of the definite description and added to the LA.

This algorithm has been checked on two different corpora. The first corpus is a fragment of LEXESP corpus different from the one used as of the training corpus. And, the second one is formed by several deeds. In order to check the last corpus the algorithm has been introduced into the information extraction system EXIT (Llopis *et al.*, 1998). The results achieved by the algorithm were: In LEXESP corpus a precision of 85.7% (210/245) in the co-reference resolution task for definite descriptions with the same head noun as their antecedent. In the deed corpus, a precision of 81.5% (159/195) was achieved.

3 Scope of the problem

3.1 Bridging references

Bridging reference is defined in Clark (1977) as the referential relation between two noun phrases with different head nouns. Semantic and pragmatic information is needed to solve this relation between antecedent and definite description. The lack of pragmatic information makes some kinds of definite descriptions impossible to solve. Based on this lack of pragmatic information and following the classification of Spanish definite descriptions in fig. 1 and Clark's definition, we use the term *bridging references* to refer to those definite descriptions *semantically related to the antecedent* and *semantically related to the verb*. In order to solve these definite descriptions a lexical resource has been used. This resource is the Spanish WordNet.

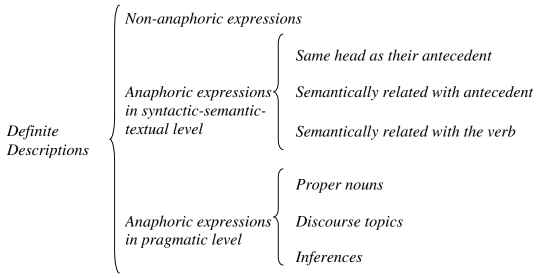

Figure 1: *Definite description's classification (Muñoz* et al.*, 2000)*

3.2 Spanish Wordnet

Spanish WordNet belongs to the EuroWordNet project[1]. As explained in Vossen (1998) EuroWordNet is a multilingual lexical database with wordnets for several European languages, and these are structured along the same lines as the Princeton WordNet (Fellbaun, 1998). WordNet contains information about nouns, verbs,

[1] EuroWordNet (LE2-4003 and LE-8328) is funded by the European Community within the Telematics Application Programme of the 4th Framework (DG-XIII, Luxembourg). The project started March 1996 and ended July 1999.

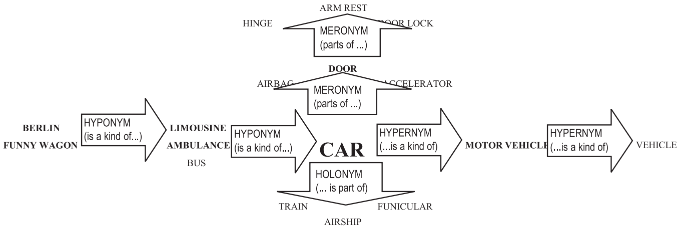

Figure 2: *WordNet's semantic relations for noun "car"*

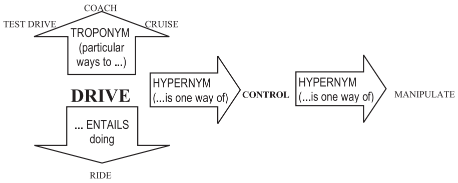

Figure 3: *WordNet's semantic relations of verb "drive"*

adjectives and adverbs in English, and is organized around the notion of a synset. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, *[car; auto; automobile; machine; motorcar]* form a synset because they can be used to refer to the same concept. A synset is often further described by a gloss: "4-wheeled; usually propelled by an internal combustion engine". Finally, synsets can be related to each other by semantic relations, such as hypernymy/hyponymy (between specific and more general concepts), meronymy/holonymy (between parts and wholes), troponymy/entailment, etc. These relations are illustrated in fig. 2 and fig. 3 for nouns and verbs respectively.

As well as these, EuroWordNet provides some additional relations such as the ROLE-INVOLVED ones. These relations will be very useful for the algorithm presented here. ROLE relation provides the synsets that are related with a verb and have a role with reference to it (agent, instrument, patient, location, direction, ...). Moreover, INVOLVED relation provides the synsets related to a noun according to the role it carries out (it is the complementary relation of the ROLE one). This way, the verb *cazar* (*hunt*) has the words *cazador*, *trampero* and *cepero* (*hunter, trapper*) as ROLE-AGENT synsets, and, consequently, *cazador* has *cazar* as INVOLVED-AGENT. This relation can be seen in fig. 4

Using these relations, the algorithm tries to relate two noun phrases without structural relations and with semantic relations, like the ones mentioned above. The following section presents the algorithm step by step

4 The algorithm

This algorithm deals with the Spanish bridging references. It is part of a resolution module of any kind of definite description references. The algorithm searches for a definite description in the text and stores all previous noun phrases in a list of candidates. We consider only the previous noun phrases in the text as possible antecedents, rejecting all others such as full sentences or paragraphs.

The text used as input to the system is made up of words and part-of-speech tags. The system analyzes the text adding syntactic tags (noun phrases, prepositional phrases and verbal phrases). Definite descriptions are identified when a

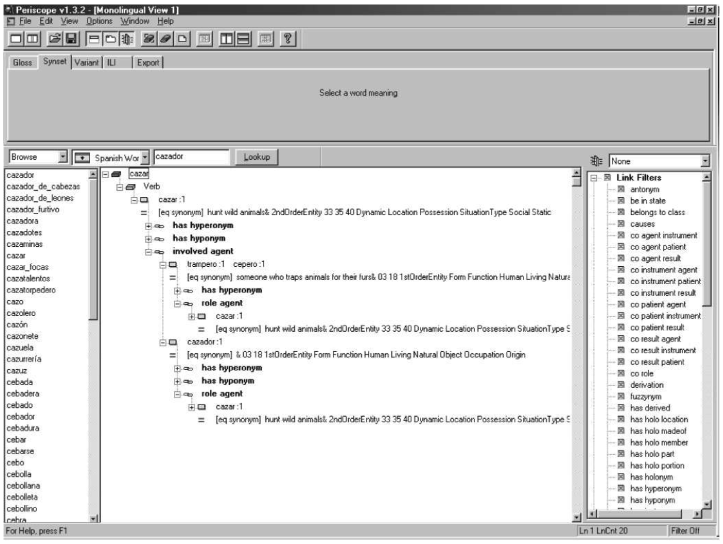

Figure 4: *ROLE/INVOLVED relations of "cazar" in Periscope tool*

noun phrase whose first constituent is a definite article or a demonstrative. Once the definite description is detected, the algorithm must identify the type of definite description (same head or bridging reference). If the definite description is classified as a bridging reference then its head noun is extracted and used for obtaining its semantic information from Spanish WordNet (synonym, hyperonym, role agent, etc.).

Below, each step of the algorithm will be explained in detail:

Step 1: Identifying the type of definite description. When the definite description is detected, its type (same head or bridging reference) is unknown. First of all, the system treats it as a same head definite description applying the algorithm developed in Muñoz & Palomar (2000). If no antecedent is proposed by the algorithm then the algorithm to solve bridging references is applied. This order is as a result of the study developed in (Muñoz *et al.*, 2000) that shows that an 81% of definite descriptions are same head definite description type.

Step 2: Obtaining semantic information. Once definite description is classified as bridging reference, semantic information is extracted from Spanish WordNet. The head noun of the bridging reference is extracted in order to obtain its semantic relations (synonym, hyperonym, role agent, involve agent, etc.) from the lexical resource. This information is stored in several semantic relation lists.

Step 3: Choosing the candidates. The algorithm goes through the list of antecedents comparing each head noun with the elements in the semantic relation lists. If an antecedent is not in any of the semantic relation lists then the candidate is rejected. Otherwise, the antecedent is stored in the list of candidates.

Step 4: Choosing the correct antecedent. If the list of candidates is empty, the definite description is classified as non-anaphoric. If the list has only one antecedent, it is chosen as the correct antecedent. Otherwise, a set of rules is applied in order to choose the correct antecedent. These rules are applied in the following order :

- **Synonym.** The algorithm prefers a candidate with a synonymy relation with the head noun of the bridging definite description.
- **Hyperonym.** The algorithm prefers a candidate with a hyperonymy relation with the head noun of the bridging definite description.

- **Thematic role.** The algorithm prefers the antecedent with an EuroWordNet's ROLE_AGENT relation with the head noun of the bridging definite description. ROLE_AGENT relation provides the verb(s) related to a noun with a subject role (cazador > cazar, *hunter* > *hunt*). The algorithm establishes co-reference between a definite description and a candidate that is the subject of a ROLE_AGENT verb of the definite description (any subject of the verb cazar –*hunt*– is considered co-referent of cazador –*hunter*–). Moreover, the algorithm combines ROLE_AGENT and ANTONYMY relations establishing that a definite description co-refers to the candidate that is the indirect object of the antonymous of the ROLE_AGENT verb of the definite description (any indirect object of the verb comprar –*buy*– is considered co-referent of vendedor –*seller*–) .

After applying these rules if the list of candidates contains more than one element, the algorithm chooses the closest.

5 Evaluation

Co-reference is a transitive relation between two or more references. If NP_i corefers to NP_j, and NP_j co-refers to NP_k then NP_i co-refers to NP_k. We take advantage of this transitive property for the definition of co-reference chains (NP_i, NP_j and NP_k make up a co-reference chain). In the resolution process, if the system proposes a noun phrase different from the correct one, but it is in the same co-reference chain, then the solution is considered correct.

This algorithm works with unrestricted texts. It has been evaluated on a fragment of LEXESP corpus[2] formed by narrative texts from several authors (8,629 words, 544 definite descriptions). Due to the lack of a word sense desambiguation tool, two different tests have been performed. First of all, the algorithm has been applied using only the first sense (the most frequent one) for each word extracted from Spanish WordNet. Secondly, we have used all senses associated to each word. This second experiment provides a wider quantity of associated terms that causes less precision in the resolution process. So, we have chosen the first approach, which gave better results and processing speed. The algorithm achieves a recall of 78% and an average precision of 60.9% divided as follows: a precision of 68% produced by the semantic relation type (synonymy, hyperonymy/hyponymy) and a precision of 56% produced by the thematic role (role agent) type. It is difficult to compare these results with other approaches because of the use of different languages and corpora. Vieira and Poesio's work is only focused on definite descriptions reaching a precision of 37% in associative anaphora (bridging references).

6 Conclusion

The main contribution of this work is the development of an algorithm to solve Spanish bridging references. We take advantage of the ROLE_AGENT semantic relation from Spanish WordNet (EuroWordNet) in order to solve thematic role definite descriptions. Moreover, the algorithm uses other relations like synonymy and hyperonymy/hyponymy to solve semantic definite descriptions. This algorithm achieves an average precision of 60.9% in bridging references and a recall of 78%.

[2]LEXESP is a Spanish corpus. This corpus is about 5 million of tagged words developed by Psychology Department from the University of Oviedo, Computational Linguistic Group from the University of Barcelona and Language Treatment Group from the Technical University of Catalonia.

The application of new semantic relations and the use of word sens desambiguation tools in order to enrich the search of the correct antecedent and improve the algorithm make up future research lines.

References

Aone, C., & Bennet, S.W. 1995. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. *Pages 122–129 of: Proceedings of the 33rd Annual Meeting of the ACL.*

Cardie, C., & Wagstaff, K. 1999. Noun Phrase coreference as Clustering. *Pages 82–89 of: Proceeding of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Christopherson, P. 1939. *The Articles: A study of their theory and use in English.* Copenhagen: E. Munksgaard.

Clark, H. H. 1977. Bridging. *Pages 411–420 of:* Johnson-Laird, P., & Wason, P (eds), *Thinking: readings in cognitive science.* London New York: Cambridge University Press.

Fellbaun, C. 1998. *WordNet, an electronic lexical database.* MIT Press.

Hawkins, J. A. 1978. *Definiteness and indefiniteness.* Atlantic Highlands, New Jersey: Humanities Press.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., & Mitchell, B. 1998. University of Sheffield: Description of the LaSIE-II System as used for MUC-7. *In:* Morgan Kaufman Publishers (ed), *Proceedings of Seventh Message Understandig Conference.*

Llopis, F., Muñoz, R., Suárez, A., & Montoyo, A. 1998. EXIT: Propuesta de un sistema de extracción de información de textos notariales. *Novática*, **133**, 26–30.

McCarthy, J. F., & Lehnert, W. G. 1995. Using Decision Trees for Coreference Resolution. *Pages 1050–1055 of: Proceedings of the Fourteenth International Conference on Artificial Intelligence.*

Muñoz, R., & Palomar, M. 2000. Processing of Spanish Definite Descriptions with the Same Head. *Pages 212–220 of:* Christodoulakis, Dimitris N. (ed), *Proceeding of NLP2000: Filling the gap between theory and practice.* Lectures Notes in Artificial Intelligence, vol. 1835. Patras, Greece: Springer-Verlag.

Muñoz, R., Palomar, M., & Ferrández, A. 2000. Processing of Spanish Definite Descriptions. *Pages 526–537 of:* O. Cairo and E. L. Sucar and F. J. Cantu (ed), *Proceeding of Mexican International Conference on Artificial Intelligence.* Lectures Notes in Artificial Intelligence, vol. 1793. Acapulco, Mexico: Springer-Verlag.

Poesio, M., & Vieira, R. 1998. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, **24**, 183–216.

Prince, E. 1981. Toward a taxonomy of given-newinformation. *Pages 223–256 of:* Cole, P. (ed), *Radical Pragmatics.* New York: Academic Press.

Quinlan, J. R. 1992. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

Vieira, R. 1999 (January). Co-reference resolution of definite descriptions. *Pages 497–503 of: Proceedings of VI Simposio Internacional de comunicación Social.*

Vieira, R., & Poesio, M. 1998. Processing definite descriptions in corpora. *In:* Botley, S., & McEnery, T. (eds), *Corpus-based and computational aproach to anaphora*. London: UCL Press.

Vossen, P. 1998. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter*, **3**(1).