NLP system oriented to anaphora resolution

Maximiliano Saiz-Noeda[†], Manuel Palomar[†] and David Farwell[‡]
[†]Dpto. Lenguajes y Sistemas Informáticos, Universidad de Alicante, Spain
[‡]Computing Research Laboratory, New Mexico State University, USA
{max,mpalomar}@dlsi.ua.es — david@crl.nmsu.edu

Abstract In this paper, we present a Natural Language Processing (NLP) system in Spanish. This system is oriented to anaphora resolution. On the one hand, we will present the linguistic sources and resources to make the resolution process easier. On the other hand, a constraint-based method for anaphora resolution is presented. We emphasise the use of semantic knowledge as a mechanism that helps the resolution of the anaphora.

Keywords: Anaphora resolution, semantics, LFG grammar and parsing

1 Introduction

It is necessary to have suitable information sources in order to propose a suitable mechanism for anaphora resolution. In the course of the last years, numerous researches have concentrated their efforts on solving the problem of lexical and lexical-morphological parsers. Also, they have been worked on the obtaining of universal information sources to provide the information adapted to each problem. With reference to the addition of semantics in the resolution of linguistic phenomena, it is not easy to find references and resources that provide good results. In this line, EuroWordNet project (see (Vossen, 1998)) has tried to make an interesting approach to help this kind of research, although the generality features of this resource continue making it difficult to use in NLP tasks.

To solve this lack of semantic resources to be applied in concrete NLP tasks, we propose in this paper the addition of specific semantic information to the anaphora resolution process. This semantic information is based on the use of the correct sense of the words and the semantic relation among them. In order to obtain the correct sense and the semantic relations for each word, we propose the use of patterns (subject-verb, verb-object) obtained by means of learning from a corpus. These patterns are formed by ontological concepts associated to nouns and verbs and establish the semantic behavior of the words in the text.

In short, we propose a NLP system that counts on as much information sources as possible for the anaphora resolution. We think that the base of the language processing is the information used for it. A modular, independent and easily expandable system with any additional information source is proposed.

Next section shows the scenario for anaphora resolution systems and algorithms, underlying the most interesting approaches in this research line. Following, the paper gives a detailed description of the NLP system and all the processes and resources related to it. Next section explains the constraint-based mechanism for the anaphora resolution and the information sources it uses. Finally, some conclusions of this work and the work in progress are presented.

2 State-of-the-art

A common point among all languages is the fact that the anaphora phenomenon involves similar strategies for its resolution (e.g. pronouns or definite descriptions). For example, they all employ different kinds of knowledge. The strategies differ only in the way in which they coordinate all the different kinds of knowledge. For example, some use just one kind of knowledge as the main selector for identifying the antecedent and the others are used merely to confirm or reject the proposed antecedent. The typical kind of knowledge used as the selector is that of the discourse structure, e.g. those that are based on the Centering Theory, such as the methods employed by Strube and Hahn (1999) or by Okumura and Tamura (1996). Other works, however, give equal importance to each kind of knowledge and generally distinguish between restrictions and preferences (Baldwin (1997), Lappin and Leass (1994), or Carbonell and Brown (1988)). Restrictions tend to be absolute and, therefore, discard any possible antecedents, whereas preferences tend to be relative and require the use of additional criteria, i.e. heuristics that are not always satisfied by all anaphors. One example of this different sort of resolution model can be found in Nakaiwa and Shirai (1996), that uses semantic and pragmatic restrictions, such as constraints which are based on modal expressions, or such as verbal semantic attributes or conjunctions.

Semantic and domain information is relatively expensive in computational processing when compared to other kinds of knowledge. Consequently, current anaphora resolution methods rely mainly on restriction and preference heuristics, which employ morpho-syntactic information or shallow semantic analysis (for example, work by Mitkov (1998)). Such approaches, nevertheless, perform notably well. An algorithm for pronominal anaphora resolution that achieves a high rate of correct analysis (85%) is described by Lappin and Leass (1994). This approach, however, operates almost exclusively on syntactic information. More recently, Kennedy and Boguraev (1996), propose an algorithm for anaphora resolution that is actually a modified and extended version of the one developed by Lappin and Leass (1994). It works from a POS tagger output and achieves an accuracy rate of 75%.

Some approaches are also based on POS tagger outputs, e.g. (Mitkov & Stys, 1997), where it is proposed another knowledge-poor approach to resolving pronouns in technical manuals in both English and Polish. The knowledge employed in these approaches is limited to a small noun-phrase grammar, a list of terms and a set of antecedent indicators, (definiteness, term preference, lexical reiteration, etc.).

So, we could say that current proposals on anaphora resolution for different languages are based on the type of information they have access to. Therefore, we considered that the anaphora resolution method is as important as the information sources used to apply it

3 NLP system: modules and resources

In this section we present a NLP system proposal and all the resources (Parser, POS Tagger and EuroWornet) and the information sources (lexical & morphological from the POS tagger, lexical & semantic information from lexicon and semantic patterns) the system uses. It is important to understand this system as a general purpose system, although in this paper we underline the anaphora resolution like one of the main problems in the NLP field. This system could be used in applications like machine translation, information extraction or text summarisation, applications that need an anaphora resolution module in order to

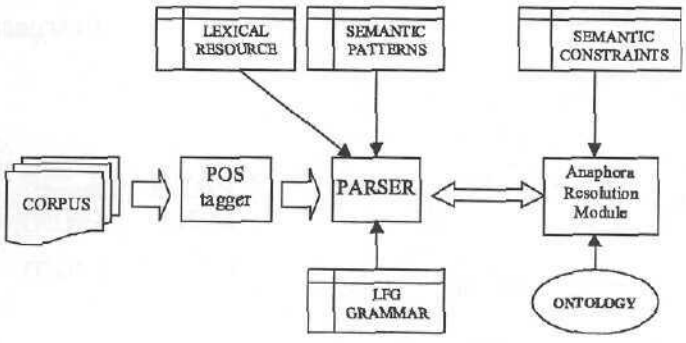

Figure 1: *Method schema*

obtain good results.

The main module of the system (see Figure 1) is the parser. It is based on the bottom-up parsing chart technique. The parser uses a LFG grammar (Lexical-Functional Grammar) that suitably defines the language features. The parser receives several information sources as input: the output of the tagger (that supplies lexical and morphological information for each word), a lexicon with lexical and semantic information (semantic features of the words) and a set of semantic patterns that make up for the possible lack of information in the lexicon. The output of the parser is the input for the anaphora resolution module, which receives for each detected anaphora a set of possible antecedents with their lexical, morphological, syntactic and semantic information that will allow solving the anaphora suitably.

Below, we describe briefly the different modules of the system.

3.1 Parser

The parser takes different input sources. On the one hand, it receives the output of a POS tagger that provides, for each word in the analysed corpus, its morphological features (gender, number, person, verbal tense,...). Also, the parser takes as input a lexicon with information about the syntactic structure of the word as well as the semantic concepts associated to it.

As said before, the parser uses the grammatical formalism LFG exposed in section 3.2 for the analysis process.

Since the lexicon is limited, if the parser does not find the word, it can take an alternative input from a set of ontological concept patterns subject-verb and verb-complement to determine if the analysis is correct. The pattern learning process will be detailed in section 3.3.

3.2 LFG Grammar

Dalrymple *et al* (1997) say that LFG assumes two syntactic levels of representation. Constituent structure (c-structure) encodes phrasal dominance and precedence relations, and is represented as a phrase structure tree. Functional structure (f-structure) encodes syntactic predicate-argument structure, and is represented as an attribute-value matrix. Both structures for the sentence "Juan come una manzana" (*"Juan eats an apple"*) are represented in Figure 2.

F-structure consists of a collection of attributes, such as PRED, SUBJ, OBJ or IOBJ, whose values can be other f-structures. In this paper, we present the structures suitable for pronouns and noun phrases, which are the required components for the anaphora resolution.

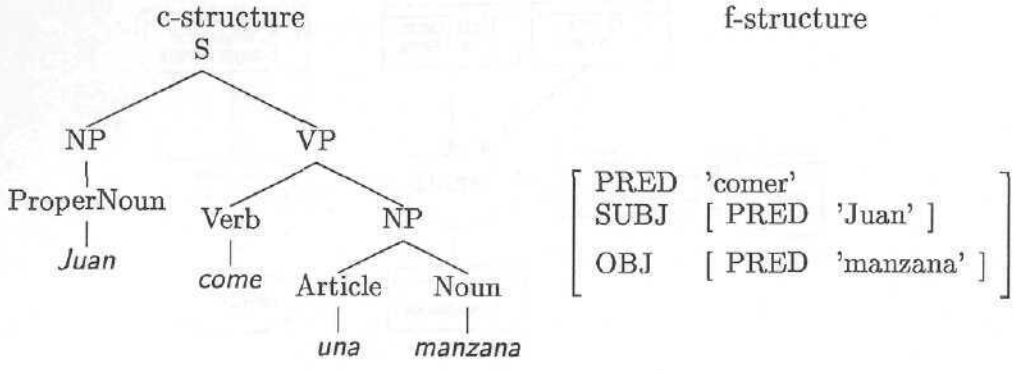

Figure 2: Constituent and Funcional structures for *"Juan come una manzana"*

A personal pronoun supplies information about its person, number and gender. We propose, like Butt *et al* (1999), that the pronouns will be analyzed as having a PRED value of 'PRO' indicating that these are anaphors awaiting resolution within the semantic component. In order to provide such a component with as much information as possible, the surface form of the pronoun is encoded in the PRO-FORM feature. Gender, number and person are also encoded because they are needed for the syntactic-semantic evaluation of the pronoun.

On the other hand, a noun phrase supplies information about its head, modifier and specifier. Moreover, the head of the noun phrase supplies information about its number and gender.

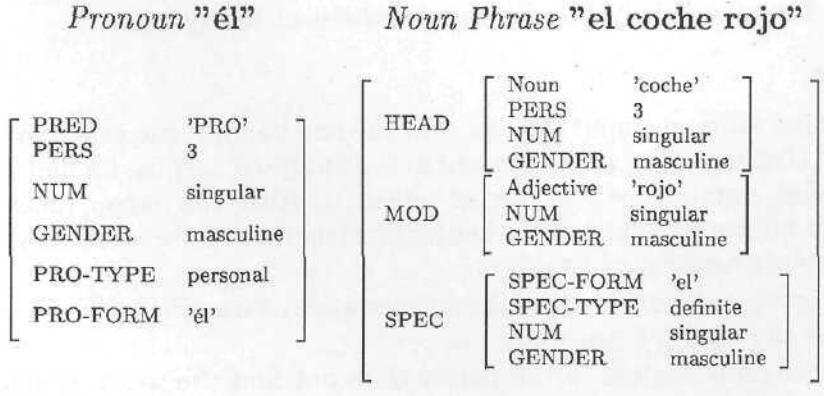

Figure 3: Information structure examples for pronoun and noun phrase

Figure 3 shows an example of structures for the pronoun 'él' (*'he'*) and the noun phrase "el coche rojo" (*the red car*).

3.3 Pattern Learning method

The pattern learning method extracts a set of *subject noun-verb* and *verb-complement noun* patterns. The semantic or ontological concepts of the noun phrase head (subject or complement) and predicate verbal phrase head in each sentence will form these patterns. These patterns define the semantic structure of these elements in the text. For the pattern construction, the Spanish version of the lexical resource WordNet, within the EuroWordNet proyect described by Vossen (1998) has been chosen. WordNet provides a main level of ontological concepts to describe all the words contained in the knowledge base. These concepts are 25

for nouns and 15 for verbs and they get the main semantic characteristic of each word sense. Table 1 shows these concepts.

Names	act, animal, artifact, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, time
Verbs	body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative, weather

Table 1: Conceptual classification for nouns and verbs in WordNet

These patterns define the semantic concept of compatibility between a noun (subject or complement) and a verb. With reference to the anaphora resolution process, this compatibility will allow choosing the correct antecedent of an anaphora among a group of noun phrases.

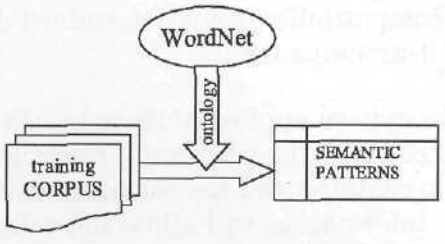

Figure 4: *Semantic pattern extraction schema*

Figure 4 shows the learning pattern extraction system. From the training corpus, the noun-verb and verb-noun pairs corresponding to the heads of the noun and verbal phrases with the subject-predicate and predicate-complement roles in a sentence are manually extracted. Both words are consulted in WordNet using their tagged sense . This way, pairs of noun-verb and verb-noun ontological concepts are formed. These new pairs define the semantic behaviour of the verbs and the nouns as subjects or complements (direct or indirect object) of those verbs.

With these patterns, the method adds the semantic information to the anaphora resolution module. For example, taking the verb (comer) (textitto eat) and two noun phrases la piedra (*the stone*) and el león (*the lion*), the patterns generated by both nouns and the verb are *object-consumption* and *animal-consumption*. Intuitively, we can deduce that the second pattern defines elements as semantically more compatible. So, if both noun phrases comprise an antecedent list of an anaphoric expression with the verb comer, it is possible to say that, from the semantic point of view, el león can be the correct antecedent because of its compatibility.

4 Anaphora resolution method

Our anaphora resolution method combines different kinds of knowledge. No knowledge based on the discourse structure is included. The reason for it is that, in order to obtain this kind of information, we would require not only semantic knowledge but also word knowledge and an almost perfect parsing (see Azzam *et al* (1998)).

From the output of the parser, that provides NP candidates to be the antecedent of the anaphoric pronoun, anaphora resolution algorithm applies morphologic, syntactic and semantic constraints in order to eliminate all the candidates that are not compatible with the pronoun.

- ***Morphological constraints:*** Morphological constraints establish gender, number and person parallelisms that demand the compatibility or agreement between the antecedent and the anaphoric pronoun.
- ***Syntactic constraints:*** Based on C-command and Minimal Governing Category restrictions as formulated by Reinhart (1983) and non-coreference conditions by Lappin and Leass (1994), we propose conditions for NP-pronoun non-coreference adapted for Spanish. A pronoun is non-coreferential with a noun phrase (NP) if any of the non-coreference conditions are fulfilled. These conditions relate reflexive, demonstrative and personal pronouns with their syntactic role and position in the sentence, as can be seen in Palomar *et al* (2000)
- ***Semantic constraints:*** From the semantic features associated to each antecedent NP through its head noun, semantic constraints eliminate those candidates that are not compatible with the verb in the anaphoric expression. Son, in this case, the compatibility is not determined directly by the pronoun, but by the verb that it accompanies.

Once these constraints have been applied, if there is only one candidate left, it is chosen as the correct antecedent of the pronoun. Otherwise, a set of semantic criteria based on the semantic patterns and the semantic structure above mentioned is applied. This semantic information will allow the selection of the antecedent that is conceptually most compatible.

5 Conclusions

In this paper we have presented a general-purpose NLP system oriented to the anaphora resolution. We consider that it is necessary to incorporate all the available information sources for a suitable resolution of the anaphora, and that the anaphora is a problem that depends, to a large extent, on the information sources and their correct use.

At this moment, our work is oriented to the research of new information sources and their application within the anaphora resolution in Spanish. Also, we are studying the structure of the discourse as a mechanism for helping this resolution. Finally we consider that applications like machine translation need suitable mechanisms for anaphora resolution, as can be seen in its resolution in the present available systems.

Referencias

Azzam, S., Humphrey, K., & Gaizauskas, R. 1998 (August). Evaluating a Focus-Based Approach to Anaphora Resolution. *Pages 74–78 of: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).*

Baldwin, B. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. *Pages 38–45 of: Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphor resolution.*

Butt, M., Holloway, T., Niño, M., & Segond, F. 1999. *A Grammar Writer's Cookbook.* Stanford University, USA: CLSI lecture notes; no. 95.

Carbonell, J.G., & Brown, R.D. 1988. Anaphora resolution: a multi-strategy approach. *Pages 96–101 of: Proceedings of 12th International Conference on Computational Linguistics (COLING'88).*

Dalrymple, M., Lamping, J., Pereira, F., & Saraswat, V. 1997. Quantifiers, Anaphora, and Intensionality. *Journal of Logic, Language, and Information,* 219–273.

Kennedy, C., & Boguraev, B. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. *Pages 113–118 of: Proceedings of 16th International Conference on Computational Linguistics,* vol. I.

Lappin, S., & Leass, H. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics,* **20**(4), 535–561.

Mitkov, R. 1998 (August). Robust pronoun resolution with limited knowledge. *Pages 869–875 of: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).*

Mitkov, R., & Stys, M. 1997 (September). Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish. *Pages 74–81 of: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP95.*

Nakaiwa, H., & Shirai, S. 1996. Anaphora Resolution of Japanese Zero Pronouns with Deictic Reference. *Pages 812–817 of: Proceedings of 16th International Conference on Computational Linguistics,* vol. I.

Okumura, M., & Tamura, K. 1996. Zero Pronoun Resolution in Japanese Discourse Based on Centering Theory. *Pages 871–876 of: Proceedings of 16th International Conference on Computational Linguistics,* vol. I.

Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M., & Muñoz, R. 2000. An algorithm for Anaphora Resolution in Spanish Texts. *Computational Linguistics (submitted).*

Reinhart, T. 1983. *Anaphora and Semantic Interpretation.* Croom Helm linguistics series. Beckenham, Kent, BR3 1AT: Croom Helm Ltd.

Strube, M., & Hahn, U. 1999. Functional Centering - Grounding Referential Coherence in Information Structure. *Computational Linguistics,* **25**(5), 309–344.

Vossen, P. 1998. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newwsletter. K.Choukri, D.Fry, M. Nilsson (des). ISSN:1026-8200,* **3**(1).