Sentence disambiguation using a restricted Spanish lexicon extracted from WordNet

Miguel Angel Ibarra Rivera,
Aurelio López López,
Jesús Favela Vara

The availability of a well structured lexicon and of a parser that takes advantage of that structure for the analysis of sentences are of fundamental importance in a natural language processing system. A syntactic-conceptual analyzer of sentences in the Spanish language based on the cognitive model of Kavi Mahesh is presented. A Spanish lexicon extracted from WordNet is used for disambiguation. The terms in the lexicon are restricted to those used in the gardening domain. The structure defining each concept was widened, including for each one of the concepts three additional attributes: the actions that the entity defined by the concept usually performs, the actions that others execute on that entity, and the actions that the entity accomplishes on itself. The analyzer itself is constituted by a finite state automata, of the augmented transition network kind, and represents a part of the Spanish grammar with which the disambiguation method is illustrated. The procedure correctly desambiguates the sentences, though is sensitive to the placement of the associated actions in the concepts within the hierarchy of the lexicon. Based on a well-structured lexicon, the method operates correctly.

1 INTRODUCTION

Since the first automatic devices for information processing were built, mechanical as well as electronic, the idea of using them for the analysis of natural language texts has existed. Mainly, the purpose has been to accomplish reliable translations from one language to another, say English-Spanish, automatically, and to facilitate the man-machine interaction using natural languages.

On the other hand, by virtue of the information explosion to which we are exposed, we see an increasing need for systems that summarize documents written in natural language or support queries in daily English or Spanish. There is a need for document summaries weather on the Internet or out of it.

Furthermore, there are other fields that require of the natural language processing, to implement conversational agents [Allen *et al*., 1995] or support language training.

To achieve this, it is required to structure the linguistic and conceptual knowledge of the most important or necessary sections of our culture and to develop methodologies so that, interacting with these hierarchical structures or lexicons, we could automatically extract the meaning of the sentences and correctly interpret the texts.

One of the best efforts of knowledge structuring is represented by WordNet, a conceptual English dictionary available now after thirty years of work [Fellbaum 1998].

In this paper we describe a syntactic-conceptual analyzer of Spanish sentences using a WordNet subset for the disambiguation, considering terms of the activity designated as gardening.

In section 2 we present the problem of performing the syntactic-conceptual analysis of a sentence. In section 3 the syntactic-conceptual analyzer is described. It presents Mahesh's cognitive model, the syntactic analyzer based on a finite state automata, the lexicon extracted from WordNet and the syntactic-conceptual analyzer that was implemented using this lexicon.

The preliminary results obtained using this strategy are described in section 4 and, finally, in section 5 we present our conclusions and proposed future work.

2 SYNTACTIC-CONCEPTUAL ANALYSIS

2.1 Syntactic Analysis

Linguistic knowledge can be organized in a grammar, in the form of syntactic and semantic rules. That is to say:

Grammar: Syntax + Semantics

There are several types of grammars. Among them we find: phrase structure grammars, containing thousands of basic types of phrases; finite state grammars, that structure the linguistic patterns in finite state automata, and are complemented with statistical methods; another type of grammar consists of constraint-based formalisms. [Uszkoreit & Zaenen 1996]. In this work we use a finite state grammar, that will be described forward in Section 2.2.

2.2 Semantic Analysis

A single word may have various meanings, depending on the context of the discourse. For example, the word "ball" can be referred to a round object, or to a lavish formal dance, between other possible meanings, such the following sentences, containing it, show:

(1) The ball traveled 90 mph on his serve
(2) She was the loveliest girl at the ball
(3) He threw nine straight balls before the manager yanked him
(4) Jim launched a ball

In the first sentence, the word "ball" represents a round object that is hit or thrown or kicked in games, while in the second it plays the roll of a lavish formal dance. In the third sentence, it means a pitch that is not in the strike zone, as can be seen in WordNet. To disambiguate the sentence is necessary to consider the context in which those words are included in that sentence. However, in (4) there is no immediate manner of deciding. Thus, for the disambiguation to be automated, the different meanings of a word should be registered in a conceptual dictionary, known as lexicon.

2.3 Disambiguation

The main problem that exists in the analysis of a sentence is the selection of the correct meaning of that sentence. The different meanings of the words give rise to several possible meanings of the sentences. It is this variety of interpretations the one which we wish to resolve. This is known as disambiguation. There are two basic approaches to achieve the disambiguation: The rule based and the probabilistic one. In the rule based approach thousands of rules that anticipate the different situations are registered. This approach can't disambiguate everything, but when they do it, they do it well. In the probabilistic approach practically every sentence is disambiguated, but sometimes erroneously. This is by virtue of the fact that it does not leave ambiguous sentences, and upon forcing the disambiguation choosing the most probable meaning a mistake can be made.

2.4 Lexicons

A lexicon is a hierarchical structure, a tree or a set of trees, in which the nodes are words that represent concepts. In the hierarchy, the father of a node, or concept, is the category to which that concept belongs; while the children of a node, or concept, are the different types of instances of that concept.

For example, take the word "tree". In WordNet there are two nodes, in different branches, for this concept:

> *tree* -- (a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms)
>
> *tree, tree diagram* -- (a figure that branches from a single root; "genealogical tree"

The location of each one of these nodes in the hierarchy is:

> *tree* => woody plant, ligneous plant => vascular plant, tracheophyte => plant, flora, plant life => life form, organism, being, living thing => entity, something
>
> *tree, tree diagram* => plane figure, two-dimensional figure => figure => shape => attribute => abstraction

Some of their children nodes are:

> *tree*
> => yellowwood, yellowwood tree
> => lancewood, Oxandra lanceolata
> => negro pepper, Xylopia
> …
> => anise tree
>
> *tree, tree diagram*
> => cladogram

Recently it has been recognized the fundamental importance that the adequate construction of lexicons has [Pustejovsky 1995]. No natural language processing project can be carried to a real system if we don't have an adequate lexicon [Guthrie *et al.*, 1996].

To develop syntactic-conceptual analyzers in Spanish language is necessary to build Spanish lexicons. Such lexicons can be built by translating WordNet subsets manually. Paradoxically, it is not possible to do this translation automatically, neither partially. The reason is that any machine translation system that attempts to translate it to the Spanish would have to rely on a syntactic-conceptual analyzer that at the same time would require a Spanish

lexicon as complete at least as WordNet itself. But this is precisely the problem that we are dealing with.

We consider of great importance the development of new methodologies for the analysis of Spanish sentences based on hierarchical lexicons such as WordNet.

Nowadays, the University of Amsterdam is the coordinator of the EuroWordNet Project, that includes WordNet type lexicons in Dutch, Spanish, Italian, German, French, Czech, and Estonian. The Spanish WordNet has 23,370 nodes.

3 SYNTACTIC-CONCEPTUAL ANALYZER

3.1 Mahesh's Cognitive Model

Within the natural language processing models proposed we find a promising one, by virtue of the fact that it takes into account the form in which we disambiguate the sentences: the *Cognitive Model*, of Kavi Mahesh [Mahesh 1995]. Although we do not tie us faithfully to this model in our work, we were inspired by its general principles, and they are:

> *Principle 1. Early Selection*: The language processor selects a unique interpretation when the necessary information is available.
>
> It does not simply produce all the possible interpretations and then lets other modules select the best, but it rather considers that this selection or ambiguities resolution is an integral part of the task of understanding a sentence. Many other models violate this principle.
>
> *Principle 2. Incremental Interpretation*: The language processor produces interpretations incrementally, on line.
>
> Similarly as with the model HPSG [Pollard & Sag 1994], the cognitive model bases its incremental interpretation on evidence provided by recent psycholinguistic theories. Not only it is required for the processor to choose an interpretation as soon as possible, but it must do so incrementally, for example after reading each word, instead of doing it until ends reading the sentence.
>
> *Principle 3. Integrated Processing*: The language processor must apply any type of knowledge, syntactic, semantic, conceptual, as soon as it be available.
>
> If a piece of knowledge is available and is not applied, the processor will not be able to reduce the set of possible interpretations. This is in

agreement with the psycholinguistic literature: humans process the language in an integrated way [Tanenhaus & Trueswell 1995].

Principle 4. Functional Independence: The language processor will be able to apply any part of the knowledge independently if other types of knowledge are available or not.

This complements the integrated processing principle.

Principle 5. Determinism: The language processor will make no decision of any kind if there is no knowledge to justify it.

Thus, the processor will not make decisions at random. It will not choose randomly the interpretation of the set of possibilities, neither will choose the first interpretation of the list.

Next we describe the syntactic analysis and syntactic-conceptual analysis method proposed, which is based on Mahesh's cognitive model. One of the main differences of the method proposed with that of Mahesh reside in principle 4, related to functional independence. Explicitly, our prototype always does first a preliminary syntactic analysis that, when not definitive, continues with the conceptual analysis. On the other hand, the structure of the nodes of the lexicon that we use is different, since they contain the actions associated with the concept, as described in section 2.3.

3.2 Syntactic Analyzer

The syntactic analyzer is based on a finite automata, and corresponds to an augmented transition network. Having a whole slew of options to build it, we proceeded gradually. First we tested simple sentences, of the type: *Los jardineros cultivan unas azaleas*, that contains two noun phrases connected by a verb, and which corresponds to the graph shown in Figure 1.

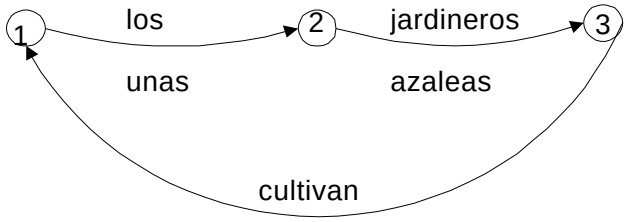

Figure 1. The first sentences.

The sentence is constituted by the subject *los jardineros*, which is the first nominal phrase, and the predicate *cultivan unas azaleas*, itself formed by the verbal phrase *cultivan* and the second noun phrase *unas azaleas*. When the input sentences does not follow this pattern, the sentences are not recognized. In the next stages, the sentences considered were of the style of: *Los eficientes jardineros nuevos cultivan hoy estas azaleas,* with graph shown in Figure 2.

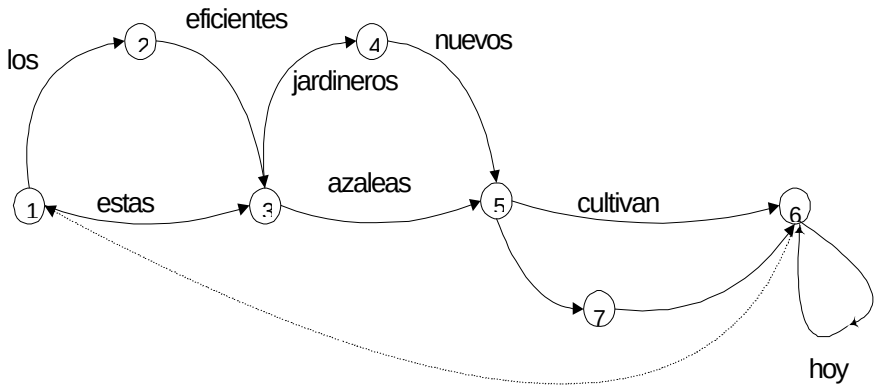

Figure 2. Example of a more complex sentence.

Finally, the sentences used were of the style: *Los hijos de los jardineros eficientes van a ir a cultivar mañana unas bonitas azaleas amarilla,* or well: *Los hijos de los jardineros piensan ir a construir una bonita mesa de madera de pino de California.*

The graph that we finally came up with has less than 20 nodes, it can analyze sentences of certain complexity. Once it has been seen that the method operates well, the graph will be increased, thus expanding gradually the grammar.

In each one of the states, the decision to go to the following state does not depend only on the entry represented in the exit arch but also of the entry to the following neighboring arch; that is to say, the path followed within the graph is determined by the reading of the current grammatical category and the one that follows.

3.3 The Lexicon

The lexicon was built based on the concepts that appear in a list of sentences sampled from gardening texts. Some of these sampled terms are: *abril, estanque, jardín, madera, sol, manos, fertilizante, esquinas, resequedad, jardinero*. In

total, there were extracted approximately 360 terms, representing all the grammatical categories, most of them nouns. These concepts were located in WordNet from which the restricted Spanish lexicon was built, based on the structure of that lexicon. The final number of terms considered was about 2000, because some additional concepts from that lexicon were included since they were necessary to define the hierarchy in the lexicon used.

For example, in WordNet it is registered that *jardinero* is an *horticulturista*, that at the same time is an *experto*, that at the same time is a *persona*, an *organismo*, an *entidad*. For this reason, in the new lexicon we must include, in addition to the term *jardinero*, the terms that are above in the hierarchy, from *horticulturista* to *entidad* (see Figure 3).

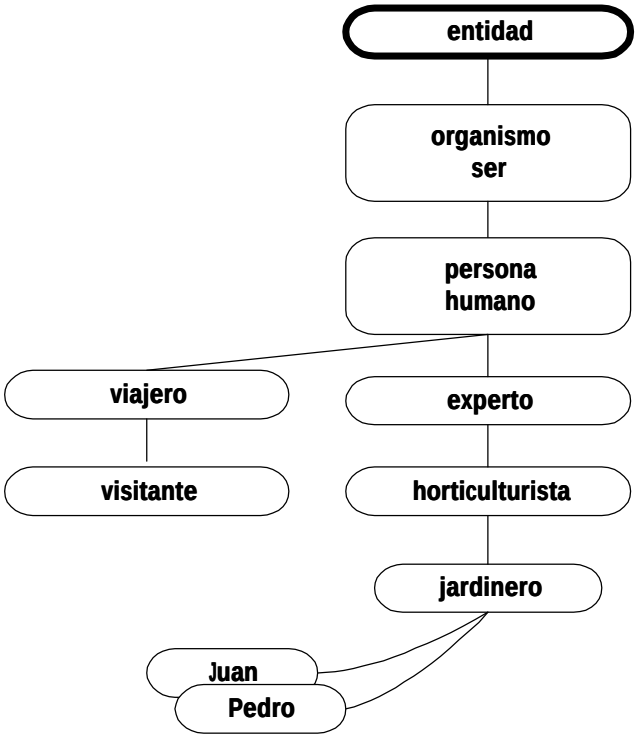

Figure 3. The hierarchy of jardinero, as extracted from WordNet.

Each node in the lexicon, an entry that represents a concept in the used lexicon, is structured as follows:

Concept:
Word
Grammatical category

Gender, number
Father concept
Children concepts
Actions that the entity accomplishes
Actions that others accomplish on it
Actions that the entity accomplishes on itself

For instance, the concept *jardinero* contains the following information:

jardinero
sustantivo
masculino, singular
horticulturista
Juan, Pedro
cultivar, sembrar, podar, …
…

Even though the *jardinero* concept has registered specific actions that it performs, in the used lexicon, it does not have verbs associated for the actions that others fulfil on him or that he accomplishes on itself in a specific way. But as jardinero has the following hierarchy: jardinero => horticulturista => experto => persona => organismo => entidad, he inherits from *organismo* actions that he can accomplish at the same time that someone can execute on him, such as *ver*, *oir*, *oler*, and from *entidad* it inherits *dañar*, *perjudicar*, among others.

WordNet does not include information on the actions that the concept accomplishes, others accomplish on it, or it accomplishes on itself. According to Pustejovsky [Pustejovsky 1995] , each node of a lexicon must ideally include information of the following four interpretive levels, known as the qualia structure: (1) constitutive, the relation between an object and its constituents or proper parts; (2) formal, that which distinguishes the object within a larger domain; (3) telic, purpose and function of the object; (4) agentive, factor involved in the origin or "bringing about" of an object. In our case, we try to include also information of the telic field, of Pustejovsky's qualia structure.

3.4 The Syntactic-conceptual Parser

In the tour through the graph that represents the grammar, the analyzer tries to identify the different parts of the sentence, without taking a definitive decision on the role that each one of the input words plays. Only assumptions are made. The posible subject and predicate are identified, with all its components, and it is checked if ambiguity exists; in this case, the lexicon is consulted.

Sometimes, a simple syntactic inspection defines the subject and the predicate, as in this sentences: *Las señoras fueron a Mexicali*, or *Las azaleas están sobre la mesa*. The presence of prepositions in a sentence, in this case *a* and *sobre*, is definitive. But in sentences as *Los jardineros cultivan unas azaleas amarillas* the syntactic analysis alone can't decide if *los jardineros* is the subject or if *unas azaleas amarillas* is it. In this circumstance, the lexicon is inquired. In it, it is registered that one of the actions that *los jardineros* are accustomed to do is *cultivar*, and that one of the actions that it is regularly applied to the *azaleas*, that are plants, is *cultivar*las.

It could be that the action, the verb *cultivar*, was not registered directly in the node corresponding to the possible kernel of the subject, *jardinero*. In this case, the father node is consulted. If the action *cultivar* is not found here, we continue seeking until arriving to the root. Every action made by a term or concept associated to a node is inherited to all the subtree below it. Also, if it were registered in the lexicon that all the plants are cultivated, then the *azaleas*, being plants, also are cultivated.

Even with this restricted grammar we can already implement the disambiguation method based on the lexicon.

4 PRELIMINARY RESULTS

The output of the analyzer is of the following type:

Sentence to analyze:
unas plantas trepadoras hoy podará el entusiasta jardinero nuevo

SENTENCE:

PROBABLE SUBJECT 1: noun phrase [1]:
Determiner[1] (fp): *unas*
Noun[1] (fp): *plantas*
Adjective_p[1] (fp **):** *trepadoras*

VERBAL PHRASE:
Adverb[1]: *hoy*
Verb[1]: *podará*

PROBABLE SUBJECT 2: noun phrase [2]:
Determiner[2] (ms): *el*
Adjective_a[2]: adjetivo(ms**):** *entusiasta*
Noun[2] (ms): *jardinero*

Adjective_p[2] (ms**):** *nuevo*

There is gender and number concordance in the noun phrase 1.
There is gender and number concordance in the noun phrase 2.

SUBJECT: noun phrase[2]
PREDICATE: VERBAL PHRASE + noun phrase[1]

In this case, (fp) is intended for the gender and number of the word considered, that is feminine plural, while (ms) stands for masculine singular. In case that there is not gender and number conformity in some noun phrase, a warning message is sent, but the analysis process is not stopped. On the other hand, when one of the words of the sentence is not registered in the lexicon the process is halted.

The analyzer is not always able to disambiguate the sentences, but only in 87% of the given sentences. However, the above mentioned sentences were specifically built for this project and we have not tested the parser against real world text. For example, in *Los jardineros nuevos prefieren las azaleas rojas* it is not possible to determine the subject since the verb *preferir* is included between the actions that execute, and are executed by, the *organismos* in general, in this lexicon.

It can be decided that only certain types of organisms could execute the action of *preferir*, in this case the persons; but this would carry us later to commit mistakes in sentences such as *Las azaleas prefieren los lugares sombreados*. One way to handle this problem is to consider the probability of use, although in this case it might still make a mistake.

This kind of examples show that the greater difficulty of building a syntactic-conceptual analyzer that disambiguates correctly is the adequate construction of the lexicon. In fact, it is not enough to translate into Spanish a subset of WordNet but one must adequately locate in the hierarchy the actions that "commonly" the instances of the concepts performs. But this task no longer corresponds to the field of computer sciences, neither to linguistics, nor to psychology. This type of tasks correspond to interdisciplinary teams of specialists as the one which built WordNet.

The execution time of the algorithm used in the analyzer depends directly on the number of words that the sentence contains and on the size of the lexicon. Basically, the time is used in the binary search within the dictionary. Using a 2,000 entries lexicon, in a personal computer, the analysis is accomplished in less than a second.

5 CONCLUSIONS AND FURTHER WORK

The method proposed to achieve the syntactic analysis of Spanish sentences seems to be adequate. The grammar can be augmented easily and gradually. Also, the procedure to disambiguate using the lexicon works fine, as long as the lexicon is well designed and complete. This indicates that to reach an efficient and reliable syntactic-conceptual parser for Spanish, we have to focus on forming multi-disciplinary teams to design and build a Spanish lexicon of reasonable size. Using the Spanish version of EuroWordNet, we estimate that in a year would be ended the construction of the augmented lexicon.

Further work to do for the prototype of language processor for Spanish includes:

Expand the grammar, to consider more general types of noun phrases and of verbal phrases. What is immediate is a complete treatment of all types of prepositional Spanish phrases, taking into consideration the different roles that each preposition can play.

Continue reviewing and expanding the lexicon, following the qualia structure proposed by Pustejovsk, based on WordNet, as well as to systematize the constructions of lexicons in different domains. A graphical environment to easily edit lexicons will be of great help for this task.

Include morphology analysis to improve the parser, mainly in verbal expressions.

Apply the prototype of syntactic-conceptual parser on a natural language interface to interact with a system for indexing and information retrieval in digital libraries [Favela & Acosta 1999]. This interface is planned to be used in a collaborative system intended for learning computer programming, to implement natural language generation and understanding of a virtual student.

REFERENCES

Acosta, R., Favela, J., 1999. MIND: *An environment for the capture, indexing and retrieval of mixed-media from a digital library of graduate thesis*. First NSF/CONACYT Workshop on Digital Libraries, Albuquerque, New Mexico, Julio 7-9, 1999.

Allen, J.F., Schubert, L.K., Ferguson, G., Heeman, P., Hwang, C.H., Kato, T., Ligth, M., Martin, N.G., Miller, B.W., Poesio, M., Traum, D.R., 1995.

The TRAINS Project: A case study in building a conversational planning agent. Journal of Experimental and Theoretical AI, 7(1995), pp. 7-48.

Fellbaum, C, (Editor), 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, 1998.

Guthrie, L., Pustejovsky, J., Wilks, Y., Slator, B. M., 1996 *The Role of Lexicons in Natural Language Processing*. Communications of the ACM, Vol. 39 (1996), Num. 1 pp.63-72.

Mahesh, K., 1995. *Syntax-Semantics Interaction in Sentence Understanding.* Ph.D. Dissertation. Computer Science Departament. Georgia Institute of Technology, 1995.

Pollard, C., Sag, I., 1994 *Head-driven Phrase Structure Grammar*. Center for the Study of Language and Information Lecture Notes. Stanford University Press, 1994.

Pustejovsky, J., 1995. *The Generative Lexicon*. MIT Press, Cambridge, 1995.

Tanenhaus, M.K. and J.C. Trueswell, 1995. *Sentence comprehension*. In J. Miller and P. Eimas, eds., *Handbook of Perception and Cognition* Vol. 11: Speech and Language. New York: Academic Press. 1995.

Uszkoreit, H., Zaenen, A., 1996. *Grammar Formalisms*. Chapter 3 in *Survey of the State of the Art in Human Language Technology*. 1996. http://cslu.cse.ogi.edu/HLTsurvey/

Miguel Angel Ibarra Rivera is a doctoral student at the Computer Science Department, Centro de Investigación Científica y de Educación Superior de Ensenada, CICESE. Km. 107 Carretera Tijuana-Ensenada, Ensenada, B.C., México. He can be reached at mairx@bahia.ens.uabc.mx

Aurelio López López is a researcher and professor at the Computer Systems Engineering Group, Instituto Nacional de Astrofísica, Optica y Electrónica, INAOE. Apdo. Postal 51 y 216, 72000, Puebla, Pue., México. He can be reached at allopez@inaoep.mx

Jesús Favela Vara is a researcher and professor at the Computer Science Department, Centro de Investigación Científica y de Educación Superior de Ensenada, CICESE. Km. 107 Carretera Tijuana-Ensenada, Ensenada, B.C., México. He can be reached at favela@cicese.mx

This work was partially supported by CONACYT under grant 29729A and scholarship Id 266 provided to the first author.