Three Mechanisms of Parser Driving for Structure Disambiguation

Sofía N. Galicia-Haro, Alexander Gelbukh, and Igor A. Bolshakov

Natural Language Processing Laboratory. Center for Computer Research,
National Polytechnic Institute. Juan de Dios Batiz s/n 07738 México D.F.
{sofia, gelbukh, igor}@cic.ipn.mx

Abstract. Structural ambiguity is one of the most difficult problems in natural language processing. Two disambiguation mechanisms for unrestricted text analysis are commonly used: lexical knowledge and context considerations. Our parsing method includes three different mechanisms to reveal syntactic structures and an additional voting module to obtain the most probable structures for a sentence. The developed tools do not require any tagging or syntactic marking of texts.

1 Introduction

Structural ambiguity while parsing takes place because the syntactic information alone does not suffice to make a unique structure assignment. Researchers use now knowledge-intensive techniques for disambiguation. These techniques required a lot of manually coded information. Thus, they could be inapplicable to unrestricted texts.

Probabilistic context-free grammars (CFG), being introduced for choosing between alternative parses gave disappointing results too, which can be explained by the fact that these grammars do not reveal dependencies between words. In research introducing lexical dependencies, some approaches towards disambiguation have been recently tried: basic noun phrase chunking and bigrams [1], mutual information to obtain lexical attraction between content words [2], etc. But the former approach remains ambiguity in the basic noun phrase, while the latter approach misses the prepositional links to previous words. In this paper we propose a model for syntactic analysis and disambiguation based on three different methods.

2 Overview of the Model

We mainly consider two kinds of disambiguation mechanisms: the first referred to as *lexical* regards dependencies between predicative words and their arguments, and the second uses a *local context* in order to solve syntactic disambiguation that is mainly constituted by adjunct arguments.

In the examples

1. *Juan lanzó una pelota sobre el puente* 'Juan threw a ball onto the bridge'
2. *Juan admitió una plática sobre el puente* 'Juan admitted a communication on the bridge,'

A. Gelbukh (Ed.): CICLing 2001, LNCS 2004, pp. 192-194, 2001.

the prepositional phrase *sobre el puente* is attached to the verb *lanzó* or to the noun *plática*. The difference is due to lexical preference. The Spanish verb *admitir* does not accept preposition *sobre,* whereas the verb *lanzar* does.

In the examples

3. *Juan bebe licores con menta* 'Juan drinks spirits with mint'
4. *Juan bebe licores con sus amigos* 'Juan drinks spirits with his friends'

the phrase *con menta* is attached to the noun *licores*, while the phrase *con sus amigos* 'with his friends,' to the verb *bebe* 'drinks.' The disambiguation could be resolved by the local context, i.e., by the fact that 'mint' is semantically closer to 'spirits' and 'friends' is closer to 'drink'. In our model both mechanisms are used in parallel.

For syntactic ambiguity resolution we propose the voting between of the outputs of different syntactic structure assignment subsystems. The overall system is presented in Figure 1. Each module gives a set of weighted variants. Those weights are based on the satisfied characteristics in each method. So, each module gives a quantitative measure of the probability of each syntactic structure in a dependency structure format.

The quantitative character gives the advantage for combining several methods as it is shown in Figure 1.

Three Mechanisms. The system includes government pattern (subcategorization frame) module, semantic proximity module, and extended CFG module. The three mechanisms require the compilation of several linguistic resources: the advanced government pattern (PMA) dictionary [3], the semantic network [4], and the extended CFG rules.

Advanced Government Patterns. This mechanism considers the linguistic knowledge contained in the so-called syntactic valences. It is the main mechanism of the model. It is the most practical to resolve most of the structure ambiguities, but does not suffice for all of them. The knowledge in this module is the lexical information about valences of verbs, adjectives, and nouns [5]. We developed a statistical version of this approach [6].

Extended CFG. It is the simplest method to compile parsing tools. We have created an extended CFG for Spanish language with agreement in gender and number and then have implemented it in a chart parser. We assumed equally weighted variants for the CFG module.

Semantic Proximity considers local context knowledge. When several structures are quite possible or adjuncts attachment is ambiguous the semantic proximity, i.e., the concepts more close related to the words in the possible constituents, could help to disambiguate structure variants. The idea behind the semantic proximity is finding the shortest paths between constituents obtained from the CFG module. For this purpose we assign different weights to relations, hierarchy concepts links and implicit relations.

Voting Module. To disambiguate syntactic structures, the module uses the weights assigned in each module, voting for the maximum among variants. The result is a ranked list of the syntactic variants.

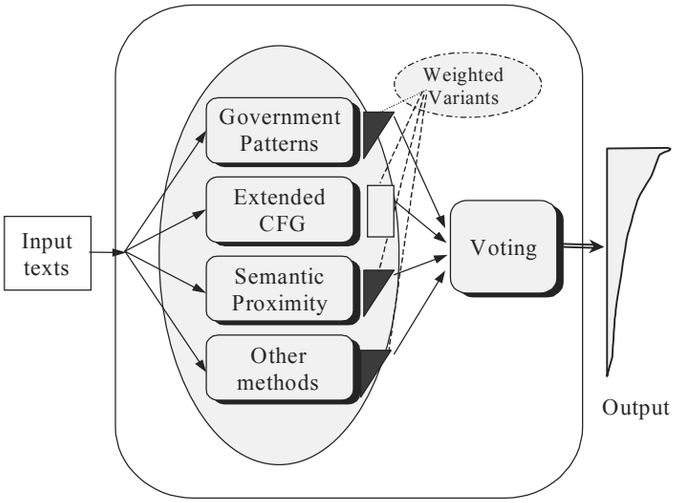

Figure 1. The model for syntactic disambiguation.

3 Conclusions

We propose three mechanisms for syntactic structuring and a voting disambiguation module for quantitative comparison of results. Each mechanism is realized as a module outputting a set of weighted parsed variants. The output is a ranked set of variants in a dependency structure format. The experiment on a corpus of 100 sentences confers the 1st rank to the correct parsing variant in 35% of cases.

References

1. Collins, M.: Head-driven Statistical Models for Natural language parsing. Ph.D. thesis. University of Pennsylvania. (1999) http:// xxx.lanl.gov/ find/cmp-lg/
2. Yuret, D.: Discovery of Linguistic Relations Using Lexical Attraction. Ph. D. thesis. Massachusetts Institute of Technology. (1998) http:// xxx.lanl.gov/ find/cmp-lg/9805009
3. Galicia-Haro, S. N, A. Gelbukh, I.A. Bolshakov: Advanced Subcategorization Frames for Languages with Relaxed Word Order Constraints. Natural Language Processing VEXTAL. (1999) 101- 110.
4. Gelbukh, A. F.: Lexical, syntactic, and referential disambiguation using a semantic network dictionary. Technical report. CIC, IPN (1998)
5. Mel'čuk, I.: Dependency Syntax: Theory and Practice. State University of New York Press (1988)
6. Gelbukh, A. F.: Extracción de un Diccionario de PMA http://www.cic.ipn.mx/ ~gelbukh/REDII/REDII-99