COMBINING DEPENDENCY AND CONSTITUENT-BASED RESOURCES FOR STRUCTURE DISAMBIGUATION *

SOFÍA N. GALICIA-HARO, ALEXANDER GELBUKH, and IGOR A. BOLSHAKOV

Center for Computing Research, National Polytechnic Institute
Av. Juan de Dios Batiz S/N, 07738 México, D. F.
{sofia,gelbukh,igor}@cic.ipn.mx

Abstract

Unrestricted text analysis requires an accurate syntactic analysis but structural ambiguity is one of the most difficult problems to resolve. Researchers have tried different approaches to obtain the correct syntactic structure from analyzed sentences but not successful results have been obtained.

Two different approaches have traditionally applied to syntactic analysis: constituent grammars and dependency grammars. We propose a model for syntactic analysis and disambiguation combining lexical dependencies and semantic proximity. Lexical dependencies are applied by means of a government pattern dictionary following the dependency approach. The semantic proximity is introduced by means of semantic closeness among constituents. Examples are given to illustrate method's contributions.

Keywords

Syntactic analysis, semantic proximity, context free grammar, dependency grammar, lexical dependencies.

1 Introduction

Structural ambiguity is one of the most difficult problems to resolve in natural language processing systems. Structural ambiguity takes place because the syntactic information alone does not suffice to make a structure assignment decision. In the constituent approach, recent researches have been developed probabilistic context free grammars (PCFG) as a means of choosing between alternative parses of the same string of words, i.e. for disambiguation. The disappointing results of PCFG can be explained, because such kind of grammars is typically unable to express dependencies between words. In addition, for a good coverage in an accurate syntactic analysis detailed lexical information is required [1, 3, 7].

In recently research lines introducing lexical dependencies, some approaches towards disambiguation have been tried: basic noun phrase chunking and bigrams [4], mutual information to obtain lexical attraction between content words [12]. However, in some languages with a wide use of simple and composed prepositions the noun phrases could include prepositional groups that increase the ambiguity of syntactic structure. Also, pair of word statistics has a high impact in the syntactic analysis of languages with stricter word order constraints.

In languages with relaxed word order constraints and wider preposition use as Spanish, the argument structure of verbs, adjectives and nouns permit to reduce the quantity of variants for noun and prepositional phrases. The argument structure of words could be mainly defined establishing their dependencies with prepositions. But besides the incorporation of lexical knowledge based on dependencies among words, structure disambiguation requires the incorporation of semantic knowledge based on certain type of local context to link adjuncts, i.e. complements not related to word's meaning.

The incorporation of all that knowledge implies a huge manual effort that restricts its application to limited domains. Even including lexical and semantic knowledge in a syntactic analyzer because of the impossibility of a complete description, a certain quantity of variants would be obtained and a disambiguation mechanism should be required.

In this work, a model for syntactic analysis and disambiguation is proposed. The model is based on lexical dependencies among predicative words and on semantic proximity. The first one follows the dependency approach by means of government patterns described in Meaning ⇔ Text Theory [8]. The semantic proximity is supported on a semantic network use to incorporate local context in a syntactic analyzer that follows the constituent approach by means of a context free grammar (CFG). For syntactic disambiguation we propose the classification of all syn-

* Work partially supported by CONACyT, SNI, and CGEPI-IPN, Mexico.

0-7803-7087-2/01/$10.00 © 2001 IEEE

Sofia N. Galicia-Haro, Alexander Gelbukh, Igor A. Bolshakov. *Combining dependency and constituent-based resources for structure disambiguation.* IEEE International Workshop on Natural Language Processing and Knowledge Engineering, NLPKE 2001, IEEE, pp. 2873–2879; www.Gelbukh.com/CV.

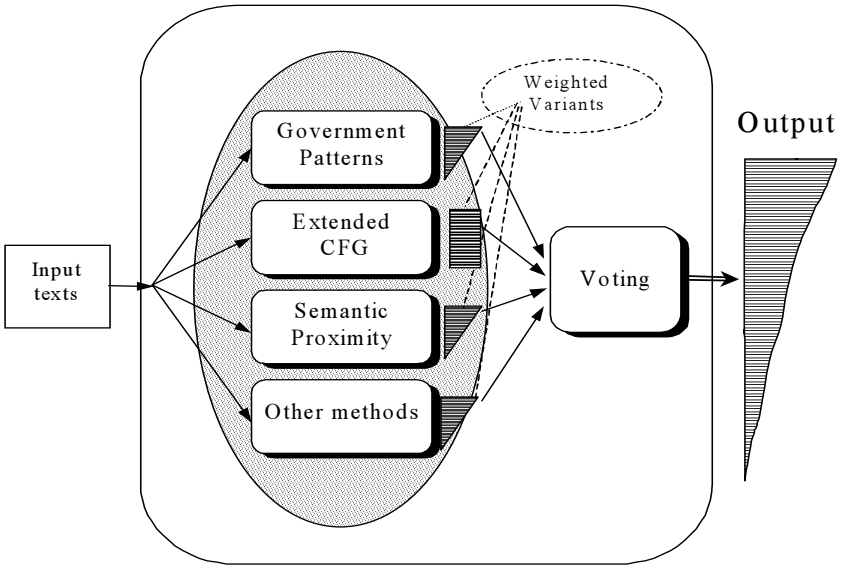

Figure 1. The Model for syntactic structure assignment and disambiguation

tactic structure variants obtained, which have a weight, based on characteristics satisfied in each method.

Our model is addressed to unrestricted text analysis. The developed tools to compile the necessary resources do not require any word tagging or any type of syntactic annotation.

The overall model is presented first and then each one of the components.

2 Overall Model

The model proposed assigns syntactic structures by means of three different modules that process unrestricted texts in parallel. Each module corresponds to one different method that represents a specific knowledge. Each module gives a set of weighted variants based on the satisfied characteristics in each method. The output of each module gives a quantitative measure of the probability of each syntactic structure, in a compatible format choused: the dependency structure format.

The three modules correspond to government patterns module, semantic proximity module and extended CFG module. The three methods require the compilation of dictionaries: the advanced government patterns (PMA) dictionary, the semantic network and the extended CFG rules.

To disambiguate syntactic structures a voting module uses the weights assigned in each module, voting for the maximum added value of variants. The result is a classified list of the syntactic variants.

The overall model is presented in Figure 1. The voting module input is the output of each module, i.e. the weighted variants obtained from each method. In the figure we emphasize that new methods could be considered in the future for the same model because of quantitative measure consideration. The three resources required for the module are:

- Extended CFG Rules
- Government Patterns dictionary
- Semantic Network

3 Extended CFG Rules

The extended Context Free Grammar module is the simplest method to compile tools and to apply them. We created an extended CFG for Spanish language (with gender, person and number concordance) and we implemented a chart parser. This parser construction considers two goals, one is its use in the extended CFG Rules module and the second one is obtaining the syntactic information for the government patterns dictionary.

The grammar does not require optimal conditions on coverage and precision because it is not the main method. Our grammar is intended to cover the most usual language constructions. It includes several improvements:

- Agreement concordance (gender, number, and person)
- Government element
- Syntactic relations
- Punctuation elements
- Semantic annotation of time
- Rule weights

The government element is required to transform constituent structures to dependency structures. The rule weights are used to filter the type of rules applied to parsing. Rules for more common constructions have higher priority.

The created grammar employs the LEXESP[1] POS tags that are not disambiguated. The tags correspond to PAROLE categories [2].

The syntactic categories alone do not help to differentiate the correct variants. More complex methods for tree structures classification or for probabilistic rules should be used to assign weights to variants in this method. So, we assume equally weighted variants for the CFG module. This assignment allows that the other methods make arise the correct variants.

4 Government patterns

The government patterns module considers the linguistic knowledge acquired by native speakers in their very early language learning. By this reason it is the main method of the model. The knowledge described in this module is the lexical information of verbs, adjectives and nouns about their valences. These lexical dependencies between predicative words express the preference of each word to link its arguments. We developed a statistical model to compile the syntactic information for this dictionary. The weights assigned are related to how well the patterns are matched.

In the *Meaning⇔Text Theory* (MTT) [8], syntactic dictionary zone describes correspondence between semantic and syntactic valences of the headword, all ways of realization of the syntactic valences, and the indication of obligatoriness of presence for each actant (if necessary). To accomplish this goal, the dictionary presents a government pattern table (GP) [11], and the GP restrictions; also examples are usually given. The restrictions considered in GP can be of any type (semantic, syntactic, or morphological). The compatibility between syntactic valences is considered among those restrictions. The examples cover all possibilities: examples for each actant, examples of all possible combination of actants, and finally examples of impossible or undesirable combinations.

The main part of a GP is the list of syntactic valences of the headword. These are listed in a rather arbitrary order, though the order of growing obliqueness is preferred: subject, then direct object, then indirect object, etc. Each headword usually imposes certain order on its actants. Also the way of expression for headword meaning[2] influences this order. For example, the expression for Spanish *acusar*$_1$: person X accuses person Y of action Z.

Other obligatory information at each syntactic valence is the list of all possible ways of expressing this valence in texts. The order of the options for a given valence is arbitrary, though the most frequent options usually go first. The options are expressed with symbols of parts of speech or specific words.

Following the notation of GP table and list of examples, we can describe the Spanish verb *acusar*$_1$ as:

1 = X	2 = Y	3 = Z
1. NP	2. *a* NP	1. *de* NP 2. *de* INF
Obligatory	Obligatory	

Possible:

C.1 + C.2	La policía *acusa* a Ana.
C.1 + C.2 + C.3.1	La policía *acusa* a Ana de robar.

Prohibited:

C.1 + C.3.1	La policía *acusa* de robar.
C.3.1	*Acusa* de robo.

Though in a complete dictionary, syntactic zone specifies all possible examples after the GP table, in our example only some elements of the complete list of possible combinations of actants and some elements of the list of examples for impossible or undesirable combinations were showed. Obligatory indication was the only consideration for impossible combination examples. We omitted the list of examples for each actant because they are implicit in the previous examples.

For the model, we propose a new structure for GPs named Advanced government patterns (AGP) that considers the inclusion of additional information besides of a modernized format required for compu-

[1] H. Rodríguez from Universidad Politécnica de Cataluña, Barcelona, Spain, kindly provided LEXESP corpus.

[2] We use a brief English description to avoid specific Spanish words.

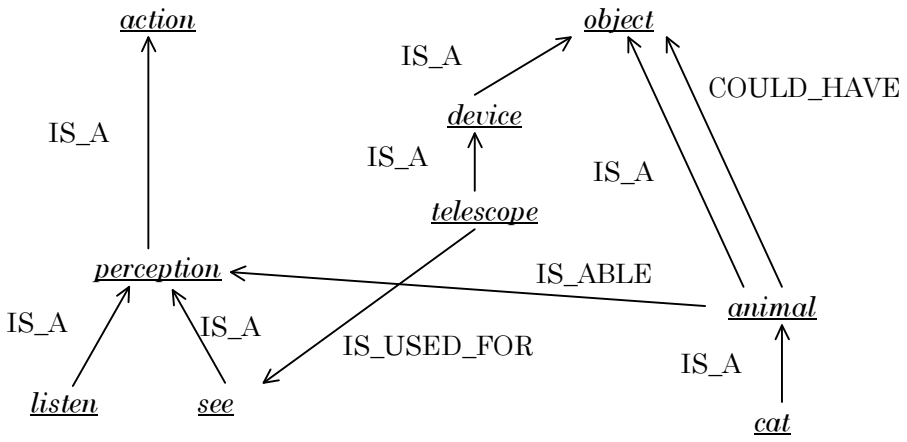

Figure 2. Semantic network fragment for the phrase *I see a cat with a telescope.*

tational dictionaries where the characteristics are represented on an attribute- value system as the modern formalisms use, for example HPSG. The new attributes represent information for some Spanish characteristics (animity, valence repetition) and probabilities of common use.

Manual work is the traditional way for syntactic information acquisition for GP's so its application has been limited. To compile this dictionary in what syntactic information refers we propose a method to obtain the objects of verbs, nouns and adjectives that is based in the statistics of analysis variants. These variants are the combinations of individual words with prepositions. Considering only POS, these combinations would be the components of subcategorization frames but specific for each word and each word could be verb, adjective or noun. The selection of this type of combinations is not random. The combinations are fixed in a good degree so their statistics are trustier than that of arbitrary pair of words. The detailed development of the method is presented in [5].

The weight assigned to each variant depends on the total number of patterns and the weights of matched valences. Also it depends on the type of patterns, on the realization valence frequencies and on the number of homonyms.

5 Semantic Network

The semantic proximity refers to 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 extended CFG module. For this purpose we assign different weights to relations, hierarchy concepts links and implicit relations.

The semantic network construction is an extreme intensive labor task with difficult realization even in a long period of time. In this work we consider the semantic network that is being developed from the FACTOTUM[3] network by means of a translation method [6]. To resolve syntactic ambiguity the links between word or group of words are realized determining how closer are they in a semantically view.

Semantic proximity is based on the semantic network characteristics, i.e., concepts, relations and paths. We define semantic proximity as a quantitative value; this idea has been employed by [10], [9]. To define it not only the length of the path is considered but also a weight assigned accordingly to relation type.

The proximity between a pair of words is a value depending on the length and the relation type, i.e. depending on:

- A value for each type of relation
- Specific values for individual links
- A higher value for implicit relations

The first assignment considers the values of the explicit relations. The second one intend to correct the problem appearing when relations are closer of the hierarchy top, as the relations are far from the top more common aspects they share.

The third assignment considers the problems of inferences. For example, *car* IS_AN *object* and *object* HAS_SUBTYPE *book*. Then the path is shorter

[3] *FACTOTUM® SemNet*, is a semantic network dictionary compiled by P. Cassidy of MICRA, INC. New Jersey, USA.

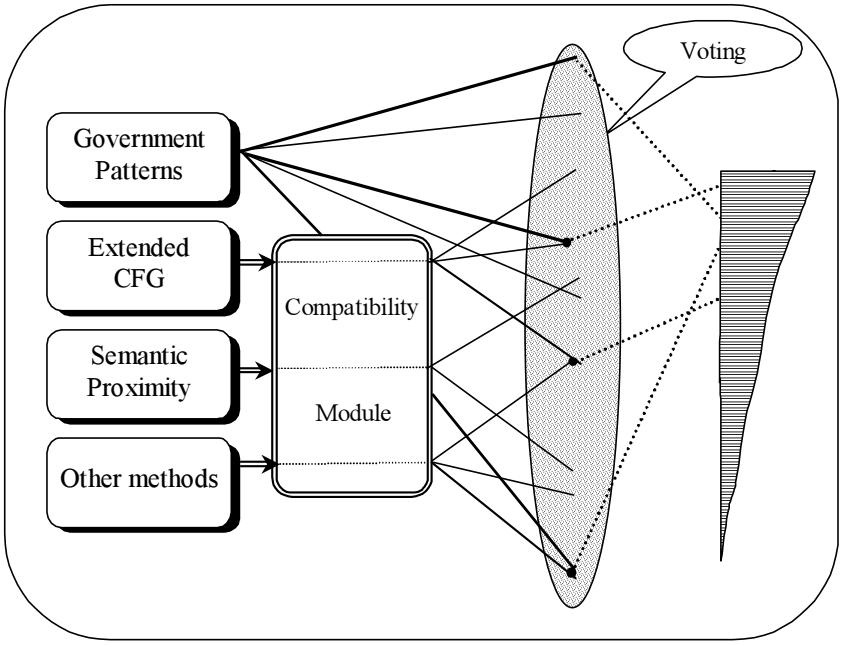

Figure 3. Disambiguation module

even there are no common aspects. To resolve this problem a higher weight is assigned to an implicit relation than an explicit one. The precision is obtained with the second assignment that makes longer the path of *car* IS_AN *object* than that of *Ford* IS_A *car*.

Syntactic disambiguation. Really we are incorporating the semantic component missing in the extended CFG module when we use the semantic network for syntactic disambiguation.

The syntactic structure in this module is taken from the output of the extended CFG module. Some grammars of the constituent approach incorporate semantic restrictions as HPSG that consider them in each dictionary entry. This is equivalent to have the internal semantic network of each word with links to the possible related words in any sentence, which implies an extreme intensive labor.

In our model, the semantic restrictions are searched in the network and they are defined through the semantic proximity that involves the shortest length between pair of words and their assigned value. The proximity evaluation not only is referred to network characteristics, it requires the syntactic type of the relation. Not all paths are admissible in a specific context. In some cases, it is necessary to find the paths with the more adequate relations to the syntactic context of the sentence. For example, the phrase *I see a cat with a telescope* has the prepositional phrase *with a telescope* where the closer relation will be USE and a relation IS_A will not be the more adequate to the context (Figure 2).

There is a very high quantity of paths connecting two words in a semantic network. Disambiguation task then is very related with the method to search the minimum acceptable paths and the syntactic context. The mathematical details for the solution and computational implementation are describe in [6].

6 Voting Module

The voting module employs the assigned weights in each module to disambiguate syntactic structures, voting for the added maximum value of variants (Figure 3). The result is a classified list of the syntactic structures. In order to realize the voting our model requires a quantitative evaluation and a compatible format.

We present an example for the phrase *El productor trasladó la filmación de los estudios al estadio universitario* ('The producer moves the filming from the studios to the university stadium'), to show the ideas behind the model. The data for each module is the following:

Government patterns

4.34896 *trasladar*,dobj_suj,obj:a,obj:de, x:?
0.436967 *trasladar*, obj:a, x:?
1.13758 *filmación*
3.29976 *estadio*

The numbers represent the obtained values from the method to compile the syntactic information for

the GP's. The mark "x:?" represents a duplicate valence by means of clitics.

Considering GP's for *trasladar* 'move', *estadio* 'stadium' and *filmación* 'movie production' the variants with the structure: "moving something from one place to another place" are favored.

Extended Context Free Grammar

8 variants equally weighted.

Semantic Proximity

We obtained the following relations:

filmación 'movie production'	→	*director* 'producer'
	→	subtype of spectacle
trasladar 'to move'	→	direction or place reference
	→	related to object translation
	→	subtype of path
estudio cinematográfico 'movie studio'	→	place
estadio 'stadium'	→	subtype of spectacle
filmación 'movie production'	→	*cine* 'movie' → *director* 'producer'

The relation between *trasladar* 'to move' and *estudio* 'studio' as a place is the only one that could be considered favoring the structure: translating from a place (*trasladar de los estudios* 'move from the studios').

For the example, the Government pattern method is the main contributor to reveal the correct variants; they emerge by means of the GP weights and then because of semantic proximity weight.

Conclusions

We proposed a syntactic analysis scheme considering three knowledge sources: lexical, syntactic and semantic. This knowledge permit to differentiate the correct variants among all variants obtained from parsing.

The three methods proposed are: government patterns that consider lexical and syntactic knowledge, extended CFG considering syntactic knowledge and semantic network that considers semantic closeness between syntactic groups.

We showed how these methods contribute to reveal the correct syntactic structure of the analyzed sentences.

Our model is intended for unrestricted texts and the developed tools do not require any manual tagging or syntactic marking of texts.

References

1. Charniak, E. *Statistical parsing with a context-free grammar and word statistics*. Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI MIT Press, Menlo Park, 1997. http://www.cs.brown.edu/people/ec/home.html #publications
2. Civit, M e I. Castellón. *Gramesp: Una gramática de corpus para el español*. Revista de AESLA, La Rioja, España, 1998.
3. Collins, M. *A new Statistical Parser Based on Bigram Lexical Dependencies*. In Proceedings 34th Annual Meeting of ACL pp.184-191. 1996.
4. Collins, M.: *Head-driven Statistical Models for Natural language parsing*. Ph.D. thesis. University of Pennsylvania. 1999. http://xxx.lanl.gov/find/cmp-lg/
5. Galicia-Haro, Sofía N. & Gelbukh, Alexander F. & Bolshakov, Igor A. *Acquiring Syntactic Information for a Government Pattern Dictionary from Large Text Corpora*. In this volume.
6. Gelbukh, A. F. *Lexical, syntactic, and referencial disambiguation using a semantic network dictionary*. Technical report. CIC, IPN, 1998.
7. Magerman, D. M. *Statistical decision-Tree Models for Parsing*. In Proceedings 33rd Annual Meeting of ACL. June 26-30 Cambridge, Massachusetts, USA, pp. 276-283, 1995. http://xxx.lanl.gov/ps/cmp-lg/9504030
8. Mel'cuk, I. A. *Dependency Syntax: Theory and Practice*. State University of New York Press. Albany
9. Rigau, G., Atserias, J. and Agirre, E. *Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp. 48-55, 1997 http://xxx.lanl.gov/ps/cmp-lg/9704007
10. Sekine, S., Carroll, J. J., Ananiadou, S. and Tsujii, J. *Automatic Learning for Semantic Collocation*. In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, pp. 104-110, 1992.

11. Steele, J. Meaning – Text Theory. Linguistics, Lexicography, and Implications. James Steele, editor. University of Ottawa press.

12. 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