ACQUIRING SYNTACTIC INFORMATION FOR A GOVERNMENT PATTERN DICTIONARY FROM LARGE TEXT CORPORA [*]

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

There are some research lines in automatic subcategorization frame acquisition and the importance of their work could not be doubted. However, almost all automatic work has been done in the constituent approach. Conversely, manual work is the traditional way for syntactic information acquisition in the dependency approach, which considers the correspondence between semantic valences and theirs syntactic realizations.

The last approximation has some advantages for description of languages with relaxed word order constraints and a vast prepositional use.

Our work is intended to compile automatically a government patterns dictionary in what syntactic information is referred to and to give a tool to facilitate linking of valences and meaning.

Keywords

Syntactic analysis, dependency grammar, constituent grammar, statistic method.

1 Introduction

Several authors have been working in automatic subcategorization frame acquisition [5, 1] and its importance could not be doubted, what's more syntactic information resources are essential for NLP. Almost all automatic work has been done in the constituent approach, i.e., considering classes that groups information for several verbs. Conversely, manual work is the traditional way for information compilation in the dependency approach.

In the last approach, semantic valences correspond to the participants (or actants) implicated in the specific meaning of verbs, adjectives and nouns. As generally each semantic valence usually have a determined syntactic realization, the recognition of these realizations for a given word leads to the specific meaning of the headword. For example government patterns in the Meaning ⇔ Text Theory (MTT) [11], which considers the correspondence between semantic valences and theirs syntactic realizations, and the relationship between valences and headword meaning.

Usually, linguists have done this kind of work and huge efforts are required to manually compile a government pattern dictionary [9]. Although manually compiled dictionaries usually have some disadvantages, e.g. new use of words is neglected, not mentioning compilation errors and specific topic domain.

There is syntactic information associated to verb meaning that is usually part of subcategorization frames (SF). For example, the use of specific prepositions for specific verbs, in Spanish "*intervenir en* NP" has the meaning *to participate* and "*intervenir a* NP" has the meaning *to operate*. As such frames, "*en* NP" and "*a* NP" are related to different verbs in the constituent approach, it is not possible, for example, to take advantage of that syntactic information to disambiguate meanings. In addition, languages with relaxed word order constraints and a vast prepositional use requires a big quantity of subcategorization frames for each specific verb.

We propose a based on large text corpora statistic method to acquire that individual SF's and a semiautomatic method to correspond syntactic information and semantic valences, which finally are related to the verb meaning.

2 Advanced Government Patterns

2.1. Government patterns in MTT

In the Meaning ⇔ Text Theory [6], 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

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

0-7803-7087-2/01/$10.00 © 2001 IEEE

Sofía N. Galicia-Haro, Alexander Gelbukh, Igor A. Bolshakov. *Acquiring syntactic information for a government pattern dictionary from large text corpora.* IEEE International Workshop on Natural Language Processing and Knowledge Engineering, NLPKE 2001, IEEE, pp. 536–542; www.Gelbukh.com/CV.

dictionary presents a government pattern table 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 influences this order. For example, the expression for Spanish *acusar*: 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 + C2	Juan *acusa* a María.
C.1 + C.2 + C.3.1	Juan *acusa* a María de robo.

Prohibited:

C.1 + C.3.1	*Juan *acusa* de robo.

Though in a complete dictionary, syntactic zone specifies all possible examples after the GP table, in our example only some examples were showed. Obligatory indication was the only consideration for impossible combination examples.

2.2. Some Spanish characteristics

In order to define what additional characteristics must be considered to link syntactic realizations and meaning of a specific verb, in the following subsections some important Spanish characteristics are analyzed.

2.2.1 Animity and direct object

In most of the languages the direct object is connected to the verb directly, without prepositions. In Spanish, quite the opposite, the animated entities are connected to the verb by the preposition *a*, for example *veo a mi vecina* (I see my neighbor) and the no animated entities are connected directly, for example, *veo una casa* (I see a house).

So, animity is a personification in Spanish. For example, government, groups of animals, organizations, politic parties, countries, etc. are animated. In other languages, as Russian, there is an animated category but groups of persons, countries, and cities are not personified in grammatical sense.

Preposition *a* has another use: to differentiate the meaning of some verbs. For example, *querer algo* (to have the desire of something) and *querer a alguien* (to love someone).

Another use of animity could help to distinguish valences when subject is postponed. The verb "to accuse", in Spanish, has two homonyms: *acusar*$_1$ (to denounce somebody as guilty of something) and *acusar*$_2$ (reveal something) according to [2]. Only *acusar*$_2$ has a syntactic realization of direct object as a noun phrase. When subject is postponed, for example, in the following phrase taken from LEXESP:[1]

En el presunto fraude aparece como principal sospechoso José Joaquín Portuondo, a quien acusaron varios testigos. (In the presumed fraud José Joaquín Portuondo appears as a main suspect to whom several witness accuse)

The animity could help to differentiate that *varios testigos* (several witness) is the subject of *acusar*$_1$ and not the direct object of *acusar*$_2$.

2.2.2 Duplicated valences

Generally, the entities referred by the diverse valences are different. This is a normal situation in natural languages; each semantic valence could be represented in the syntactic level by only one actant.

However, there are languages, as Spanish, which permits the duplication of valences. The next examples show, in each sentence, two disjoined groups in bold face that represents the same object:

- ***Arturo le*** *dio la manzana a Víctor.*
- ***El disfraz de Arturo, lo*** *diseñó Víctor.*
- ***A Víctor le*** *acusa el director.*

While the first sentence object duplicate the indirect object, the two next sentences duplicate the direct object. While the repetition in the first sentence is optional it is obligatory in the others.

[1] Kindly provided by Dr. Horacio Rodriguez O. of Universidad Politécnica de Cataluña, Spain.

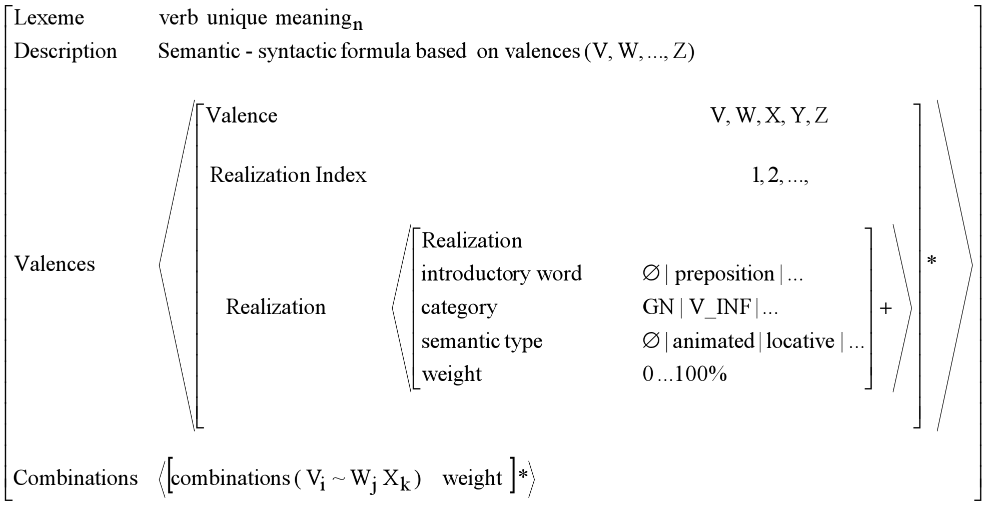

Figure 1. Formal structure for AGP

Sometimes repetition is obligatory. The word order and the specific verbs impose some constructions. For example, the change of dative and accusative arguments before verb presents a complication, the repetition showed in the examples.

2.2.3 Advanced format for government patterns

We propose a new structure that considers the Spanish characteristics mentioned in the previous section besides a modernized format. The new structure called Advanced Government Patterns (AGP) is based on an attribute- value system.

Figure 1 shows the complete description of AGP. The first attribute called *Lexeme* corresponds to the headword. Its value is the numerated headword with a specific: meaning and syntactic realizations. The second attribute called *description* corresponds to the semantic explanation of the situation related to the headword.

The third attribute is the GP table where syntactic valences realizations are described recursively. Each realization could have the following attributes: introducing word, grammatical category, semantic descriptor and weight. The introducing words are mainly prepositions, simple and composed, but they could be also words that introduce subordinated clauses as *que.* The grammatical categories are of any type. Now, the semantic descriptors considered are animated and locative, but they could be of other nature.

The weight defines the filling probability of the diverse valences. This value has immediate application in syntactic analysis and in filters to reject intermediate results impossible or not desired. The obligatory indication is marked with 100% weight.

The last attribute corresponds to the examples for GP. The possible combinations of actants and the prohibited combinations are considered in another way, by probabilistic measure.

3 Statistical Method

The information necessary to fill in these frames is the information of usual subcategorization frames plus syntactic valence information of the verb, semantic features (animity, locative, etc.), and statistics of common use.

For Spanish, there are no dictionaries with complete subcategorization information. There are some attempts for automatic acquisition [8]. To compile a dictionary of AGPs, connecting syntactic, statistical, and semantic knowledge is required.

Taking into account that Spanish has relaxed word order constraints, combinations of complements become a serious problem. A statistical method of disambiguation in text analysis is proposed. The method is based on the statistics of errors that a specific parser used for text analysis makes on specific types of documents. For each phrase, a *weight* (or probability) of each variant is determined. It is based on the statistics of individual subcategorization frames that we called *features,* correctly used in the language and in the wrong variants generated by the specific parser. The variant with the largest weight is considered the best.

One of the most difficult problems of natural language processing is ambiguity, especially syntactic ambiguity. Let us consider a simple English phrase: *John puts the block in the box on the table*,

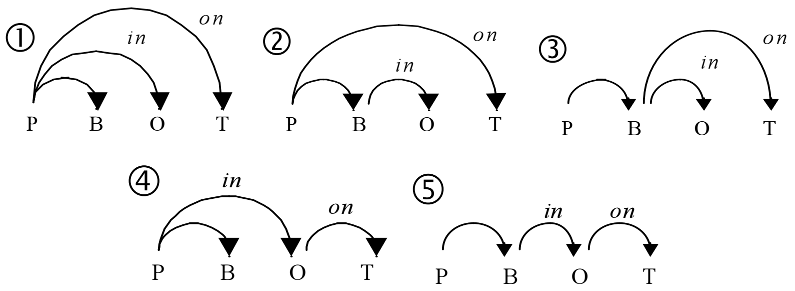

Figure 2. Syntactic structure variants for the phrase[2] *John puts the block in the box on the table.*

which can be represented as *NG V NG P NG P NG*, where *NG* is a noun group or a pronoun, *V* is a verb, *P* is a preposition. Figure 2 shows five possible variants of this phrase's syntactic structure.

A native speaker would choose the structure 4 as the most probable interpretation by taking into account some lexical information. However, it is possible to give examples of phrases with the same basic structure *V NG P NG P NG* in such a way that each of the other variants would be the correct one correspondingly.

Thus, a parser needs similar lexical information to disambiguate the phrase, i.e., to choose for the correct structure from the variants shown on. There are many works devoted to this difficult problem; we can mention [12]. However, the problem is far from having being solved to an acceptable degree. In our opinion, such a choice should be made in quantitative terms: we assign some weight, or a probability, to each variant, the more the weight the more probable variant to be the correct one.

In this work, we deal with the statistics of features that are combinations of individual words with prepositions. Such combinations are similar to so-called subcategorization frames; we apply this notion to words of any part of speech. E.g., the tree number 1 in Figure 2 contains only one non-trivial combination: *put* + *in* + *on*; the tree number 2 contains two such combinations: *put* + *in*, *box* + *on*, etc. The choice of this type of combinations, is not random, they are to a good degree fixed for each specific word, so their statistics are more reliable than that of, say, arbitrary word pairs.

The proposed disambiguation method is based on the frequencies of such combinations both the correct use in phrases of the text (p_i^+) and in the parser errors (p_i^-), i.e., in the trees generated by the parser, but rejected by human proofreaders. What is more, the calculation of the weights should be redone each time a significant modification is made to the parsing algorithm or the grammar.

It is not a problem to calculate such weights provided that there is a marked-up text corpus available, i.e., a corpus where the correct syntactic relationships are marked manually or by some analyzer. However, for a specific subject area, or a specific genre, or a specific language such as Spanish or Russian, or just for a specific set of texts, such information is often absent. The goal of our algorithm is to determine the correct parsing of each phrase of the text corpus, having no such information beforehand.

Thus, our procedure has two interrelated goals: first, to determine the correct structure of each phrase of the corpus, and second, to compile a dictionary for such disambiguation. This dictionary depends on the type of the text and on the specific analyzer. The procedure that we use is iterative. It approaches the two goals in alternating steps: first it estimates the hypotheses on the basis of the current weights in the dictionary, then re-evaluates the weights in the dictionary on the basis of the current weights of the hypotheses of each phrase, and repeats the process.

The process starts with an empty dictionary. In the first iteration, for each phrase, all the hypotheses produced by the parser have equal weights. Then, the frequencies p_i^+ and p_i^- are determined for each combination found at least once in any of the variants produced by the parser for the phrases of the corpus. Since at this stage it is unknown which variants are correct, when determining the number p_i^+ of occurrences of the combination in the correct variants, we sum up the weights w_j of each variant j where the combination was found, since these weights represent the probability for a variant to be the correct one. Similarly, to determine p_i^- we sum up the values $(1 - w_j)$ since this value represents the probability that the given variant is incorrect. We can summarize these expressions as follows:

[2] P = puts, B = block, O = box, T = table

p_i^+/p_i^-	Combination
11.351	acusar,obj:con,obj:de,x:?
11.351	acusar,dobj_suj:?,obj:de,x:?
11.351	acusar,dobj_suj:?,obj:con,obj:de,x
11.351	acusar,dobj_suj:?,dobj_suj:?,obj:de,x
11.351	acusar,dobj_suj:?,dobj_suj:?,obj:con,obj:de,x:?
4.4825	acusar,obj:de,x:?
3.5906	acusar,obj:de,obj:de,x:?
3.0085	acusar,x:?
1.3241	acusar,dobj_suj:?,dobj_suj:?,obj:a,obj:de
1.2941	acusar,dobj_suj:?,dobj_suj:?,obj:a
0.6912	acusar,dobj_suj:?,obj:a,obj:de
0.6206	acusar,dobj_suj:?,obj:a
0.4080	acusar,obj:a,obj:de
0.3788	acusar,obj:a,x:?
0.3788	acusar,obj:?,obj:a,x

Table 1. Statistics.

$$p_i^+ = \sum w_j / S,$$
$$p_i^- = \left[\sum\left(1 - w_j\right) + \lambda\right] / (V - S),$$
$$w_j = C \times \prod\left(p_i^+ / p_i^-\right), \quad \sum w_k = 1,$$

where S is the total number of sentences, V is the total number of variants, i.e., hypotheses, in the corpus. In the first two lines, the addition is done only by the variants where the combination i appears. In the third line, the multiplication is done by all the combinations that appear in the variant j, and the addition is done by all the variants of the structure of one and the same phrase. The divisors in the first two lines, and the constant C in the third line, are introduced only for normalization: S is the total number of correct variants and ($V - S$) incorrect ones. Details of formulae developing were presented in [3], [4].

To solve this system of equations: first the initial two lines are calculated, then the third line. When calculating the expression in the third line, we first ignore the constant C, and then divide all the values w_j by the sum shown in the third line.

We observed that a significant cause of errors were the combinations that appeared in only one or few sentences. Since they never appeared in the wrong variants, their quotient p_i^+/p_i^- reached huge values and became a cause of errors. Artificial addition of some noise in the form of a few fictive "occurrences" of each combination in false variants solved the problem. In the formulae introducing an additional constant λ reflects this addition. We experimented with several values but the best results were obtained with $\lambda \approx 10^{-10}$.

Table 1 shows the results for the verb "*acusar*" produced by the algorithm entering the LEXESP corpus. Despite this results were obtained with a very few sentences, we could compare to that of section 2.1 and only 3 values are wrong, those considering an object introduced by preposition *con* which would be eliminated with more analyzed sentences. The sign "?" indicates a noun phrase. There is no order, so *acusar,dobj_suj:?,obj:?* also represents *acusar, obj:?,dobj_suj:?*

The algorithm is non-supervised and produces a list of prepositions used with each word; at the same time the algorithm resolves the syntactic ambiguity in the corpus.

The technique to acquire such a lexical attraction differs from those known methods for prepositional phrase (PP) attachment [10] because they are addressed to only link patterns of the type V N1 P N2, neglecting multiple PP attachment, or they use texts with syntactic marks [7].

4 Semi-automatic compilation of the dictionary

It is possible to use this raw data for semi-automatic compilation of a classic GP dictionary. We use a dialogue procedure. The algorithm of partitioning of the set of prepositions into actants for one entry works in the following steps:

1. The prepositions are grouped together so that no group contains two prepositions that belong to the same combination. These groups correspond to the hypothetical actants of the word.
2. Of all such possible partitions, the ones resulting in the minimal number of groups are chosen.
3. All the possible orders of the set of the groups are considered, with the restriction that the group containing the direct object must be the second actant. For each one of the orders, a measure is calculated according to the word order in the combinations.
4. The ordered partitioning with the best score is presented to the user. The user can remove some of the prepositions related to circumstances rather than to actants, or move a preposition to another group. After each user action, the calculations are repeated taking into account the restrictions introduced by the user, and an improved version of partitioning is presented to the user.

The process repeats until the user accepts one version or chooses to continue manually. At each stage, the user is presented with the examples, which are also included in the final dictionary.

After the actants have been determined, two kinds of hypotheses are presented to the user:

- The hypotheses on obligatory actants. If some actant is present in all the available examples, the program suggests to mark it as obligatory.
- The hypotheses on incompatibility of actants. Only pairs of prepositions are considered, and if two prepositions belonging to different actants are not found together in examples, the program suggests that they are incompatible.

The program collects examples for the same combination. They are chosen from the text corpus based on a compound criterion: (1) they cover the corpus approximately proportionally and (2) the examples are kept with the best scores. The latter scores order the examples and the best one is the first to be presented to the user, but the user can view all of them and choose the best one, remove some of them, search the corpus for other examples or enter a new one manually.

Of course, the user manually adds the semantic interpretation of the word and the semantic roles corresponding to the syntactic valences.

Conclusions

We have described a method to acquire the syntactic information of government patterns. The advantages of the use of an advanced government pattern (AGP), the required information that AGPs take into account, and the initial steps for acquiring an AGP dictionary from a corpus were discussed.

While usual SFs are defined as a unique representation for all possible complements of verbs, the suggested AGPs are defined for each specific verb to specify its valences and some specific semantic information. The same AGPs can be used for nouns and adjectives that subcategorize for some complements. AGPs provide the information necessary to differentiate the meaning of verbs, to discriminate the valences of the verb, etc. They also help in more accurate parsing and introduce relations of the syntactic valences to the semantic valences of the verb.

The method that we propose has the following characteristics:

- No hand preparation is required to train the model. However, morphological and syntactic analyzers are required since the method is devoted to disambiguate the results of such parsing.
- The method is tuned to a specific parser, taking into account the balance between the correct and wrong assignments of prepositions to words.

The results obtained applying our method on LEXESP corpus were used on parsing a 100-sentence test corpus and the true structures for the input sentences proved to be classified in the first rank ca. 35%.

References

1. Briscoe, E. & Carroll, J. *Automatic extraction of subcategorization from corpora*. In Proceedings of the 5th ACL Conference on Applied Natural Language Processing. Washington, DC. 1997
2. DEUM *Diccionario del Español Usual en México*. Edit. Colegio de México. México, 1996.
3. Gelbukh, A.F., I.A. Bolshakov, and S.N. Galicia-Haro. *Automatic* learning *of a syntactical* government *patterns dictionary from Web-retrieved texts*. Proc. of CONALD-98, Pittsburgh, PA, 1998.
4. Gelbukh, A.F., I.A. Bolshakov, and S.N. Galicia-Haro. *Syntactic disambiguation by learning weighted government patterns from a large corpus*. Technical report, CIC, IPN, Mexico, 1998.
5. Manning, C. *Automatic acquisition of a large subcategorisation dictionary from corpora.* In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus, Ohio. 1993.
6. Mel'cuk, I. A. *Dependency Syntax: Theory and Practice*. State University of New York Press. Albany. 1988.
7. Merlo, P., Crocker, M. and Berthouzoz, C. *Attaching Multiple Prepositional Phrases: Generalized Backed-off Estimation*. In Proceedings of the EMNLP-2, 1997. http://xxx.lanl.gov/find/ cmp-lg/9710005
8. Monedero, J. et al. *Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS.* En Actas del XI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 95: Bilbao), págs. 241-254. 1995.
9. Polguère, A. *Observatory of Meaning-Text Linguistics* (OMTL). Université de Montréal. Faculté des Arts et des Sciences. 1998. http://www.fas.umontreal.ca/ling/olst/indexE.html
10. Ratnaparkhi, A. *Statistical Models for Unsupervised Prepositional Phrase Attachment.* In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Montreal, Quebec, Canada. 1998.
11. Steele, J. 1990. *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