Automatic Enrichment of a Very Large Dictionary of Word Combinations on the Basis of Dependency Formalism*

Alexander Gelbukh[1,2], Grigori Sidorov[1], San-Yong Han[2], and Erika Hernández-Rubio[1]

[1] Natural Language and Text Processing Laboratory, Center for Computing Research, National Polytechnic Institute, Av. Juan Dios Batiz s/n, Zacatenco 07738, Mexico City, Mexico
{gelbukh,sidorov}@cic.ipn.mx, www.gelbukh.com
[2] Department of Computer Science and Engineering, Chung-Ang University, 221 Huksuk-Dong, DongJak-Ku, Seoul, 156-756, Korea
hansy@cau.ac.kr

Abstract. The paper presents a method of automatic enrichment of a very large dictionary of word combinations. The method is based on results of automatic syntactic analysis (parsing) of sentences. The dependency formalism is used for representation of syntactic trees that allows for easier treatment of information about syntactic compatibility. Evaluation of the method is presented for the Spanish language based on comparison of the automatically generated results with manually marked word combinations.

Keywords: Collocations, parsing, dependency grammar, Spanish.

1 Introduction

There is a growing demand for linguistic resources in modern linguistics and especially in natural language processing. One of the important types of resources is the dictionary that reflects mutual combination of words.

The problem concerning the types of information about compatibility of words that should be stored in the dictionary has a rather long history – first papers appeared in 50s. Basically, the focus of attention of the researchers was the concept of collocation and its usage. The main research direction was integration of this concept into lexicographical practice and methods of teaching of foreign languages – how many examples should the dictionaries or textbooks contain, are the examples of collocations just examples of usage or essential part of knowledge of a language (language competence).

* Work was done under partial support of Mexican Government (CONACyT, SNI, CGPI-IPN, PIFI-IPN), Korean Government (KIPA Professorship for Visiting Faculty Positions in Korea), and ITRI of CAU. The first author is currently on Sabbatical leave at Chung-Ang University. We thank Prof. Igor A. Bolshakov for useful discussions.

R. Monroy et al. (Eds.): MICAI 2004, LNAI 2972, pp. 430–437, 2004.

After many discussions, the common point is that it is very difficult to find a concise and formal definition of collocation. Nevertheless, it seems that the majority of investigators agree that collocations are very important part of knowledge of language and they are useful for different tasks of automatic natural language processing like automatic translation, text generation, intelligent information retrieval, etc. All this causes the necessity of compilation of specialized dictionaries of collocations and even of free word combinations. See next section for more detailed discussion of the concept of collocations.

There exist many methods of extracting collocations that are based on the analysis of a large corpus [1, 3, 8, 10, 11, 18]. Still, the majority of them are oriented to searching of highly repetitive combinations of words based on measuring of their mutual information. These methods do not guarantee finding the collocations if they do not have sufficiently high frequency. Usually, the great number of collocations does not have this frequency. Besides, the corpus size for such search should be larger than the existing corpora (now measured in gigabytes).

There are several attempts to apply the results of automatic syntactic analysis (parsing) for compilation of dictionaries of collocations [3, 7]. For example, in a recent work Strzalkowski [17] uses the syntactic analysis for improving results of information retrieval by enriching the query. One of the classic works on the theme is [15]. The system *Xtract* is presented that allows for finding repeated co-occurrences of words based on their mutual information. The work consists in three stages, and, at the third stage, the partial syntactic analysis is used for filtering out the pairs that do not have a syntactic relation. Unfortunately, all these methods are applied to word pairs obtained by frequency analysis using a threshold. The aim of all these methods is collocations, and not free word combinations (see below). As far as the *Xtract* system is concerned, it is reported rather high precision (80%) and recall (94%), nevertheless, the evaluation was done by comparison of results with opinion of only one lexicographer, and what is collocation according to the system remains unclear, obviously, they did not process free word combinations.

There are already some resources of the described type available. One of the largest dictionaries of collocations and free word combinations is *CrossLexica* system [4, 5, 6]. It contains about 750,000 word combinations for Russian with semantic relations between the words and the possibilities of inference. There is also this type of resources for the English language, e.g., Oxford dictionary of collocations [14] (170,000 word combinations) or Collins dictionary [2] (140,000 word combinations), though they do not contain semantic relations. This is the lower bound of the dictionary of word combinations, which justifies the term *very large dictionary* in the title of this paper.

In the rest of the paper, we first discuss the concept of collocation and its relation with free word combination, and then we describe the method of enrichment of the dictionary based on automatic syntactic analysis with dependency representation. After this, we evaluate its performance, and finally draw some conclusions.

2 Idioms, Collocations, and Free Word Combinations

Now let us discuss the concept of collocation in more detail. Intuitively, collocation is a combination of words that has certain tendency to be used together. Still, the

strength of this tendency is different for different combinations. Thus, collocations can be thought of as a scale with different grades of strength of the inter-word relation, from idioms to free word combinations.

On the one side of the scale there are complete idioms like "*to kick the bucket*", where neither the word "*to kick*", nor "*the bucket*" can be replaced without destroying the meaning of the word combination. In this case, the meaning of the whole is not related with the meaning of the components. In certain much more rare cases, the meaning of the whole has the relation with the meaning of the components, but also it has an additional part that cannot be inferred, e.g., "*to give the breast*" that means "*to feed a baby using breast*". Though the physical situation is described correctly by using the words "*to give*" and "*the breast*", the meaning of "*feeding*" is obtained from the general knowledge about the world. This case can be considered as a little shift on the scale towards free word combinations, nevertheless, this type of combinations are still idioms ("nearly-idiom" according to Mel'chuk [13]).

On the other side of the scale there are free combinations of words, like "*to see a book*", where any word of the pair can be substituted by a rather large class of words and the meaning of the whole is the sum of the meanings of the constituent words.

Somewhere in the middle on this scale, there are lexical functions[1] [13] like "*to pay attention*". In this case, the meaning of the whole is directly related only with one word (in the example above, the word *attention*), while the other word expresses a certain standard semantic relation between actants of the situation. The same relation is found, for example, in combinations like "*to be on strike*", "*to let out a cry*", etc. Usually, for a given semantic relation and for a given word that should conserve its meaning, there is a unique way to choose the word for expressing the relation in a given language. For example, in English it is *to pay*, while in Spanish it is *prestar atención* (lit. *to borrow attention*), in Russian – *obratit' vnimanije* (lit. *to turn attention to*), etc.

As far as free word combinations are concerned, it can be seen that some free word combinations are "less free" than the others, though they are still free word combinations in a sense that the meaning of the whole is sum of meanings of the constituent words. The degree of freedom depends on how many words can be used as substitutes of each word. The less is the number of substitutes, the more "idiomatic" is the word combination, though these combinations will never reach neither idioms nor lexical functions where the meaning cannot be summed.

It is obvious that the restrictions of freedom in free word combinations are the semantic constraints, for example, "*to see a book*" is less idiomatic than "*to read a book*" because there are much more words that can substitute *a book* combining with the verb *to see*, than with the verb *to read*. Namely, practically any physical object can be *seen*, while only objects that contained some written information (or its metaphoric extension, like "*to read signs of anger in his face*") can be *read*.

Another important point is that some free word combinations can have associative relations between its members, e.g., *a rabbit* can *hop*, and *a flea* also, but *a wolf* usually does not *hop*, though potentially it can move in this manner. This makes some combinations more idiomatic because the inter-word relation is strengthened by association.

---

[1] Lexical functions were discovered by Mel'čuk in 70s. Unfortunately, till now they are not reflected systematically even in good dictionaries. This is because the work of finding these functions is rather laborious and it needs very high-level lexicographic competence.

In a strict sense, only lexical functions are collocations, but the common treatment of this concept also expands it to free word combinations that are "more idiomatic". Since there is no obvious border between more idiomatic and less idiomatic, the concept of collocation finally can cover all free word combinations as well, though this makes this concept useless because its purpose is to distinguished idiomatic word combinations from the free ones. Thus, in our opinion, the difficulties related with the concept of collocations are related with impossibility to draw the exact border between it and free word combinations.

Note that the obvious solution to treat collocations only as lexical functions contradicts to the common practice. This demonstrates that, in any case, we need something to distinguish between more and less idiomatic free word combinations. If it is not collocation, then the other term should be invented.

3 Automatic Enrichment of the Dictionary

Traditionally, free word combinations are considered of no interest to linguistics, though, in fact, practically any free word combination is "idiomatic" to a certain grade, because the majority of them have certain semantic restrictions on compatibility. In our opinion, it is so, because, according to the famous Firth idea "you shall know the word by the company it keeps", any word combination is important. For example, in automatic translation, some wrong hypothesis can be eliminated using the context [16]; in language learning, the possibility to know the compatibility allows for much better comprehension of a word; not speaking about automatic word sense disambiguation, where one of the leading approaches is analysis of the context for searching of the compatible words, etc.

Note that manual compilation or enrichment of the dictionary of free word combinations is very time-consuming, for example, *CrossLexica* [5] was being complied during more than 13 years and it is very far from completion yet.

We suggest the following method of automatic enrichment of such kind of dictionaries. Obviously, the method needs some post-verification, because we cannot guarantee the total correctness of the automatic syntactic analysis, still, it is much more efficient than to do it manually.

We work with the Spanish language, but the method is easily applicable for any other language depending on the availability of a grammar and a parser. First, we apply the automatic syntactic analysis using the parser and the grammar of Spanish developed in our laboratory [9]. The results of the syntactic analysis are represented using the formalism of dependencies [12]. It is well known that the expressive power of this formalism is equal to the formalism of constituents, that is much more commonly used, but the procedure of treatment of word combinations is much more easy using dependencies.

The idea of the formalism of dependencies is that any word has dependency relations with the other words in a sentence. The relations are associated directly with word pairs, so it is not necessary to pass the constituency tree in order to obtain the relation. One word always is a head of relation, and the other one is its dependant. Obviously, one headword can have several dependencies.

The problems that are to be solved even using this formalism are the treatment of coordination conjunctions and prepositions, and filtering of some types of relations and some types of nodes (pronouns, articles, etc.).

We store the obtained combinations in the database. All members of the pairs are normalized. Still, some information about the form of the dependant is saved also. In our case, for nouns, we save the information about its number (singular or plural), say, "*play game Sg*" and "*play game Pl*" (the word combination in both cases is "*play game*", and it has an additional mark); for verbs, the information if it is a gerund or a participle is important, etc.

The coordinative conjunctions are heads in the coordinative relation; still, the word combinations that should be added to the dictionary are the combinations with their dependants. For example, *I read a book and a letter*, the combinations that should be extracted are *read book* and *read letter*. Thus, the algorithm detects this situation and generates two virtual combinations that are added to the dictionary.

Treatment of prepositional relation is different from other relations. Since the prepositions usually express grammar relations between words (for example, in other languages these relations can be expressed by grammar cases), the important relation is not relation with the preposition, but the relations between two lexical units connected by the preposition. Still, the preposition itself is also of linguistic interest, so we reflect this relation in the dictionary by the word combination that contains three members: the headword of the preposition, the preposition, and its dependant, e.g., *He plays with a child* gives the combination *play with child*.

Filtering of determined types of nodes is very easy. Since the parser uses the automatic morphological analysis, the morphological information for every word is available. It allows for filtering out the combinations without significant lexical contents, i.e., if at least one word in the combinations belongs to one of the following categories with mainly grammatical meaning. The following categories are discarded in the actual version of the algorithm: pronouns (personal, demonstrative, etc.), articles, subordinate conjunctions, negation (*not*), and numerals. Since the combinations with these words have no lexical meaning, they have no semantic restrictions on compatibility, and can be considered as "absolutely" free word combinations. These combinations are of no interest for the dictionary under consideration.

The other filter is for the types of relations. It depends on the grammar that is used. In our grammar, the following relations are present: *dobj* (direct object), *subj* (subject), *obj* (indirect object), *det* (determinative), *adver* (adverbial), *cir* (circumstantial), *prep* (prepositional), *mod* (modifying), *subord* (subordinate), *coord* (coordinative). Among these relations, the prepositional and coordinative are treated in a special mode, as mentioned above. The only relations left that are of no use for detecting of word combinations are subordinate relation and circumstantial relation.

One of the advantages of the suggested method is that it does not need corpus for its functioning, and, thus, there is no dependency of the corpus size or corpus lexical structure.

Let us have a look at the example of the functioning of the method. The following sentence is automatically parsed.

Conocía todos los recovecos del río y sus misterios.
(*I knew all detours of the river and its mysteries*.)

The following dependency tree corresponds to this sentence. The hierarchy of depth in the tree corresponds to the relations (number of spaces at the beginning of each line[2]). For example, V(SG,1PRS,MEAN) [*conocía*] is head of the sentence and its dependants are CONJ_C [*y*] and $PERIOD. CONJ_C has dependants N(PL,MASC) [*recovecos*] and N(PL,MASC) [*misterios*], etc. Each line corresponds to a word and contains the word form and its lemma, e.g., *conocía : conocer* (*knew : know*), etc.

```
V(SG,1PRS,MEAN) ->  () // Conocía : conocer (knew : know)
   CONJ_C ->  (obj)  // y : y (and : and)
      N(PL,MASC) ->  () // recovecos : recoveco (detours : detour)
         PR ->  (prep) // del : del  (of the : of the)
            N(SG,MASC) ->  (prep) // río : río (river : river)
         ART(PL,MASC) ->  (det) // los : el  (the : the)
         #*$$todo# ->  () <*$$todo> // todos : todo (all : all)
      N(PL,MASC) ->  (coord_conj) // misterios : misterio (mysteries : mystery)
         DET(PL,MASC) ->  (det) // sus : su (its : it)
   $PERIOD ->  () // . : .
```

The following word combinations were detected:

conocer (obj) recoveco (*to know detour*)
conocer (obj) misterio (*to know mystery*)
recoveco (prep) [del] río (*detour of the river*)

It can be seen that the relation (*obj*) corresponds to coordinative conjunction, and then it is propagated to its dependants: *recoveco* (*detour*) and *misterio* (*mystery*). The preposition *del* is part of the 3-member word combination. The articles and pronouns are filtered out (*el, todo, su*), though the algorithm found the corresponding word combinations.

4 Evaluation

We conducted the experiments on the randomly chosen text in Spanish from Cervantes Digital Library. Totally 60 sentences were parsed that contain 741 words, average 12.4 words per sentence. For evaluation, we manually marked all dependency relations in the sentences. Then we compared the automatically added word combinations with manually marked word combinations.

Apart, we used as a baseline a method of gathering the word combinations that takes all word pairs that are immediate neighbors. Also we added certain intelligence to this baseline method – it ignores the articles and takes into account the prepositions. Totally, there are 153 articles and prepositions in the sentences, so the number of words for baseline method is 741-153 = 588.

The following results were obtained. The total number of correct manually marked word combinations is 208. From these, 148 word combinations were found by our

[2] Usually, the arrows are used to show the dependencies between words, but it is uncomfortable to work with arrows in text files, so we use this method of representation. Besides, this representation is much more similar to constituency formalism.

method. At the same time, the baseline method found correctly 111 word combinations. On the other hand, our method found only 63 incorrect word combinations, while the baseline method marked as a word combination 588*2 – 1 = 1175 pairs, from which 1175 – 111 = 1064 are wrong pairs.

These numbers give us the following values of precision and recall. Let us remind that precision is the relation of the correctly found to totally found, while the recall is the relation of the correctly found to the total that should have been found. For our method, precision is 148 / (148+63) = 0.70 and recall is 148 / 208 = 0.71. For the baseline method, precision is 111 / 1175 = 0.09 and recall is 111 / 208 = 0.53. It is obvious that precision of our method is much better and recall is better than these parameters of the baseline method.

5 Conclusions

A dictionary of free word combinations is very important linguistic resource. Still, compiling and enriching of this dictionary manually is too time and effort consuming task. We proposed a method that allow for enrichment of such dictionary semi-automatically. The method is based on parsing using the dependency formalism and further extraction of word combinations. Some types of relations and some types of nodes are filtered because they do not represent substantial lexical information. Special processing of coordinative and prepositional relations is performed. The method requires post-processing of obtained word combinations, but only for verification that no parser errors are present.

The results are evaluated on a randomly chosen text in Spanish. Proposed method has much higher precision and better recall than the baseline method.

References

1. Baddorf, D. S. and M. W. Evens. Finding phrases rather than discovering collocations: Searching corpora for dictionary phrases. In: *Proc. of the 9th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS'98)*, Dayton, USA, 1998.
2. Bank of English. Collins. http://titania.cobuild.collins.co.uk/boe_info.html
3. Basili, R., M. T. Pazienza, and P. Velardi. Semi-automatic extraction of linguistic information for syntactic disambiguation. *Applied Artificial Intelligence*, 7:339-64, 1993.
4. Bolshakov, I. A. Multifunction thesaurus for Russian word processing. In: *Proceedings of 4th Conference on Applied Natural language Processing*, Stuttgart, October 13-15, 1994, p. 200-202.
5. Bolshakov, I. A., A. Gelbukh. A Very Large Database of Collocations and Semantic Links. In: Mokrane et al. (Eds.) *Natural Language Processing and Information Systems, 5th International Conference on Natural Language Applications to Information Systems NLDB-2000*, Versailles, France, June 2000. Lecture Notes in Computer Science No. 1959, Springer Verlag, 2001, p. 103-114.
6. Bolshakov, I. A., A. Gelbukh. Word Combinations as an Important Part of Modern Electronic Dictionaries. *Revista SEPLN (Sociedad Español para el Procesamiento del Lenguaje Natural)*, No. 29, septiembre 2002, p. 47-54.

7. Church, K., W. Gale, P. Hanks, and D. Hindle. Parsing, word associations and typical predicate-argument relations. In: M. Tomita (Ed.), *Current Issues in Parsing Technology*. Kluwer Academic, Dordrecht, Netherlands, 1991.
8. Dagan, I., L. Lee, and F. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1), 1999.
9. Gelbukh, A., G. Sidorov, S. Galicia Haro, I. Bolshakov. Environment for Development of a Natural Language Syntactic Analyzer. In: Acta Academia 2002, Moldova, 2002, pp.206-213.
10. Kim, S., J. Yoon, and M. Song. Automatic extraction of collocations from Korean text. *Computers and the Humanities* 35 (3): 273-297, August 2001, Kluwer Academic Publishers.
11. Kita, K., Y. Kato, T. Omoto, and Y. Yano. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing*, 1(1):21-33, 1994.
12. Mel'čuk, I. *Dependency syntax*. New York Press, Albany, 1988, 428 p.
13. Mel'čuk, I.Phrasemes in language and phraseology in linguistics. In: *Idioms: structural and psychological perspective*, pp. 167-232.
14. *Oxford collocation dictionary*, Oxford, 2003.
15. Smadja, F. Retrieving collocations from texts: Xtract. *Computational linguistics*, 19 (1):143-177, March 1993.
16. Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1-38, 1996.
17. Strzalkowski, T. Evaluating natural language processing techniques in information retrieval. In: T. Strzalkowski (ed.) Natural language information retrieval. Kluwer, 1999, pp. 113-146.
18. Yu, J., Zh. Jin, and Zh. Wen. Automatic extraction of collocations. 2003.