Automatic Syntactic Analysis for Detection of Word Combinations*

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

[1] 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 for automatic detection of "non-trivial" word combinations in the text. It is based on automatic syntactic analysis. The method shows better precision and recall than the baseline method (bigrams). It was tested on a text in Spanish. The method can be used for enrichment of very large dictionaries of word combinations.

1 Introduction

The concept of word combination is related to the possibility of different words to appear together in the text connected by a syntactic link. The task is not computationally trivial because syntactically connected words can be linearly far from each other, i.e., separated by other words.

There are different types of word combinations. Some word combinations are fixed, like idioms, e.g., *to kick the bucket* or lexical functions like *to pay attention* [14]. In case of idioms and lexical functions, the meaning of the whole cannot be deduced from the meaning of the constituent words. In idioms, usually all words loose theirs meanings. As far as lexical functions are concerned, only one word (in case of our example, *attention*) keeps its meaning, while the other word (*to pay*) expresses standard semantic relation between actants of the situation. Detailed description of lexical functions can be found, for example, in [14] or other works by Mel'čuk. Since the meaning of the combinations is not a sum of the meanings of the words, there are severe restrictions for compatibility in lexical functions. Namely, if we want to express the given meaning and the words that conserves its meaning is known, then usually the choice of the other word is predetermined.

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

+ Corresponding author.