An Experiment in Detection and Correction of Malapropisms through the Web[*]

Igor A. Bolshakov

Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico
igor@cic.ipn.mx

Abstract. Malapropism is a type of semantic errors. It replaces one content word by another content word similar in sound but semantically incompatible with the context and thus destructing text cohesion. We propose to signal a malapropism when a pair of syntactically linked content words in a text exhibits the value of a specially defined Semantic Compatibility Index (SCI) lower than a predetermined threshold. SCI is computed through the web statistics of occurrences of words got together and apart. A malapropism detected, all possible candidates for correction of both words are taken from precompiled dictionaries of paronyms, i.e. words distant a letter or a few prefixes or suffixes from one another. Heuristic rules are proposed to retain only a few highly SCI-ranked candidates for the user's decision. The experiment on malapropism detection and correction is done for a hundred Russian text fragments—mainly from the web newswire—in both correct and falsified form, as well as for several hundreds of correction candidates. The raw statistics of occurrences is taken from the web searcher Yandex. Within certain limitations, the experiment gave very promising results.

1 Introduction

One of the most important applications of computers is automated checking of text accuracy. The problem of out-of-context orthographic correction of letter strings not existing in the language is practically solved.

Syntactic errors leave correct all separate words but violate the sentence structure by using words with wrong morpho-syntactic features (POS, number, gender, case, person, etc.) or violating habitual word order. There is some advance in grammar checkers, though they still need more powerful and robust parsers.

In raw texts, semantic errors also occur. They are of various types and usually violate neither orthography nor grammar. Being expressed by correct words inappropri-

[*] Work done under partial support of Mexican Government (CONACyT, SNI) and CGEPI-IPN, Mexico. Many thanks to Denis Filatov for help with manuscript preparation.