Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power*

Igor A. Bolshakov[1], Elena I. Bolshakova[2], Alexey P. Kotlyarov[1], and Alexander Gelbukh[2]

[1]Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico
{igor,gelbukh}@cic.ipn.mx
[2]Moscow State Lomonosov University
Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
bolsh@cs.msu.su, koterpillar@gmail.com

Abstract. For extracting collocations from the Internet, it is necessary to numerically estimate the cohesion between potential collocates. Mutual Information cohesion measure (MI) based on numbers of collocate occurring closely together (N_{12}) and apart (N_1, N_2) is well known, but the Web page statistics deprives MI of its statistical validity. We propose a family of different measures that depend on N_1, N_2 and N_{12} in a similar monotonic way and possess the scalability feature of MI. We apply the new criteria for a collection of N_1, N_2, and N_{12} obtained from AltaVista for links between a few tens of English nouns and several hundreds of their modifiers taken from Oxford Collocations Dictionary. The nounits own adjective pairs are true collocations and their measure values form one distribution. The nounalien adjective pairs are false collocations and their measure values form another distribution. The discriminating threshold is searched for to minimize the sum of probabilities for errors of two possible types. The resolving power of a criterion is equal to the minimum of the sum. The best criterion delivering minimum minimorum is found.

1 Introduction

During the two recent decades, the vital role of collocationsin any their definitionwas fully acknowledged in NLP. Thus great effort was made to develop methods of collocation extraction from texts and text corpora. As pilot works we can mention [3, 6, 17, 18]. However, up to date we have no large and humanly verified collocation databases for any language, including English. The only good exception is Oxford Collocations Dictionary for Students of English (OCDSE) [11], but even in its electronic version it is oriented to human use rather than to NLP. So the development of the methods of collocation extraction continues [4, 5, 9, 12–16, 19].

* Work done under partial support of Mexican Government (CONACyT, SNI, CGEPI-IPN) and Russian Foundation of Fundamental Research (grant 06-01-00571).