ThaiBinh Luong
CBB545 – Data Mining
Final project description
3/27/07

In the field of text mining, current natural language processing algorithms generally divide a testing sentence into words, each tagged with a part-of-speech (e.g. noun, verb). To test such system requires a “gold standard” that was already human-tagged. Obviously, this can be very time-consuming. A different type of algorithm, “ADIOS” (automatic distillation of structure)[1], creates grammar rules based on a corpus of sentences, in which assigning parts-of-speech is not required. Instead, words are mapped into a type of tree where nodes represent categories of words. To calculate the precision of the algorithm, the system creates sentences based on the grammar it created, and the user indicates whether the generated sentence is indeed a valid sentence.

After recreating their experiment on the established set of corpuses they mention in the paper, I would like to use this system in the context of information retrieval in biomedical literature. I will input a corpus of sentences that each contain two or more genes/proteins, and see if ADIOS can generate a grammar that correctly represents how one gene/protein affects another gene/protein(s). Further, I would have ADIOS create a grammar based only on sentences that co-mention one specific gene and another gene. In theory, the grammar should generate sentences that correctly make inferences about the specific gene.

[1] Solan, Z., D. Horn, E. Ruppin, and S. Edelman. Unsupervised learning of natural languages. *PNAS, August 16, 2005: 11629-11634.*