論文の内容の要旨

論文題目　Mining Literature for Disease-Gene Relations
（遺伝子–疾患関係概念の文献からのマイニング）

氏　名　全 弘 宇

Background: Automatic extraction of relations between a specific disease name and its relevant gene or protein names is an important practice of bioinformatics. Considering the utility of the results of this approach, we identified disease and gene names with the ID tags of public biomedical databases. Moreover, considering that genetics experts will use our results, we classified them based on topics that can be used to analyze the type of disease-gene relations.

Methods: We developed a Maximum Entropy Markov Model (MEMM)-based disease and gene name recognizer, a relation extractor and a topic-classified relation extractor applied them to a corpus-based approach. We collected corpus from MEDLINE with respect to prostate cancer and gastric cancer, and constructed an annotated corpus of disease and gene relations based on two topics: etiology and clinical marker. To recognize disease and gene names and extract any relations between them, we used rich information that was obtained from an analysis of syntactic structures of the input data. Moreover, to extract relations based on the topics, we collected various features considering aliases, synonyms, acronyms and full names of candidate disease and gene names that were obtained from abstracts (vocabulary and context extension). We used them to train the Maximum Entropy Markov Model (MEMM)-based disease and gene name recognizer, relation extractor, and topic-classified relation extractor.

Results: Topic-classified relation extraction achieved encouraging results. For the relations between prostate cancers and genes, the performance of relation extraction based on etiology obtained 77.6% F-measure (increases of 75.0, 74.4, and 2.2% from that obtained in experiments using the dictionary matching, disease and gene name filtering, and relation filtering methods, respectively. Each method used all the previous methods. In other words, the relation filtering method used the results of the dictionary matching and disease and gene name filtering methods.) and that based on clinical marker obtained 77.0% F-measure (increases of 29.1, 24.5, and 7.3% from that obtained in experiments using the dictionary matching, disease and gene name filtering, and relation filtering methods, respectively.).

For the relations between gastric cancers and genes, the performance of relation extraction based on etiology obtained 74.0% F-measure (increases of 46.0, 43.2, and 8.0% from that obtained in experiments using the dictionary matching, disease and gene name filtering, and relation filtering methods, respectively.) and that based on clinical marker obtained 65.1% F-measure (increases of 51.8, 49.6, and 9.8% from that obtained in experiments using the dictionary matching, disease and gene name filtering, and relation filtering methods, respectively.).

Conclusions: A series of experimental results revealed three important findings: (1) A carefully designed named entity filtering and relation filtering methods can improve the performance of topic-classified relation extraction. (2) Features that were obtained by extension of context and vocabulary improved the performance of topic-classified relation extraction, and (3) The Maximum Entropy Markov Model-based topic-classified relation extraction approach achieved the encouraging results for both prostate cancer- and gastric cancer-related instance sentences.