CHAPTER 5

CONCLUSION

This thesis study focus on improving the field of information extraction by using text mining, natural language processing and graphical models to capture functional genomics named entities and extract biological knowledge via machine learning and statistical methods. In the preliminaries of the research, the characteristics of the biological literature and the problems in biological information extraction which relate to functional genomics were studied to place our research objectives in context. The scopes of thesis and data sources were designed in **Chapter1**.

Information extraction techniques which are related to the goals of this thesis were surveyed, in **Chapter 2**. The biological text mining, the problems, motivations and state of the challenge task were revised. After that we studied the effectiveness of general terms in biological literature. According to information extraction is the subtask of Information retrieval. General terms form biological literature was studied with Latent Semantic Indexing (LSI) to improve information retrieval performance but the general terms were still not represented each class as well. Then we focused on the feature selection models (information gain (IG), mutual information (MI) and Odd ratio) was trained to generate the model prediction based on three of the most important and effective machine learning algorithms that are often used in text classification; *k-nearest* neighbors, Naive Bayes, and Support Vector Machine (SVMs) were reviewed. The test data was taken from the NCBI corpus which is a collection of bioinformatics related

scientific papers. The results showed that the SVMs model with odds ratios returned the highest accuracy prediction at 75%. The main errors came from mistake in the prediction of the word boundaries. We can conclude that biological texts characteristics are different from others. General terms still could not well represent texts. Feature selection methods based on computational approaches were not enough for biological text in information extraction. These resulted lead to the sequence labeling problem. It was considered as part of the biomedical named entity recognition models. The models helped to locate the boundary of biological terms which was the critical steps for information extraction from the literature data.

After surveyed the biological information extraction techniques based on machine learning in **Chapter 2**, the error results were leaded to the new topics which we called biological terms recognition in the next chapter.

In **Chapter 3**, many techniques for biological named entity recognition; Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM) and Conditional Random Fields (CRFs) were referenced. The contents of each model described how to train the data set and how to predict the classes of named entities. The assumption of this thesis is compatible with the basic assumptions of the CRFs which is a performance labeling model. Its feature functions were comprised of the state and transition feature functions which enabled the CRFs to capture the dependency between each word in the biological terms. The first contribution of this thesis is the work which designed to improve suitable potential functions for named entity recognition problem. The word surfaces and labels were considered to be a feature selection which had a focus to define

the potential function and the performance of the models prediction. The potential function was defined as 1st order and 2nd potential function in the CRFs. The results showed that the prediction models which were generated from the 2nd potential function CRFs returned a higher recall value at 99.73%, precision value at 99.88% and F-measure at 99.81% than the recall, precision value and F-measure of the 1st potential function which were 56.18%, 69.50 % and 62.13%, respectively. To validate the performance of CRFs, we extended the experiment in chapter 2, biological information retrieval using LSI and CRFs. From the experimental results, the information retrieval performance increased from 65.98% to 75.98%. We can conclude that the LSI methods help to find latent concepts in whole documents, not only important keywords. Features selection based on CRFs can find the exactly the names entities from literature. The last of this chapter, the model was compared with another techniques and state of the art which our model returned the higher performance than the others.

After improving the named entities recognition models, another contribution is knowledge extraction. In **Chapter 4**, relation extraction was studied as lexical and semantics. For lexical-relation extraction from biological literature, we found that the occurrence of terms together in text can serve as an indication of a relationship between them. For example, the co-occurrence of two protein names within context can suggest an interaction between the proteins. It can consider as co-occurrence of named entities. This assumption of this chapter was assumed that biological terms which relate each other always occur together in the same documents. We applied collocations techniques which is the one of Natural Language Processing (NLP) process for co-occurrence problems.

The biological terms collocation can be found by applying the Poisson distribution. A Poisson collocation was considered in this subject. The significance values based on the Poisson indicated that Possion collocations could help to detect the co-occurrence of biological terms. We tested the performance with other statistical methods for collocation and discussed why Possion preferred suitable than while the semantic relation extraction, NLP and bio-molecular event was studied. Part of speech (POS) was used to find relations of named entities which occurred together. It's leads to structure prediction not sequence as the last chapter. Tree structure was applied with POS to extract relations. In the validation step, the extracted relations were shown to correspond to known existing biological knowledge for example Gene Ontology (GO). But it's still manual validation from others sources which made this difficult and some relations was not corresponding with biological semantics.

According to results above, to focus on biological semantics, bio-molecular event topic was studied. The related works and state of the art were reviewed first then this work was designed into two parts. The first step was to transform input data into a tree structure data based on Context Free Grammar (CFG) while the other step was the suitable predicate arguments based on biological semantics for classification were selected and generated models with Tree kernels Support Vector Machine. The performances of the models' results were 77.48% for the F-measure scale, 78.18% for precision and 76.79% for the recall values. We can conclude that POS could help to extract pre semantic-relation extraction from biological literature but still not cover all biological semantics while semantic-relation extraction based on Tree Kernels method,

defined the suitable predicate argument for feature selection could improve the model performance. The last chapter was the aggregation of the functional genomics knowledge base from each phase. Finally, the knowledge base extraction was depicted in the semantics graph network. The last contribution, the various biological data sources now are increasing. That's difficult to collaborate for specific knowledge base. That's motivation of this thesis to serve the researcher who wants to establish functional genomics knowledge base with machine learning.