CHAPTER 1

INTRODUCTION

Functional genomics is the field of molecular biology that attempts to make use of the vast wealth of data produced by genomic projects to describe gene and protein function and interactions. Unlike genomics and proteomics, functional genomics focuses on dynamic aspects such as gene transcription, translation, and protein-protein interactions, as opposed to the static aspects of the genomic information such as DNA sequence or structure. Functional genomics includes function related aspects of the genome itself such as mutation and polymorphism (such as SNP) analysis, as well as the measurement of molecular activities such as transcriptomics, proteomics and metabolomics. These measurement modalities quantify the various biological processes and power the understanding of gene, protein functions and protein-protein interactions.

The literature related to functional genomics that is available on the web has experienced unprecedented growth in recent years. The demand for efficient processing of these documents is also increasing rapidly. It has become difficult to locate, and report all relevant information [1, 2, and 3]. The MEDLINE corpus, which is the largest component of [1]PubMed, is a freely accessible online database of functional genomics journals, citations and abstracts created by the U.S. National Library of Medicine (NLM®).

[1]http://pubmed.gov

Approximately 5,200 journals published in the United States and more than 80 other countries have been selected and indexed. MEDLINE contains abstracts with a growth rate of about 400,000 articles per year. Hence, in the extraction for establishing the specific functional genomics knowledge base parameters from the literature or source texts still needs various techniques such as text mining and natural language processing. Many research efforts have been made concerning information extraction [1, 2 and 3] in order to identify and pull out a keyword that represents topics of information from a given sequence of words. The available information extraction tools that help researchers to cope with the overwhelmed information are for example, TextPresso Interface (National Human Genome Research Institute: an information Extracting and processing) [4], Chilibot (The U.S. National Institutes of Health, NIH: to search PubMed literature) [5], UBC Bioinformatics Center (Database on specific relations between proteins and genes by returning the result as a graph) [6], BBP (Brucella genome annotation with literature mining and a gateway to search and analyze data originating from public databases and literature) [7],

Knowledge extraction can be difficult and challenging due to the huge quantity of source material and due to the complex nature of the name entities and relations in literature [8, 9 and 10].

In the functional genomics literature, many complex named entities appear. To specify the word boundaries, the problem is considered to be Named Entity Recognition (NER) topics which uses a labeling based method (which is the one of the information extraction techniques) to specify the best boundary of the feature [3]. NER in the biological domain

has been studied for approximately ten years. Functional genomics named entity recognition remains a challenging task and an active area of research for many reasons as follows:

- The same word or phrase can refer to different entities depending upon their contexts. Conversely, many biological NEs have various spelling forms.
- Some modifiers are often placed before the basic NE, and sometimes functional genomics NEs are very long. These factors highlight the difficulties for identifying the boundary of NE.
- Abbreviations are frequently used in functional genomics domain. Since abbreviations do not have many evidences for certain NE class, it is difficult to classify them correctly.

According to the above, NER is an important process in functional genomics information extraction [11 and 12]. It helps to specific entities and features. There are many machine learning models that deal with NER. Recently, the Conditional Random Fields (CRFs) model has been proposed for Biomedical Named Entity Recognition [1, 11 and 13] and methods such as ABNER [14], an open source tool for automatically tagging genes, protein, DNA, RNA, cell-line, and cell-type entities [10]. The efficiency of this model indicated by an F-Measure was about 70% [1, 13]. Also for relation extraction, the trend of past research has been to focus on the detection of protein relationships. Recently in the systems biology recognizes in particular the importance of interactions between biological systems with known events. [15].To computationally search the literature for such events, text mining methods that can detect, extract and recognize are required. The

state of the art in relation extraction for the model performance is approaching a practically applicable level and revealing some remaining challenges. [15]

This thesis aims to improve biological information extraction with machine learning by using text mining, natural language processing and graphical models to capture functional genomics named entities, extract semantic relations and establish a specific knowledge base from the biological literature. Finally, the knowledge base is represented in a semantics graph network.

1.1 Objectives of this thesis

This thesis focuses on improving information extraction from the biological literature by using machine learning techniques. To establish knowledge, the Named entity recognition model is generated based on a graphical model (Conditional Random Fields). This work aims to recognize the classes of biological terms as protein, DNA, RNA, Cell-line, Cell-type while relation extraction is detected based on lexical and semantics by statistic methods and structure prediction (Tree kernel Support Vector Machine). In this work, relations refer to the bio-molecular events which relate with the functional genomics. The knowledge in this thesis refers to named entity recognition and the semantic relation extraction which is extracted from the biological literature.

1.2 Scope and Outline of this thesis

From the objectives of this thesis, the scope is designed as information extraction backgrounds, named entity recognition, biological relation extraction and establishing a

knowledge base. The validation is based on recall, precision in F-measure scale performance. Each part is calculated under scope of the [1]platform as **Table 1.1** and the framework as **Figure 1.1**.

Table1.1 Scope of the platform in thesis

Processes	Dataset	Description
Information Extraction Background	[2]NCBI corpus	• 120 Pubmed Abstracts (Evolution, Genetics, Genomes studies and Molecular Biology)
Named Entity Recognition	[3]GENIA corpus	• Training data 2,000 PubMed abstracts with term annotation • Evaluation data 404 PubMed abstracts, with annotation and one without for each. • Evaluation tool Updated evaluation tool. Use this tool to get the evaluation equivalent to that of the shared task.
Relation Extraction	NCBI corpus GENIA corpus [4]BioNLP2009	• 275 PubMed abstracts (Cyanobacteria) • 404 PubMed abstracts, with one file without term annotation • 2,402 abstracts files which relates bio-molecular events particularly on proteins or genes
Establishing Specific Knowledge base	-	• Knowledge(Named entity recognition and relation extraction) collaboration in semantic graph network

[1]The Model training and testing in named entity recognition process runs on platform as Xenon Dual Core, 4GB, HD 280 GB, Linux, Redhat while the others run on Intel Core i3-231M CPU2.1GHz RAM 6 GB 64-bit.
[2]http://www.ncbi.nlm.nih.gov/ [3]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
[4]http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/

The framework of thesis is designed as **Figure 1.1**

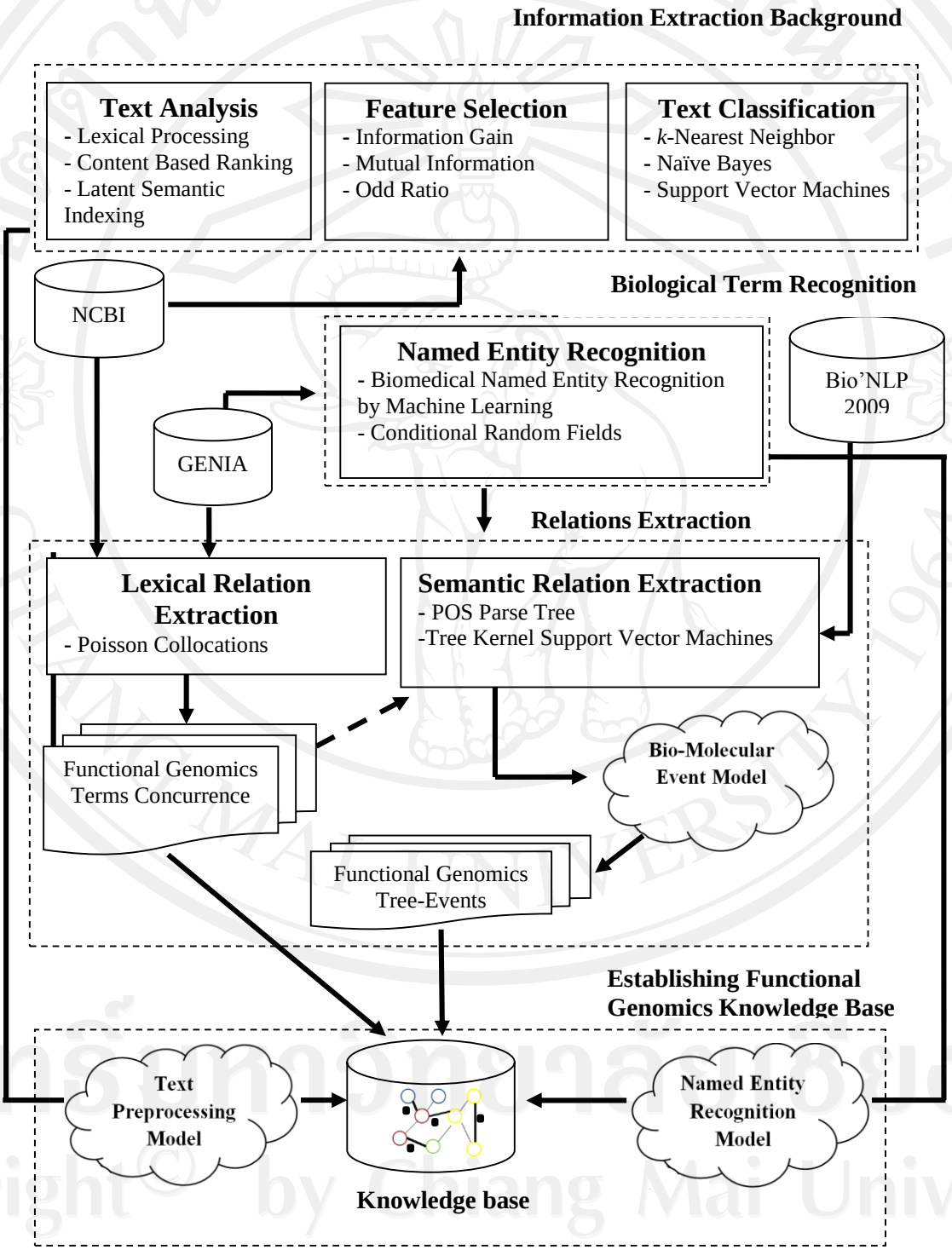

Figure 1.1 The proposed framework under this thesis

From **Figure 1.1,** the thesis outline is identified as follows,

Chapter 2 is about information extraction background for biological literature, and the related work in the bio-text mining fields, this chapter also presents the main ideas and experiments examined for the effectiveness of the general terms representing the biological literature by machine learning such as text analysis, scaling approach. This chapter shows the effectiveness of a general term to information retrieval based on Latent Semantic Indexing (LSI) in probabilistic retrieval. The last section, the feature selection techniques (Information Gain, Mutual Information and Odd Ratio) are studied with text classification approaches such as *k*-Nearest Neighbor, Naïve Bayes and Support Vector Machines (SVMs).

Chapter 3 relates to biological term recognition. From the motivations of **Chapter 2**, the Named Entity Recognition (NER) technique is considered in the biological information extraction process. Hence, the first part is the study of the concepts and comparison about the biological named entity recognition model based on machine learning and graphical models (Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Fields (CRFs)). To improve the named entity recognition model prediction, potential functions of CRFs were tested with the GENIA Corpus to compare the models performance.

After the NER step, biological terms are identified from **Chapter 3**. **Chapter 4** presents the relation extraction. The points of this chapter are to identify the relation extraction to the lexical and semantic relation extraction. For lexical relation extraction, a Poisson collocation is applied from the Poisson distribution to find biological terms

concurrence in biological literature. Then, Natural Language Processing (NLP) and structure prediction based on Tree Kernel Support Vector Machines (SVMs) is studied as bio-molecular event or semantic relation extraction. For the last section, knowledge is established using a semantic graph network to represent relations between biological terms which correspond with the biological knowledge base exist.

Chapter 5 presents the contributions and conclusions of this thesis.