# ENHANCED WORD EMBEDDING TECHNIQUE FOR BIOMEDICAL NAMED ENTITY RECOGNITION

MAAN TAREQ ABD

DISERTATION SUBMITED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTERS
OF COMPUTER SCIENCE
(ARTIFICIAL INTELLIGENCE)

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2017

# TEKNIK PERWAKILAN PERKATAAN DIPERTINGKATKAN UNTUK PENGECAMAN ENTITI NAMA BIOPERUBATAN

MAAN TAREQ ABD

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH
SARJANA SAINS KOMPUTER
(KECERDASAN BUATAN)

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2017

# DECLARATION

I hereby declare that the work in this thesis is my own, except for quotations and summaries, which have been duly acknowledged.

22 Jun 2017

MAAN TAREQ ABD
P84263

## ACKNOWLEDGMENTS

First and foremost, praise be to Almighty Allah for all his blessings by giving me patience and good health throughout the duration of this research.

I would like to express my sincere gratitude towards my supervisor, Assoc. Prof. Dr. Masnizah MOHD, for giving me the opportunity to work in this topic and for her continued guidance and advice during the course of this work.

I would like to thank all the post-graduate students of the UKM Natural Language Processing Research Group for their help, useful suggestions, and the pleasant working environment they provided throughout my years in UKM.

I give my utmost appreciation to my parents for supporting and encouraging me through all the years. I am grateful for their efforts in paving the way for my continued learning. Finally, I would like to thank my brother and sister for their interest and support. I would also like to thank all my friends and colleagues for their encouragement, cooperation, and support.

## ABSTRACT

Biomedical named entity recognition is the process to identify and classify technical entities in the domain of molecular biology such as protein, gene names, cell types, virus names and DNA sequence. Biomedical named entities have inherently complex structures which poses a big challenge for their identification and classification. The state of the art in supervised machine learning models still suffer from low performance in biomedical entity recognition task where there is still a wide gap between their performance in news-wire domains (≈ 91%) and their performance in biomedical domains (≈ 78%). To handle this problem, this research explores different effective word representations with support vector machine learning method to deal with the special characteristics of biomedical named entities. First, this research identifies and evaluates a set of morphological and contextual features with support vector machine learning method for biomedical named entity recognition. In addition, this research studies the effect of using prototypical word embedding technique (PWE) on the performance of support vector machine learning method. Furthermore, this research proposes a new model based on support vector machine and extended distributed prototypical word embedding technique (EDRWE) for biomedical named entity recognition. These models are evaluated on widely used standard biomedical named entity recognition dataset namely GENIA corpus. The results show that support vector machine model with morphological and contextual features achieves a good results with an overall F-measure of 70.6%. In addition, experimental results also show that both PWE and EDRWE word embedding technique achieve higher performance with an F-measure of 76.97% and 82.8% respectively, and significantly improves the overall performance of support vector machine learning for biomedical named entity recognition over traditional features representation technique. In general, results show that word representation is a key factor in constructing suitable recognition method.

# ABSTRAK

Pengecaman entiti nama bioperubatan adalah proses untuk mengenal pasti dan mengklasifikasi entiti teknikal dalam domain biologi molekul seperti protein, nama gen, jenis sel, nama virus dan urutan DNA. Entiti nama bioperubatan mempunyai struktur yang kompleks. Ini menyebabkan perkembangan model pembelajaran mesin mengalami prestasi yang rendah untuk mengecam entiti nama bioperubatan. Masih terdapat jurang yang besar antara prestasi pengecaman entiti nama dalam domain berita (≈ 91%) dan dalam domain bioperubatan (≈78%). Maka kajian ini meneroka perbezaan keberkesanan teknik perwakilan perkataan dengan dibantu oleh kaedah pembelajaran mesin sokongan vektor dalam usaha untuk menangani ciri-ciri khas entiti nama dalam bidang bioperubatan. Pertama, kajian ini dilakukan untuk mengecam dan menilai satu set ciri morfologi dan kontekstual dengan kaedah pembelajaran mesin sokongan vektor untuk mengenal pasti entiti nama bioperubatan. Di samping itu, kajian ini turut mengkaji kesan penggunaan *prototypical* dengan menerapkan teknik perwakilan perkataan (PWE) kepada prestasi kaedah pembelajaran mesin sokongan vektor. Satu model baharu berdasarkan mesin sokongan vektor iaitu *prototypical* lanjutan yang menerapkan teknik perwakilan perkataan (EDRWE) dicadangkan untuk mengecam entiti nama bioperubatan. Model PWE dan EDRWE ini dinilai dengan menggunakan korpus bioperubatan piawai GENIA korpus. Keputusan menunjukkan bahawa model mesin sokongan vektor dengan ciri-ciri morfologi dan kontekstual mencapai satu keputusan yang baik dengan keseluruhan *f-measure* sebanyak 70.6%. Keputusan eksperimen juga menunjukkan bahawa kedua-dua model mencapai prestasi yang tinggi dengan *f-measure* masing-masing sebanyak 76.97% (PWE) dan 82.8% (EDRWE). Keseluruhannya prestasi mesin sokongan vektor dalam pengecaman entiti nama bioperubatan meningkat dengan menggunakan teknik perwakilan perkataan berbanding menggunakan fitur tradisional. Ini membuktikan perwakilan perkataan adalah faktor utama dalam membina kaedah pengecaman yang sesuai.

## TABLE OF CONTENTS

# LIST OF TABLES

**Table Number** **Page**

## LIST OF FIGURES

**Figure Number** **Page**

# LIST OF ABBREVIATIONS

| | |
|---|---|
| BioTM | Biomedical Text Mining |
| NER | Named Entity Recognition |
| POS | Part Of Speech |
| SVM | Support Vector Machine |
| HMM | Hidden Markov Model |
| CRF | Conditional Random Field |
| NLP | Natural Language Processing |
| IR | Information Retrieval |
| IE | Information Extraction |
| PWE | Prototypical Word Embedding |
| EDRWE | Extended Distributed Prototypical Word Embedding |
| TM | Text Mining |
| ML | Machine Learning |
| MeSH | Medical Subject Headings |
| GSC | Gold Standard Corpora |
| SNN | Shared Nearest Neighbor |
| ME | Maximum Entropy |
| WR | Word Representation |
| PCA | Principal Component Analysis |
| PMI | Pointwise Mutual Information |
| RNN | Recurrent Neural Network |
| CNN | Convolutional Neural Network |
| SDP | Shortest Dependency Path |
| BI | BI-directional |

| | |
|---|---|
| LSTM | Long Short-Term Memory |
| Bio-NER | Biomedical Named Entity Recognition |
| EPMI | Extended Pointwise Mutual Information |
| EMI | Extended Mutual Information |

# CHAPTER I

# INTRODUCTION

## 1.1 INTRODUCTION

The huge volume of biomedical text, partly due to the massive growth of biomedical information available in the literature and the growing number of biomedical publications in recent years, has made the development of Biomedical Text Mining (BioTM) solutions indispensable. For instance, the MEDLINE literature database contains over 20 million references to journal papers, covering a wide range of biomedical fields. The MEDLINE database is currently growing at the rate of 500,000 new citations each year, In order to organize and manage these data, several manual creation efforts have been set up to identify, in texts, information regarding entities (e.g. genes and proteins) and their relations (e.g. protein-protein interactions). The extracted information is stored in structured knowledge resources, such as Swiss-Prot and GenBank. However, the effort required to continually update these databases makes this a very demanding and expensive task, naturally leading to increasing interest in the application of Natural Language Processing systems to help perform those tasks. The automatic recognition of biomedical entities from biomedical texts can markedly reduce the time that experts spend populating biomedical knowledge bases and annotating papers and patents. Biomedical entity recognition which sometimes referred to as biomedical concept identification aims to identify and classify technical terms in the domain of molecular biology that are of interest to biologists and scientists in order to understand the prevailing occurring names.

Technically speaking, biomedical entity recognition in bio-literature is similar to common named entity recognition (NER) systems. However, the task of biomedical entity recognition differs from other common NER problems because of the complex situations such as irregular expression, made of very long compounded words, hardly distinguished boundaries, same word or phrase can refer to different named entities and daily changing group members.

Three main approaches are generally used to recognize biomedical named entities ranging from dictionary matching, rule-based, to supervised approaches. Thanks to the availability of annotated biomedical corpora, supervised learning methods have been widely adopted, have yielded great success, and conquer other approaches. Recent supervised machine learning techniques could efficiently find gene names and clinical problems from certain type of texts with above 0.8 F-score(Zhang 2004). These techniques used different types of features: word level features (such as n-grams, numerical items and digits, word length and part-of-speech), lookup features extracted from dictionaries and gazetteers and document features (for example, co-occurrences of mentions). However, the supervised nature of supervised machine learning techniques relies on a fairly large amount of training data which must be annotated by humans. As a result, they are only applicable to a limited number of settings. Rule-based approaches are also used in the form of lexical patterns which use the grammar based (such as part of speech (POS), grammatically based (such as word priority) and orthographic based features (i.e. capitalization of the word) with the use of vocabularies in dictionary to detect biomedical named entities boundaries. In spite of its simplicity, rule-based technique requires a high understanding of biomedical naming standards as well as annotation guidelines and it carries a high overhead when the data changes. It means new rules for identification have to be created. It is known to be non-portable, non-dynamic and non-robust compared to other approaches. Dictionary or lexicon-based approaches are used where domain specific resources and gazetteers are used. Dictionary or lexicon-based approaches do not require training data, but usually involve ad hoc rules and resources and gazetteers that may limit the type of entities which they could extract.

In biomedical named entities recognition, the release of the GENIA corpus has pushed forward related research using various supervised learning models, including Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and Conditional Random Fields (CRFs). Among them, support vector machines have been recognized as a reliable, high-performance algorithm for different biomedical named entity recognition datasets.

Another important aspect for biomedical named entity recognition based on machine learning techniques is features sets used for designing the classification models. Previous works show the difficulty of biomedical named entities recognition by general features. Since it is necessary to explore more evidential word representation to recognize biomedical entities, we need to employ the contribution word embedding technique with powerful machine learning technique.

Word embedding can help make better performance of supervised learning models in natural language processing by grouping similar words. In word embedding, similar words are likely to have similar vectors. Word embedding can capture various dimensions of meanings and phrase information relevant to the potential features of words within the vector. As a result, word embedding is less sensitive to data insufficiency. Recently, word embedding has been used in different NLP task such a named entity recognition, sentiment analysis or parsing. Our hypothesis is that the incorporation of word embedding features into efficient supervised machine learning can improve their performance for biomedical named entity recognition task. Thus, this research will explore different strategies for deriving distributed word representations from word embedding in biomedical NER tasks.

## 1.2 PROBLEM STATEMENT

The exponential growth in the size of available biomedical literature has encouraged a lot of interest in developing efficient techniques for biomedical text mining. It has become extremely difficult for biologists to keep up with the relevant journals in their own discipline, let alone publications in other, related disciplines (Chowdhury & Mahbub 2013). Bio-medical literature considered as a source of authentic medical knowledge which is critical for e-health applications. Biological researchers are very considerable on the reality of use the knowledge that is founded inside bio-medical literature. For instance, there are above 22 million abstracts in the domains of medicine, bio-medical sciences and laboratory sciences in Medline alone. Information Retrieval (IR) tools help biologists by retrieving relevant articles in response to search queries. Information Extraction (IE) methods can automatically extract useful information to help in phenomenally reducing manual effort in creation tasks such as databases of proteins, genes, diseases, drugs and biological pathways.

Data representation is a fundamental task in machine learning. The representation of data affects the performance of the whole machine learning system. In a long history, the representation of data is done by feature engineering, and researchers aim at designing better features for specific tasks. Recently, the rapid development of word embedding has brought new inspiration to various domains and becomes one of the strongest trends in Natural Language Processing (NLP) at the moment. Word embeddings have been exceptionally successful in many NLP tasks and replaced traditional distributional features (Zamani & Croft 2017). Previous works show the difficulty of Bio-NER using traditional features such as morphological, syntactic and semantic information of words. These features heavenly account to the problem of data sparsity and they differ between entity types, which makes their development costly. Furthermore, these features are complex hand designed and often optimized for a specific gold standard corpus, which makes extrapolation of quality measures difficult. This motivates us to explore effective word representation and learning methods to deal with the special characteristics in the biomedical domain.

The problem that will be addressed in this research is biomedical entity recognition which aims to identify and classify technical entities in the domain of molecular biology. These entities are of interest to biologists and scientists such as protein and gene names, cell types, virus name, DNA sequence, and others. Biomedical named entities have inherently complex structures which poses a big challenge for their identification and classification in biomedical information extraction. The biomedical entity recognition is vast, but there is still a wide gap in performance between the systems developed for the news-wire domains and the existing systems in biomedical domains (Campos et al. 2012; Campos et al. 2013; Tang et al. 2014; Yang & Zhou 2014; Li et al. 2015; Li et al. 2015; Wang et al. 2015). Therefore, there is a room for improvement as recognition accuracy of name entity has basically hovered around 10 points in their F-measure. In addition, the ability of biomedical researchers to manage, integrate and analyze biomedical data is often limited due to a lack of tools, accessibility, and training (Yao et al. 2015). Moreover, the biomedical named entity recognition is more difficult than general named entity recognition because of the complex situations such as irregular expression, consist of long compounded words, hardly distinguished boundaries, same word or phrase can refer to different named entities and daily changing group members. The difficulty and potential importance of this task attract many researchers. These above factors make NER in the biomedical domain difficult. Therefore, it is necessary to explore effective word representation and learning methods to deal with the special characteristics in the biomedical domain.

## 1.3 RESEARCH OBJECTIVES

The aim of this work is to propose a biomedical named entity recognition model which combines a supervised machine learning method with word embeddings. To achieve this aim, the following objectives are proposed to be achieved:

i. To propose an extended distributed prototypical word embedding technique (EDRWE) for biomedical named entity recognition.

ii. To evaluate the proposed EDRWE and compare it with other representation techniques.

## 1.4 RESEARCH SCOPE

This research will focus on designing a new biomedical named entity recognition model for biomedical literature stored in electronic document form. The scope of this research focuses on biomedical data mining precisely biomedical named entities recognition task. Additionally, this research will centred on the supervised machine learning. This research explores different effective word representations with support vector machine learning method to deal with the special characteristics of biomedical named entities. First, this research identifies and evaluates a set of morphological and contextual features with support vector machine learning method for biomedical named entity recognition. In addition, this research studies the effect of using prototypical word embedding technique (PWE) on the performance of support vector machine learning method. Furthermore, this research presents a supervised biomedical named entity recognition model based on support vector machine and extended distributed prototypical word embedding technique (EDRWE) for biomedical named entity recognition. This research will explore different strategies for deriving distributed word representations from word embedding in biomedical NER tasks.

This research evaluates the proposed models upon a widely accepted dataset namely the GENIA corpus. The GENIA corpus is the primary collection of biomedical literature compiled and annotated within the scope of the GENIA project. The corpus was created to support the development and evaluation of information extraction and text mining systems for the domain of molecular biology. The corpus contains Medline abstracts, selected using a PubMed query for the three MeSH terms "human," "blood cells," and "transcription factors." The corpus has been annotated with various levels of linguistic and semantic information. The original GENIA corpus contains 36 classes of entities. A more widely used version of GENIA corpus is the one simplified for the BioNLP/NLPBA shared task, in which entities are grouped into only five major classes: protein, DNA, RNA, cell line, cell type.

## 1.5 OUTLINE OF THE THESIS

This thesis is divided into five chapters, which are the following:

**Chapter II:** **Literature review:** Focus on studying and analysing literature review. First, it covers the concepts and the state-of-the-art in biomedical named entity recognition. Second, it deals with the understanding of the challenges and problems of biomedical named entity recognition.

**Chapter III:** **Methodology:** Introduce the proposed extended distributed prototype word embedding technique with two levels of support vector machine learning technique Bio-NER methodology which is applied in this research.

**Chapter IV:** **Experimental Results:** The discussion of experimental results of different word representations and the proposed extended distributed prototype word embedding technique method.

**Chapter V:** **Conclusion:** This chapter present the summary of this research. It describes the contribution and the suggestions for future work of this research.

# CHAPTER II

# LITERATURE REVIEW

## 2.1 INTRODUCTION

It is well known that the rapid growth and spreading of the Internet has resulted in huge amounts of information generated and shared, available in the form of textual data, images, videos or sounds. A huge amount of biomedical documents and publications is becoming publicity available due to the recent adoption of electronic health records, the growing number of biomedical publications, and the exploding occurrence of health information online. The MEDLINE literature database contains over 20 million references to journal papers, covering a wide range of biomedical fields. In order to organize and manage these data, several manual annotation efforts have been set up to identify, in texts, information regarding entities (e.g. genes and proteins) and their relations (e.g. protein-protein interactions). The extracted information is stored in structured knowledge resources, such as Swiss-Prot (Boeckmann et al. 2003). However, these databases require continuous updating and the task to achieve this is quite expensive and demanding. This has led to a growing interest in the use of Text Mining (TM) systems to perform those tasks.

A crucial initial step in information extraction called NER is the major focus of TM research. It helps to identify chunks of text that relate to particular entities of interest, including the names of a protein, gene, drug and disease. Also, such systems can be incorporated in larger biomedical Information Extraction (IE) pipelines that make use of the automatically extracted names to perform various tasks such as classification, relation extraction or/and topic modelling. However, the task of text recognition becomes difficult as most biomedical names contain different characteristics (Zhang et al. 2004; Zhou et al. 2004) such as:

- Descriptive nature of entity names (e.g. ‘normal thymic epithelial cells’).
- One head noun shared by two or more entity names (e.g. ‘91 and 84 kDa proteins’ refers to ‘91 kDa protein’ and ‘84 kDa protein’).
- Several spelling forms of one entity name (e.g. ‘N-acetylcysteine’, ‘N-acetyl-cysteine’ and ‘NAcetylCysteine’).
- Frequent use of ambiguous abbreviations (e.g. ‘TCF’ may refer to ‘T cell factor’ or to ‘Tissue Culture Fluid’).

Consequently, several NER systems have been developed for the biomedical domain, using different approaches and techniques that can generally be categorized as being based on rules, dictionary matching or Machine Learning (ML). Each approach fulfills different requirements, depending on the linguistic characteristics of the entities being identified. Such heterogeneity is a consequence of the predefined naming standards and how faithfully the biomedical community followed them. Thus, it is recommended to take advantage of the approaches that better fulfill the requirements of each entity type. Thus, the approach that meets most of the requirements of each entity type must be leveraged:

- Rule-based approach: Matches names with a strongly defined morphological and orthographic structure.
- Dictionary-based approach: Matches with closely defined vocabulary names (e.g. species and diseases).
- Machine learning-based approach: Matches highly dynamic vocabulary and strong variability of names (e.g. proteins and genes).

As each approach has different technical requirements, employing the best approaches is not always possible for all cases (Campos et al. 2012). However, the advantages of machine learning methods are more when compared to other methods, and provide the best performance results as well when appropriate resources are available.

Various complex steps incorporating different processing pipelines are involved in the development of entity recognition methods, a machine learning-based system. In the past few years, different techniques, frameworks and strategies have been employed to develop a variety of systems. Thus, this chapter presents an outline of the Bio-NER methods and a detailed description of the latest and most crucial research techniques. It also gives an in-depth analysis of the available frameworks and systems, which takes into consideration the provided features, performance outcomes and technical characteristics.

## 2.2 GENES AND PROTEINS NAMES

Most of the developed methods are focused on two main corpora, GENETAG and JNLPBA. GENETAG is not restricted to a specific domain, containing annotations of proteins, DNA and RNA (grouped in only one semantic type), which were performed by experts in biochemistry, genetics and molecular biology. This corpus was used in the BioCreative II challenge (Smith et al 2008), and it contains 15000 sentences for training and 5000 sentences for testing. For evaluation, the matching is performed allowing alternative names provided by the expert annotators.

On the other hand, the JNLPBA corpus is a sub-set of the GENIA corpus, containing 2404 abstracts extracted from MEDLINE using the MeSH terms “human”, “bloodcell” and “transcription factor”. The manual annotation of these abstracts was based on five classes of the GENIA ontology, namely protein, DNA, RNA, cell line, and cell type. This corpus was used in the BioEntity Recognition Task in BioNLP/NLPBA 2004 (Kim et al 2004), which encompasses 2,000 abstracts for training and another 404 for assessment. For this test, evaluations were accomplished with the use of precise pairing. GENETAG is not centred on any particular domain of biomedical and thus its entries are more heterogeneously annotated than that of JNLPBA. A succinct analysis with consideration of proteins and DNA and RNA types demonstrate that GENETAG comprises nearly 65% dissimilar names in comparison to the 36% discovered in JNLPBA.