UNIVERSITI PUTRA MALAYSIA

IMPROVING NAMED ENTITY RECOGNITION ACCURACY FOR GENE AND PROTEIN IN BIOMEDICAL TEXT

HOSSEIN TOHIDI

FSKTM 2011 26

IMPROVING NAMED ENTITY RECOGNITION ACCURACY FOR GENE AND PROTEIN IN BIOMEDICAL TEXT [redacted]

By

HOSSEIN TOHIDI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirement for the Degree of Master of Science

AUGUST 2011

بســــــــــــــــم الله الـرحمن الـرحيم

كَمَا أَرْسَلْنَا فِيكُمْ رَسُولا مِنْكُمْ يَتْلُو عَلَيْكُمْ آيَاتِنَا وَيُزَكِّيكُمْ وَيُعَلِّمُكُمُ
الْكِتَابَ وَالْحِكْمَةَ وَيُعَلِّمُكُمْ مَا لَمْ تَكُونُوا تَعْلَمُونَ.
سوره البقره – آیه 151

عاقلان نقطه ی پرگار وجودند
ولی عشق داند که در این دایره سرگردانند

The sages are the center of the compass of existence; but
Love knoweth that, in this circle, they head-revolving are.
Hafez Shirazi, Grate Iranian Poet

DEDICATION

To my dear parents, that I owe them each moment of my life
To my dear and gracious wife Sima for her support and her love which is my
motivation. To my country IRAN and my people which I respect them and I love them.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

IMPROVING NAMED ENTITY RECOGNITION ACCURACY FOR GENE AND PROTEIN IN BIOMEDICAL TEXT [redacted]

By

HOSSEIN TOHIDI

JANUARY 2011

Chairman : Assoc. Prof. Hamidah Ibrahim, PhD

Faculty : Computer Science and Information Technology

ABSTRACT

The plethora of biomedical material on the WWW is one of the factors that have sustained interest in automatic methods for extracting information from biomedical document, which can help biologists in their research. To extract useful knowledge from the biomedical literature, we must be able to recognize names of biomedical entities, such as genes, proteins, cells, and diseases which are called Named Entity. The task of recognizing entity-denoting expressions, or named entities (NE), in natural language documents is called Named Entity Recognition (NER). Among the biomedical types such as gene, protein, virus, cells, and etc, the most important biomedical types for recognition are gene and protein, which is the scope of this research. The most important reason why most researchers focus on the gene and protein named entities is due to the complexity nature of such types. This complexity includes the issues of *character-level variation, word-level variation,* and *word order variation* in biomedical text literature.

Typically there are four approaches for Named Entity Recognition, namely: Dictionary-Based, Rule-Based, Statistical and Machine Learning, and Hybrid approaches. In this study, to handle the above issues in recognizing gene and protein names, a statistical similarity measurement as a pattern matching function is proposed. Our approach is based on an assumption that a named entity occurs among a noun group which is extracted using *Brill Part of Speech* tagger. The strength of our proposed approach for recognizing biomedical named entity is based on a *Statistical Character-Based Syntax Similarity (SCSS)* algorithm which measured similarity between all extracted candidates and the well-known biomedical named entities from a corpus. For this study, we have used the GENIA V3.0 corpus, which is the largest annotated corpus in the molecular and biology domain. The proposed approach is evaluated based on two measures: *recall* and *precision* which are used to calculate a balanced *F*-test. We have compared our pattern matching function with the other methods and result is satisfied as *precision* is 98.5% and *recall* is 96.4%, while the *F*-test is 97.5 for both gene and protein names recognizing and *precision* is 99.3% and *recall* is 99.1%, while the *F*-test is 99.1 for protein names recognizing.

ABSTRAK

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai
memenuhi keperluan untuk ijazah Master Sains

PENGECAMAN ENTITI BERNAMA UNTUK GEN DAN PROTEIN DALAM [redacted] TEKS BIOPERUBATAN

Oleh

HOSSEIN TOHIDI

JANUARI 2011

Pengerusi : Assoc. Prof. Hamidah Ibrahim, PhD

Fakulti : Fakulti Sains Komputer dan Teknologi Maklumat

Kelimpahan bahan bioperubatan di WWW merupakan salah satu faktor yang telah menyebabkan minat berterusan dalam kaedah automatik untuk mengekstrak maklumat dari dokumen bioperubatan, yang mana boleh membantu ahli biologi dalam penyelidikan mereka. Untuk melombong pengetahuan yang berguna dari literatur bioperubatan, kita mestilah mampu untuk mengecam nama entiti bioperubatan, seperti gen, protein, sel, dan penyakit yang dikenali sebagai entiti bernama. Tugas mengecam ungkapan perwakilan entiti, atau entiti bernama, dalam dokumen bahasa tabii dipanggil Pengecaman Entiti Bernama. Alasan paling penting mengapa kebanyakan penyelidik fokus ke atas entiti bernama bioperubatan adalah disebabkan oleh sifat kerumitan teks tersebut. Kerumitan ini termasuk isu *variasi tahap-aksara*, *variasi tahap-perkataan*, dan *variasi urutan kata* dalam literatur teks bioperubatan.

Kebiasaannya terdapat empat pendekatan untuk Pengecaman Entiti Bernama, iaitu: pendekatan Dictionary-Based, Rule-Based, Statistical and Machine Learning, dan Hybrid. Di dalam kajian ini, untuk menggendalikan isu di atas dalam mengecam nama gen dan protein, pengukuran statistical similarity sebagai fungsi padanan corak dicadangkan.

Dalam kajian ini satu pendekatan bagi Pengecaman Entiti Bernama bioperubatan yang mengendalikan isu di atas untuk mengecam nama gen dan protein telah dicadangkan. Pendekatan kami adalah berdasarkan kepada suatu andaian di mana satu entiti bernama ujud antara satu kumpulan kata nama yang diekstrak menggunakan penanda *Brill Part of Speech*. Kekuatan pendekatan yang dicadangkan oleh kami untuk mengecam entiti bernama bioperubatan adalah berdasarkan kepada algoritma *Statistical Character-Based Syntax Similarity* (SCCS) yang mengukar persamaan antara kesemua calon yang diekstrak dan entiti bernama bioperubatan terkenal dari sebuah korpus. Untuk kajian ini, kami telah menggunakan korpus GENIA V3.0, yang merupakan korpus beranotasi terbesar dalam domain molekul dan biologi. Pendekatan yang dicadangkan dinilai berdasarkan dua ukuran: *perolehan kembali* dan *ketepatan* yang digunakan untuk mengira suatu ujian-*F* yang seimbang. Hasil adalah memuaskan di mana *ketepatan* ialah

98.5% dan *perolehan kembali* ialah 96.4%, sementara ujian-*F* ialah 97.5 untuk pengecaman kedua-dua nama gen dan protein dan *ketepatan* ialah 99.3% dan *perolehan kembali* ialah 99.1%, sementara ujian-*F* ialah 99.1 utuk pengecaman nama protein.

ACKNOWLEDGEMENTS

In the name of *ALLAH*, the most merciful and most compassionate. Praise to *ALLAH* S. W. T. who granted me strength, courage, patience and inspiration to complete this work.

First and foremost, I heartiest would to sincere thanks my supervisor Associate Professor Dr. Hamidah Ibrahim, for her incredible guidance, continues support, and encouragement. Always having time for me and readily providing her technical expertise throughout the period of my study. I owe more than I can ever repay. Only has the successful completion of this work become possible due to her supervision, she is the first person to thank for making my Master program at the Universiti Putra Malaysia a very enjoyable experience. She is the one who gave me invaluable and worthwhile advice on this research, and gave me a panorama of observance into the knowledge in my field. Moreover I want to appreciate her tenuous and professional comments for my English writing.

To my advisor Dr. Masrah Azrifah Azmi Murad, I would like to express appreciation for her insightful comments and suggestion on the wok. Her critical comments for my paper are extremely valuable for the improvement in my thinking.

I would like to thank the many people whom I have met during my stay in Malaysia for their friendship, enjoyable, discussion and good times.

I would like to express my love and deepest thanks to my noblest father Mohammad Tohidi and my great mother Zahrah Hakini were the reason of my success, I am indebted to them.

Finally, I would like to express my thanks to my beloved wife, Sima, who has offered moral support, encouragement and patient companionship during my study.

Hossein Tohidi

January 2011

I certify that an Examination Committee met on [date of viva] to conduct the final examination of Hossein Tohidi on his thesis entitled "Named Entity Recognition for Gene and Protein in Biomedical Text [redacted]" in accordance with Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the candidate be awarded the Master of Science.

Members of the Examination Committee are as Follows:

Chairman
Title: Assoc. Prof. Dr. Rusli bin Hj Abdullah
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Examiner 1
Title: Dr. Lilly Suriani Affendey
Faculty of
Universiti
(Internal Examiner)

Examiner 2
Title: Assoc. Prof. Dr. Md Nasir b Sulaiman
Faculty of
Universiti
(Internal Examiner)

Examiner 3
Title: Y. Bhg. Prof. Dr. Zainab binti Abu Bakar
Faculty of
Universiti
(External Examiner)

NORIATAH OMAR, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee are as follows:

Hamidah Ibrahim, PhD
Associate Professor
Department of Computer Science
Universiti Putra Malaysia
(Chairman)

Masrah Azrifah Azmi Murad, PhD
Lecturer
Department of Information Systems
Universiti Putra Malaysia
(Member)

HASANAH MOHD GHAZALI, PhD
Professor/Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citation which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

HOSSEIN TOHIDI

Date:

TABLE OF CONTENTS