UNIVERSITI PUTRA MALAYSIA

LEXICAL PARAPHRASE EXTRACTION WITH MULTIPLE SEMANTIC INFORMATION

HO CHUK FONG

FSKTM 2012 2

LEXICAL PARAPHRASE EXTRACTION WITH MULTIPLE SEMANTIC INFORMATION

HO CHUK FONG

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2012

LEXICAL PARAPHRASE EXTRACTION WITH MULTIPLE SEMANTIC INFORMATION

By

HO CHUK FONG

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

August 2012

DEDICATION

TO

My beloved parents who have devoted their life to their children,

My beloved mother who has been giving me support along my research journey,

My brother, my sister and my sister in law,

And

To all people who live with peace and wisdom.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

LEXICAL PARAPHRASE EXTRACTION WITH MULTIPLE SEMANTIC INFORMATION

By

HO CHUK FONG

August 2012

Chairperson: **Masrah Azrifah Azmi–Murad, PhD**

Faculty: **Computer Science and Information Technology**

Natural language processing (NLP) refers to the interaction that happens between humans and computers where computers try to understand and make sense of human languages. However, human beings tend to express similar meanings using sentences with different structures or different surface wordings. Due to this phenomenon called variability, NLP becomes a difficult task. Since paraphrases are different words, phrases or sentences that express the same or almost the same meaning, a variety of paraphrase extraction methods have been proposed believing that paraphrases can be used to capture this variability.

In general, paraphrase extraction methods can be categorized into corpus–based and knowledge–based. A corpus–based method is dependent on syntax information (rules that govern the arrangement of words to form phrases in sentences) while a knowledge–based method is dependent on semantic information (the study of meanings). However, previous studies have shown that depending on syntax information alone can result in

mistakenly extracting antonyms and barely related or unrelated words as paraphrases. Semantics on the other hand is a complex study of meanings. Therefore, extracting paraphrases based on shallow or a single instance of semantic information only such as synonyms or semantic relations would be ineffective as it has no difference from solving a complex problem based on incomplete information.

The main purpose of this thesis is to propose a new model, called Multilayer Semantic–based Validation Paraphrase Extraction (MSVPE), which relies on the use of different types of semantic information. In particular, MSVPE collects paraphrase candidates from multiple instances of lexical resources. Then, it validates the candidates using word similarity method, sentence similarity method and domain matching technique which correspond to the use of semantic relations, definitions and domains respectively.

However, there are some flaws in the existing sentence similarity methods and word similarity methods. In particular, sentence similarity methods determine the semantic similarity between sentences based on the incorrect interpretation of meaning from each sentence and with incomplete information. Word similarity methods on the other hand derive the semantic similarity between words based on multiple features which have not been processed and combined properly. Consequently, similarity judgments produced by them are not reliable. To address these problems, we also proposed: 1) a new sentence similarity method (SSMv1) that compares the actual meaning of each sentence, 2) another sentence similarity method (SSMv2) that takes into consideration multiple pieces of information, and 3) a new word similarity method (WSM) that makes use of optimally processed and combined features.

In order to evaluate MSVPE, SSMv1, SSMv2 and WSM, four different experiments have been conducted on three different data sets. SSMv1, SSMv2 and WSM were tested on two standard data sets which consist of 30 pairs of definitions and 65 pairs of nouns respectively that ranged from highly synonymous to semantically unrelated and which have widely been applied for evaluation purposes. In contrast, MSVPE was tested on a data set created in this study which consists of 85 words and 56 sentences.

Experimental results showed that compared with the two benchmarks based solely on syntax information, MSVPE can extract paraphrases more effectively. This is probably because semantic information is more related to meanings than syntax information. Results further showed that MSVPE with multiple instances of semantic information outperforms MSVPE with only a single instance of semantic information. Although the effectiveness of different semantic information varies, they are complementary.

Experimental results also showed that SSMv1, SSMv2 and WSM outperform all of their benchmarks significantly, thus indicating that they can better simulate human inferring capability. The reason is that SSMv1 has the correct understanding of the meaning of each sentence while SSMv2 makes use of information that is complementary. WSM on the other hand consists of the optimized transformation of different types of features and the optimized combination of them representing the nearest replica of human thinking behavior.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENGEKSTRAKAN LEKSIKAL PARAFRASA DENGAN GANDAAN MAKLUMAT SEMANTIK

Oleh

HO CHUK FONG

Ogos 2012

Pengerusi: **Masrah Azrifah Azmi–Murad, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Pemprosesan bahasa tabii (NLP) merujuk kepada interaksi yang berlaku di antara manusia dan komputer di mana komputer cuba memahami bahasa manusia. Walau bagaimanapun, manusia menunjukkan kecenderungan untuk mengekspresi makna yang lebih kurang sama dengan menggunakan struktur ayat yang berlainan atau ayat yang mengandungi perkataan–perkataan permukaan yang berbeza. Disebabkan oleh fenomena ini yang dipanggil kebolehubahan, NLP telah menjadi satu tugas yang sukar. Ini kerana parafrasa adalah perkataan–perkataan, frasa–frasa atau ayat–ayat yang berbeza tetapi membawa makna yang sama atau makna yang lebih kurang sama, maka, pelbagai kaedah untuk mengekstrak parafrasa telah dicadangkan di atas kepercayaan bahawa parafrasa boleh digunakan untuk menangani kebolehubahan tersebut.

Secara umum, kaedah–kaedah yang digunakan untuk mengekstrakan parafrasa boleh dikategorikan kepada kaedah yang berasaskan korpus dan kaedah yang berasaskan pengetahuan. Kaedah yang berasaskan korpus bergantung kepada maklumat sintaks

(peraturan yang menentukan susunan perkataan–perkataan untuk membentuk frasa–frasa di dalam ayat) manakala kaedah yang berasaskan pengetahuan bergantung kepada maklumat semantik (kajian berkaitan dengan makna). Walau bagaimanapun, kajian dahulu telah menunjukkan bahawa ketergantungan kepada maklumat sintaks sahaja boleh menyebabkan kaedah–kaedah tersebut mengekstrak antonim, kata–kata yang hampir tidak berkaitan atau kata–kata yang tidak berkaitan sebagai parafrasa dengan tidak sengaja. Sebaliknya, semantik adalah kajian makna yang kompleks. Oleh sebab itu, pengekstrakan parafrasa dengan menggunakan maklumat semantik yang cetek atau terhad sahaja seperti sinonim atau hubungan semantik adalah tidak berkesan kerana cara tersebut serupa dengan menyelesaikan masalah yang kompleks dengan menggunakan maklumat yang tidak lengkap.

Tujuan utama tesis ini adalah untuk mencadangkan satu model baharu yang dipanggil pengekstrakan paraphrases dengan menggunakan pelbagai lapisan pengesahan yang berasaskan maklumat semantik (MSVPE) yang bergantung kepada perlbagai jenis maklumat semantik yang berlainan. Pertama, MSVPE mengumpulkan calon–calon parapfrasa daripada pelbagai sumber leksikal. Kemudian, MSVPE akan mengesahkan calon–calon dengan menggunakan kaedah yang mengukur persamaan di antara perkataan–perkataan, kaedah yang mengukur persamaan di antara ayat–ayat dan kaedah yang membandingkan kesepadanan di antara domain di mana kaedah–kaedah tersebut bergantung kepada penggunaan hubungan semantik, definisi dan domain.

Walau bagaimanapun, terdapat beberapa kelemahan dalam kaedah–kaedah sedia ada yang mengukur persamaan di antara perkataan-perkataan dan di antara ayat–ayat.

Khususnya, kaedah–kaedah yang mengukur persamaan semantik di antara ayat–ayat membandingkan persamaan di antara makna ayat–ayat dengan berdasarkan makna yang salah difahami dari setiap ayat. Selain itu, mereka membuat perbandingan dengan menggunakan maklumat yang tidak lengkap. Kaedah–kaedah yang mengukur persamaan semantik di antara perkataan–perkataan bergantung kepada pelbagai ciri–ciri yang tidak diproses dan digabungkan dengan tidak betul. Akibatnya, keputusan persamaan yang dibuat oleh kaedah–kaedah tersebut adalah tidak boleh dipercayai. Untuk menangani masalah ini, kami juga mencadangkan: 1) kaedah baru yang mengukur persamaan di antara ayat–ayat (SSMv1) dengan mengambil kira makna sebenar yang disampaikan oleh setiap ayat, 2) satu lagi kaedah baru yang mengukur persamaan di antara ayat–ayat (SSMv2) dengan mengambil kira maklumat–maklumat yang saling melengkapi, dan 3) kaedah baru yang mengukur persamaan di antara perkataan–perkataan (WSM) dengan menggunakan pelbagai ciri–ciri yang telah diproses dan digabungkan secara terbaik.

Untuk menilai MSVPE, SSMv1, SSMv2 dan WSM, empat uji kaji yang berbeza telah dijalankan dengan menggunakan tiga set data yang berlainan. SSMv1, SSMv2 dan WSM telah diuji dengan menggunakan dua set data piawai yang mengandungi 30 pasang definisi dan 65 pasang kata nama yang telah digunakan secara meluas untuk tujuan penilaian. Kedua–dua data tersebut terdiri daripada pasangan yang sangat serupa secara semantik sehingga pasangan yang tidak berkaitan secara semantik. Sementara itu, MSVPE telah diuji dengan menggunakan set data yang dicipta dalam kajian ini yang mengandungi 85 perkataan dan 56 ayat.

Keputusan ujikaji menunjukkan bahawa MSVPE boleh mengekstrak parafrasa dengan lebih berkesan apabila dibandingkan dengan tanda–tanda aras yang berdasarkan maklumat sintaks sahaja. Ini kerana maklumat semantik adalah lebih berkaitan dengan makna daripada maklumat sintaks. Keputusan juga menunjukkan bahawa MSVPE yang bergantung kepada pelbagai maklumat semantik boleh mengekstrak parafrasa dengan lebih berkesan daripada MSVPE yang hanya bergantung kepada satu maklumat semantik. Walaupun keberkesanan maklumat–maklumat semantik yang berbeza adalah berlainan, mereka adalah saling melengkapi.

Keputusan ujikaji juga menunjukkan bahawa SSMv1, SSMv2 dan WSM mengatasi prestasi–prestasi tanda–tanda aras mereka secara penting. Ini mengesahkan bahawa mereka boleh menyamarkan keupayaan manusia untuk membuat kesimpulan dengan lebih baik. Sebabnya adalah SSMv1 mampu mencapai pemahaman maksud setiap ayat yang sebenar manakala SSMv2 menggunakan pelbagai jenis maklumat yang saling melengkapi. WSM mengandungi ciri–ciri yang telah diproses and digabungkan secara terbaik untuk mewakili replika yang paling dekat dengan keupayaan manusia untuk membuat kesimpulan.

ACKNOWLEDGEMENTS

First and foremost, all praise to the almighty God for his blessings and merciful that enable me to learn.

I am sincerely grateful to my supervisor, Assoc. Prof. Dr. Masrah Azrifah Azmi–Murad, for giving me the great opportunity and confidence to work under her professional and thorough supervision, for her genuine interest in my research and career, for never being too busy to set regular time aside for getting together, for stimulating conversations, for providing invaluable advices on many topics and for her patience. Besides, I would like to express my sincere thanks and appreciation to the supervisory committee members, Assoc. Prof. Dr. Shyamala Doraisamy and Dr. Rabiah Abdul–Kadir for their continued patience, guidance, suggestions and advices throughout the journey.

My PhD study was supported by a grant fellowship (GRF) from the School of Graduate Studies at University Putra Malaysia. I would like to take this opportunity to thank the University for the generous financial support.

I cannot end without thanking my family. I owe so much to my dear mother and father, who are always my source of inspiration and who encourage me to learn and support me throughout my life. I would also like to thank my brother, my sister and my sister in law for their patience.

Ho Chuk Fong
April 2012

APPROVAL

I certify that a Thesis Examination Committee has met on 14 August 2012 to conduct the final examination of Ho Chuk Fong on his thesis entitled "Lexical Paraphrase Extraction with Multiple Semantic Information" in accordance with the Universities and University College Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Rusli bin Hj Abdullah, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Ali bin Mamat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Aida binti Mustapha, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Christopher Hinde, PhD
Professor
Loughborough University
United Kingdom
(External Examiner)

_____
SEOW HENG FONG, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of philosophy. The members of the Supervisory Committee were as follows:

Masrah Azrifah Azmi Murad, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairperson)

Shyamala Doraisamy, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Rabiah Abdul Kadir, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

_____
BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at University Putra Malaysia or other institution.

HO CHUK FONG

Date: 14th August 2012

TABLE OF CONTENTS