UNIVERSITI PUTRA MALAYSIA

QURANIC ONTOLOGY FOR RESOLVING QUERY TRANSLATION DISAMBIGUATION IN ENGLISH-MALAY CROSS-LANGUAGE INFORMATION RETRIEVAL

ZULAINI BINTI YAHYA

FSKTM 2012 27

QURANIC ONTOLOGY
FOR RESOLVING QUERY TRANSLATION DISAMBIGUATION
IN ENGLISH-MALAY CROSS-LANGUAGE INFORMATION RETRIEVAL

By

ZULAINI BINTI YAHYA

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

November 2012

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

**QURANIC ONTOLOGY
FOR RESOLVING QUERY TRANSLATION DISAMBIGUATION
IN ENGLISH-MALAY CROSS-LANGUAGE INFORMATION RETRIEVAL**

By

ZULAINI BINTI YAHYA

November 2012

Chairman: Muhamad Taufik bin Abdullah, PhD

Faculty: Computer Science and Information Technology

This research proposed a Cross Language Information Retrieval (CLIR) method based on specific domain/ontology using specific concepts for disambiguating translation of the query. This research experiment the use of specific domain/ontology: Quran, written in English and Malay languages as a bilingual parallel-corpora and specific concepts: Quran, as a resource for cross-language query translation along with dictionary-based translation.

This study evaluates the effectiveness of query translation using dictionary-based and ontology for CLIR system. For translation, we use two basic approaches as benchmark: 1) first translation listed in the dictionary; and 2) all translation candidates listed in the dictionary. For the proposed CLIR method, we use three approaches: 1) based on verse list; 2) based on concepts similarity; and 3) based on concepts expansion. For concepts

matching before and after query translation, we used two approaches: 1) query concepts; and 2) translation concepts.

The experimental result shows that retrieval performance using dictionary-based is lower than monolingual either in English or Malay document collections. Direct translation involved in returning many possibility results which can affect the decreasing in document retrieval performance either in English or Malay document collections.

For the proposed CLIR method, performance of CLIR query translation based on verse list approach, concepts similarity approach and concepts expansion approach, obtained a better result either using query concepts or translation concepts matching compared to dictionary-based for English document collections but not in Malay document collections. In Malay document collections the retrieval performance only improved in concepts expansion approach. English language has a better structure compared to Malay language which affects the retrieval performance. A single Malay word may have a variety of meaning, not only by the word itself but also depends on the meaning of the verse or chapter. This is one of the reasons why retrieval performance decreasing in Malay document collections.

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

ONTOLOGI QURAN UNTUK MENYELESAIKAN PENYAHTAKSAAN TERJEMAHAN PERTANYAAN DALAM DAPATAN SEMULA MAKLUMAT SILANG BAHASA INGGERIS-MELAYU

Oleh

ZULAINI BINTI YAHYA

November 2012

Pengerusi: Muhamad Taufik bin Abdullah, PhD

Fakulti: Sains Komputer dan Teknologi Maklumat

Kajian ini mencadangkan kaedah Dapatan Semula Maklumat Silang Bahasa (DSMSB) berdasarkan domain/ontologi khusus dengan menggunakan konsep khusus untuk penyahtaksaan terjemahan pertanyaan. Kajian ini menggunakan domain/ontologi khusus: Al-Quran yang ditulis dalam bahasa Inggeris dan bahasa Melayu sebagai korpus-selari dwibahasa, kamus dwibahasa dan konsep khusus: Quran yang ditulis dalam bahasa Inggeris dan bahasa Melayu sebagai sumber untuk merentas terjemahan pertanyaan.

Kajian ini menilai keberkesanan terjemahan pertanyaan dengan menggunakan kamus dwibahasa dan ontologi untuk sistem DSMSB. Untuk terjemahan, kami menggunakan dua pendekatan sebagai penanda aras, iaitu: 1) terjemahan pertama yang tersenarai dalam kamus; dan 2) semua terjemahan yang tersenarai dalam kamus. Dalam kaedah DSMSB

yang dicadangkan, kami menggunakan tiga pendekatan iaitu: 1) berdasarkan senarai surah; 2) berdasarkan persamaan konsep; dan 3) berdasarkan pengembangan konsep. Untuk penggunaan padanan konsep sebelum dan selepas terjemahan pertanyaan, kami menggunakan dua pendekatan, iaitu: 1) konsep pertanyaan; dan 2) konsep terjemahan.

Hasil kajian menunjukkan prestasi terjemahan pertanyaan menggunakan kamus dwibahasa pada sistem DSMSB lebih rendah berbanding dengan dapatan semula maklumat satu bahasa. Terjemahan secara langsung menghasilkan pelbagai kemungkinan jawapan yang menyebabkan penurunan prestasi bagi koleksi Inggeris ataupun Melayu.

Bagi pendekatan DSMSB cadangan, prestasi terjemahan pertanyaan menggunakan pendekatan berdasarkan senarai surah, berdasarkan persamaan konsep and berdasarkan pengembangan konsep, mendapat keputusan yang lebih baik sama ada dengan menggunakan konsep pertanyaan atau konsep terjemahan berbanding dengan kamus bagi koleksi dokumen Inggeris tetapi tidak dalam koleksi dokumen Melayu. Prestasi terjemahan pertanyaan hanya baik dengan menggunakan pendekatan berdasarkan pengembangan konsep bagi koleksi dokumen Melayu. Bahasa Inggeris mempunyai struktur yang lebih baik berbanding dengan bahasa Melayu yang mana ia memberi kesan kepada prestasi terjemahan pertanyaan. Satu perkataan bahasa Melayu boleh mempunyai pelbagai makna, bukan sahaja dari perkataan itu sendiri, tetapi juga

bergantung kepada makna ayat atau bab. Inilah satu sebab mengapa prestasi dapatan semula maklumat menurun dalam koleksi dokumen Melayu.

ACKNOWLEDGEMENTS

Alhamdulillah, praise to Allah Almighty. With His gift and permission, I was given the strength and perseverance in completing this thesis as fulfillment of the assignment for the degree of Master of Science (Information Retrieval) successfully. However, this success could not have achieved without the guidance, assistance, cooperation, support, and encouragement of certain people. I would like to extend my appreciation and gratitude to the institutions and individuals who have jointly helped me to achieve this success.

Foremost, I would like to thank my supervisor Dr. Muhamad Taufik bin Abdullah, who shared with me a lot of his time, expertise and research insight. His faithful encouragements, keen insight, worthy guidance, and valuable suggestions throughout the academic period have helped me immensely to achieve success in both my research and completing this thesis. I am also grateful to my committee members, Dr. Azreen bin Azman and Dr. Rabiah Binti Abdul Kadir, for their helpful suggestions and comments on my research.

I also acknowledge my lecturer, Assoc. Prof. Hj. Mohd. Hasan bin Selamat, who shared with me an ideas, knowledge and experience in understanding the research method. My thanks are also extended to all colleagues in the Faculty of Computer Science and Information Technology at Universiti Putra Malaysia, for their sincere cooperation by spending time, sharing knowledge and giving ideas in understanding and completing the requirements of the compulsory subjects in my studies.

Acknowledgements also directed to Dato' Hj. Termuzi bin Hj. Abdul Aziz, Director General of Institute of Language and Literature, Mr. Kamarul Zaman bin Shaharudin, deputy director general of Institute of Language and Literature for their permission and financial support. Special thanks also given to Mr. Sulaiman bin Kaiat, head of Information System Department for giving me an opportunity to attend this program. My thanks are extended to

all members of Information System Department for their support, enthusiasm and guidance.

Finally, and the most importantly, infinite thank are given to my parents, Hj. Yahya bin Ismail and Hjh. Puteh binti Jusoh, for their love, guidance and their prayer are always on their lips and hearts. Sincere thanks are given to my brothers and sisters, for their love and understanding. Their presence and support most deeply felt and appreciated, makes my life more meaningful. To them I dedicate this thesis.

Thank you Allah

I certify that a Thesis Examination Committee has met on **22 November 2012** to conduct the final examination of Zulaini binti Yahya on her thesis entitled "**Quranic Ontology for Resolving Query Translation Disambiguation in English-Malay Cross-Language Information Retrieval**" in accordance with the 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 student awarded the degree of Master of Science (Information Retrieval).

Members of the Thesis Examination Committee were as follows:

Rahmita Wirza O.K. Rahmat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

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

Hamidah binti Ibrahim, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Zainab binti Abu Bakar, PhD
Professor
Faculty of Information Technology and Mathematical Sciences
University Technology Mara
(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 fulfilment of the requirement for the degree of Master of Science (Information Retrieval). The members of Supervisory Committee were as follows:

Muhamad Taufik bin Abdullah, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Azreen bin Azman, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Rabiah binti 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 Universiti Putra Malaysia or at any other institution.

ZULAINI BINTI YAHYA

Date: 22 November 2012

TABLE OF CONTENTS