COMPARISON OF ONTOLOGY LEARNING TECHNIQUES FOR QUR’ANIC TEXT

CHEW KIM MEY

A project report submitted in partial fulfilment of the requirements for the award of the degree of Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

APRIL 2010

Dedicated, in thankful appreciation to my beloved
father, CHEW CHONG PING
mother, TING NAN YUAN
brother, CHEW KIM SHING
sisters, CHEW KIM YEH and
CHEW KIM LIN
friend, YONG CHING YEE

ACKNOWLEDGEMENT

First, I give thanks to God for giving me this opportunity to further my study in such a good condition environment. A special appreciation dedicates to my supervisor, Professor Dr. Naomie bte Salim, for her patient and guidance in this study and the writing of the thesis. I also would like to thanks all the members in Ontology Learning group especially to Mr. Syafrullah and Madam Saidah Saad for their helping and guidance in this project. I also give thanks to all the staff in the faculty especially Madam Lijah who always provide the latest information for me. I thank you to all my friends as well, who always give me both mentally and physically support. Finally, I would like to give thanks to my family for their financial support along the years while I continue my study. Thank you.

ABSTRACT

Currently, ontology plays an important role in semantic Web technology and defines the concepts and relationships among these concepts. Ontology learning approach is to distinguish according to the type of input such as text, dictionary, knowledge, policies, schemes and schemes of semi-structured relations. Ontology learning can be explained as extract information subtask and ontology learning objectives is to dig the relevant concepts and relationships from the corpus or a particular type of data sets. In this project, I will focus on ontology learning from text using Qur'anic text as input data. The approaches which used to extract Qur'anic text in this project are Alfonseca and Manandhar's method and Gupta and Colleagues's approach. After completed the project, I hope to exit with an appropriate method or technique which suitable to extract the ontologies from Qur'anic text. With this ontology extraction tool, I hope can help more people to understand the true meaning of this language and teach the Qur'an.

ABSTRAK

Saat ini, ontologi memainkan peranan penting dalam teknologi web semantik untuk mendefinisikan konsep-konsep dan hubungan antara konsep-konsep. Pendekatan pembelajaran ontologi berbezakan mengikut jenis input seperti teks, kamus, pengetahuan, polisi, dan skim-skim semi-berstruktur hubungan. Pembelajaran ontologi dapat dijelaskan sebagai subtask untuk mengekstrakan maklumat dalam ontologi. Tujuan pembelajaran ontologi adalah untuk menggali konsep-konsep serta hubungan antara konsep-konsep daripada korpus atau data set yang tertentu. Dalam projek ini, saya akan menumpu pada pembelajaran ontologi daripada teks dengan menggunakan teks Qur'anic sebagai data input. Dua pendekatan yang akan digunakan untuk mengekstrak teks Qur'anic dalam projek ini adalah kaedah Alfonseca dan Manandhar dan kaedah Gupta dan Colleagues. Pada akhir projek, saya berharap dapat mengeluarkan satu kaedah atau teknik yang sesuai yang dapat membantu dalam menghasilkan applikasi atau alat yang boleh mengekstrakkan teks Qur'anic. Saya berharap penghasilan applikasi ini dapat membantu lebih banyak orang dalam memahami makna sebenarnya teks Qur'anic.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
DOE	Differential Ontology Editor
FCA	Formal Concept Analysis
XML	Extensible Markup Language
HTML	Hyper Text Markup Language
DTD	Document Type Definitions
UPM	Universidad Politécnica de Madrid
NL	Natural Language
NLP	Natural Language Processing
CICYT	Interministerial Comission of Science and Technology
TDIDF	Term Frequency, Inverse Document Frequency
PDF	Portable Document Format
UPM	University of Paris Nord
SPPC	Shallow Processing Production Center
SWP	Similarity with Parent Principle
SWS	Similarity with Siblings Principle
DWS	Difference with Siblings Principle
DWP	Difference with Parent Principle
GNER	General Name Entity Recognition
NER	Name Entity Recognition
WSD	Word-Sense Disambiguation

DAML	DARPA Agent Markup Language
RAM	Random-access memory
GB	Gigabyte
ADJP	Adjective Phrase
ADVP	Adverb Phrase
CONJP	Conjunction Phrase
NP	Noun Phrase
PP	Prepositional Phrase
IN	Preposition or Subordinating Conjunction
INTJ	Interjection
LST	List Marker
SBAR	Subordinating conjunction
JJ	Adjective
JJR	Adjective, Comparative
JJS	Adjective, Superlative
NN	Noun, Singular or Mass
NNS	Noun, Plural
NNP	Proper Noun, Singular
NNPS	Proper Noun, Plural
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Ontology learning is a subtask of information extraction. The goal of ontology learning is to extract relevant concepts and relations semi-automatically from a given corpus or other kinds of data sets to form ontology.

Ontologies play a key role in semantic web and technologies by defining concepts and relations among these concepts. Maedche and Staab (2001) distinguish different ontology learning approaches according to the types of used input. These ontologies learning are ontology learning from texts, ontology learning from dictionary, ontology learning from knowledge base, ontology learning from semi-structured schemata and ontology learning from relational schemata. Many significant studies had been held based on ontology learning topic and these useful techniques.

Based on these studies, ontology learning from texts consists of extracting ontologies by applying natural language analysis technique to texts. Among the most well-known approaches are pattern-based extraction (Morin, 1999; Hearst, 1992), association rules (Maedche and Staab, 2001), conceptual clustering (Faure et al., 2000), ontology pruning (Kietz et al., 2000), concept learning (Hahn et al., 2000) and lexico-syntactic patterns (Bruno Bachimont, 2002).

According to Asunción (Asunción et al, 2003), ontology learning from dictionary based its performance on the use of a machine readable dictionary to extract relevant concepts and relations among them. Ontology learning from a knowledge base aims to learn an ontology using an existing knowledge-base as source. Ontology learning from semi-structured data looks for eliciting an ontology from sources which have some predefined structure, such as XML schemas. Ontology learning from relation schemas aims to learn an ontology by extracting relevant concepts and relations from knowledge in a database.

Almost all the techniques require a good knowledge and infrastructure of natural language processing (NLP) to obtain effective results.

1.2 Problem Background

Ontology has become an important mean for structuring knowledge and building knowledge-intensive systems. Ontology also refers to the shared understanding of some domains of interest, which is often conceived as a set of concepts, relations, functions, axioms and instances. The aim of domain ontology is to reduce the conceptual and terminological confusion among the members of a virtual community of users that need to share electronic documents and information of various kinds.

Ontology engineers treat ontology learning as an useful method because this method helps them to construct ontology more easily. It also been regarded as one of the most important fields in the semantic web related to research work. So, ontology learning can defined as the set of method and techniques used for building ontology from scratch or, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources.

For the time being, there are many existing approaches with various kind of input. Among these existing approaches, none are using Qur'anic text as input. Qur'anic text is an input which has a very unique pattern. The Qur'anic text seems to have no beginning, middle, or end, it's nonlinear structure being akin to a web or net (Saidah, 2009). The textual arrangement is sometimes considered to have lack of continuity, absence of any chronological or thematic order, and presence of repetition. So, majority of the readers find that it is impossible for them to continue the reading, for they encounter a text unlike any they have ever read.

In the multilayered coherence of the Qur'anic text, all its themes emerge in short passages, creating an inimitable interplay between its imagery, oaths, parables, chronicles, warnings, and glad things. With so many elements of text coalescing, separating, reuniting, and reemphasising one another at numerous levels, the result can be a total incomprehensibility and confusion (Saidah, 2009). So, not much research had being made base on Qur'anic text as the input. Besides that, there is not much consensus within the ontology learning community on the concrete tasks, and this make a comparison of approaches difficult.

1.3 Problem Statement

As has been mentioned in the introduction, know that ontology learning has different categories according to the type of input data. In this project I want to study the ontology learning from text using Qur'anic text as the input. Through the study, I found that there are different methods and approaches that have been used in ontology learning from text. These methods and approaches are Kietz and colleagues' method (2000), Nobécourt approach (2000), Alfonseca and Manandhar's method (2002), Bachimont's method (2002), Missikoff and colleagues' method (2002) and others. These methods and approaches had been used in some of the ontology learning from text tools such as Text-To-Onto, TERMINAE, Welkin, Differential Ontology Editor (DOE), and OntoLearn.

In this project I going to compare the techniques based on the term, synonym and concept layers. Alfonseca and Manadhar's method and Gupta and Colleagues's approach are two of the methods which involve in this area. This is the reason why these two techniques been selected to do comparison. Both of these techniques are up to date and suitable for the time being.

Therefore, the problem statements are:

a) How well the existing Alfonseca and Manadhar's method and Gupta and Colleagues's approach can be use to extract ontology from text using Qur'anic text as input?

1.4 Project Objectives

a) To evaluate the application of two techniques, Alfonseca and Manandhar's method and Gupta and Colleagues's approach to extract ontology from text using Qur'anic text as the input.

b) To explore the possibility of enhancing or combining the two techniques to meet the need of ontology extraction from text using Qur'anic text as the input.

1.5 Scope of the Project

Ontological primitive can be organized into a few layers according to the increasingly complex subtasks within ontology learning to acquire them. On the other hand, ontology learning from text is a highly error-prone process. It seems clear that the success of ontology learning from text lies exactly in the combination of different technique. This helps to compensate for each other's erroneous predictions, thus increase the overall accuracy. So, this project is going to cover:

a) Ontology learning from text using the translated Qur'anic text from Yusof Ali's English translation.
b) Use of Alfonseca and Manandhar's method for ontology extraction.
c) Use of Gupta and Colleagues's approach for ontology extraction.
d) The performance measurement will defined by the search engine call recall.
e) Use the "Gold Standard for Islamic Knowledge Ontology (focus on Salah/Prayer)" (Saidah, 2009) as the benchmark.

1.6 Conclusion

In this chapter, we know ontology is a very wide area. There are a lot of things we can discover and explore. In this project, the area I interested is on the ontology learning from text because a lot of useful documents are in text form.

Nowadays there are a lot of techniques already been use on ontology learning from text. But different kinds of input data use different kind of technique. In order to distinguish a suitable technique to extract the input data, Qur'anic text for this project, I will do the comparison on the existing techniques. The result can help to figure out the leaking of the existing techniques and produce a new technique which suitable to this project. So, the research is needed to allow this project to be continuing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
DOE	Differential Ontology Editor
FCA	Formal Concept Analysis
XML	Extensible Markup Language
HTML	Hyper Text Markup Language
DTD	Document Type Definitions
UPM	Universidad Politécnica de Madrid
NL	Natural Language
NLP	Natural Language Processing
CICYT	Interministerial Comission of Science and Technology
TDIDF	Term Frequency, Inverse Document Frequency
PDF	Portable Document Format
UPM	University of Paris Nord
SPPC	Shallow Processing Production Center
SWP	Similarity with Parent Principle
SWS	Similarity with Siblings Principle
DWS	Difference with Siblings Principle
DWP	Difference with Parent Principle
GNER	General Name Entity Recognition
NER	Name Entity Recognition
WSD	Word-Sense Disambiguation

DAML	DARPA Agent Markup Language
RAM	Random-access memory
GB	Gigabyte
ADJP	Adjective Phrase
ADVP	Adverb Phrase
CONJP	Conjunction Phrase
NP	Noun Phrase
PP	Prepositional Phrase
IN	Preposition or Subordinating Conjunction
INTJ	Interjection
LST	List Marker
SBAR	Subordinating conjunction
JJ	Adjective
JJR	Adjective, Comparative
JJS	Adjective, Superlative
NN	Noun, Singular or Mass
NNS	Noun, Plural
NNP	Proper Noun, Singular
NNPS	Proper Noun, Plural
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Ontology learning is a subtask of information extraction. The goal of ontology learning is to extract relevant concepts and relations semi-automatically from a given corpus or other kinds of data sets to form ontology.

Ontologies play a key role in semantic web and technologies by defining concepts and relations among these concepts. Maedche and Staab (2001) distinguish different ontology learning approaches according to the types of used input. These ontologies learning are ontology learning from texts, ontology learning from dictionary, ontology learning from knowledge base, ontology learning from semi-structured schemata and ontology learning from relational schemata. Many significant studies had been held based on ontology learning topic and these useful techniques.

Based on these studies, ontology learning from texts consists of extracting ontologies by applying natural language analysis technique to texts. Among the most well-known approaches are pattern-based extraction (Morin, 1999; Hearst, 1992), association rules (Maedche and Staab, 2001), conceptual clustering (Faure et al., 2000), ontology pruning (Kietz et al., 2000), concept learning (Hahn et al., 2000) and lexico-syntactic patterns (Bruno Bachimont, 2002).

According to Asunción (Asunción et al, 2003), ontology learning from dictionary based its performance on the use of a machine readable dictionary to extract relevant concepts and relations among them. Ontology learning from a knowledge base aims to learn an ontology using an existing knowledge-base as source. Ontology learning from semi-structured data looks for eliciting an ontology from sources which have some predefined structure, such as XML schemas. Ontology learning from relation schemas aims to learn an ontology by extracting relevant concepts and relations from knowledge in a database.

Almost all the techniques require a good knowledge and infrastructure of natural language processing (NLP) to obtain effective results.

1.2 Problem Background

Ontology has become an important mean for structuring knowledge and building knowledge-intensive systems. Ontology also refers to the shared understanding of some domains of interest, which is often conceived as a set of concepts, relations, functions, axioms and instances. The aim of domain ontology is to reduce the conceptual and terminological confusion among the members of a virtual community of users that need to share electronic documents and information of various kinds.

Ontology engineers treat ontology learning as an useful method because this method helps them to construct ontology more easily. It also been regarded as one of the most important fields in the semantic web related to research work. So, ontology learning can defined as the set of method and techniques used for building ontology from scratch or, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources.

For the time being, there are many existing approaches with various kind of input. Among these existing approaches, none are using Qur’anic text as input. Qur’anic text is an input which has a very unique pattern. The Qur’anic text seems to have no beginning, middle, or end, it’s nonlinear structure being akin to a web or net (Saidah, 2009). The textual arrangement is sometimes considered to have lack of continuity, absence of any chronological or thematic order, and presence of repetition. So, majority of the readers find that it is impossible for them to continue the reading, for they encounter a text unlike any they have ever read.

In the multilayered coherence of the Qur’anic text, all its themes emerge in short passages, creating an inimitable interplay between its imagery, oaths, parables, chronicles, warnings, and glad things. With so many elements of text coalescing, separating, reuniting, and reemphasising one another at numerous levels, the result can be a total incomprehensibility and confusion (Saidah, 2009). So, not much research had being made base on Qur’anic text as the input. Besides that, there is not much consensus within the ontology learning community on the concrete tasks, and this make a comparison of approaches difficult.

1.3 Problem Statement

As has been mentioned in the introduction, know that ontology learning has different categories according to the type of input data. In this project I want to study the ontology learning from text using Qur'anic text as the input. Through the study, I found that there are different methods and approaches that have been used in ontology learning from text. These methods and approaches are Kietz and colleagues' method (2000), Nobécourt approach (2000), Alfonseca and Manandhar's method (2002), Bachimont's method (2002), Missikoff and colleagues' method (2002) and others. These methods and approaches had been used in some of the ontology learning from text tools such as Text-To-Onto, TERMINAE, Welkin, Differential Ontology Editor (DOE), and OntoLearn.

In this project I going to compare the techniques based on the term, synonym and concept layers. Alfonseca and Manadhar's method and Gupta and Colleagues's approach are two of the methods which involve in this area. This is the reason why these two techniques been selected to do comparison. Both of these techniques are up to date and suitable for the time being.

Therefore, the problem statements are:

a) How well the existing Alfonseca and Manadhar's method and Gupta and Colleagues's approach can be use to extract ontology from text using Qur'anic text as input?

1.4 Project Objectives

a) To evaluate the application of two techniques, Alfonseca and Manandhar's method and Gupta and Colleagues's approach to extract ontology from text using Qur'anic text as the input.
b) To explore the possibility of enhancing or combining the two techniques to meet the need of ontology extraction from text using Qur'anic text as the input.

1.5 Scope of the Project

Ontological primitive can be organized into a few layers according to the increasingly complex subtasks within ontology learning to acquire them. On the other hand, ontology learning from text is a highly error-prone process. It seems clear that the success of ontology learning from text lies exactly in the combination of different technique. This helps to compensate for each other's erroneous predictions, thus increase the overall accuracy. So, this project is going to cover:

a) Ontology learning from text using the translated Qur'anic text from Yusof Ali's English translation.
b) Use of Alfonseca and Manandhar's method for ontology extraction.
c) Use of Gupta and Colleagues's approach for ontology extraction.
d) The performance measurement will defined by the search engine call recall.
e) Use the "Gold Standard for Islamic Knowledge Ontology (focus on Salah/Prayer)" (Saidah, 2009) as the benchmark.

1.6 Conclusion

In this chapter, we know ontology is a very wide area. There are a lot of things we can discover and explore. In this project, the area I interested is on the ontology learning from text because a lot of useful documents are in text form.

Nowadays there are a lot of techniques already been use on ontology learning from text. But different kinds of input data use different kind of technique. In order to distinguish a suitable technique to extract the input data, Qur'anic text for this project, I will do the comparison on the existing techniques. The result can help to figure out the leaking of the existing techniques and produce a new technique which suitable to this project. So, the research is needed to allow this project to be continuing.