Dedication

To my soul mate, Joe and my family

"In science, the credit goes to the man who convinces the world, not to the man to whom the idea first occurs" by Sir William Osler

"*We can only see a short distance ahead, but we can see plenty there that needs to be done*" by Alan Turing

Acknowledgements

Graduating with a Master's degree through research in any subject is a non-trivial task. With the support in certain aspects from the people mentioned below, life has been made a little easier. First of all, I would like to thank my supervisor Associate Professor Goh Ong Sing for giving me the opportunity to pursue this project. Going through this research alone is a very daunting task but with the companionship and assistance from other individuals in the University Press, things went along much smoother. Not leaving out people from the Faculty of Information and Communication Technology, Study Leave Unit and Postgraduate Study Centre, a very big hug for everyone. A special thanks to Dr. Mohammad Ishak Desa, Dr. Shahrin Sahib and Dr. Nanna Suryana. Thank you for always being there for my questions. I would also like to attribute whatever academic success that I have achieved to the Government of Malaysia. I would not have come this far without the opportunity and financial support by the Government since the very first day I pursued my undergraduate studies. Not forgetting, my family back in Penang. There are no words that could express my heartfelt-gratitude for the day-in day-out prayers and support from afar. With your blessings, I'm bound to achieve greatness from the very first day I left home. As for my pseudo-family members in Kuala Lumpur namely Shu Yueh, Ah Hoe, Sau Lun, Mr. and Mrs. Nakayama and others, a big thank you. On a more personal level, I would like to give a million hugs and kisses to my soul mate Liaw Sau Joe for sticking with me across ten thousand miles of ocean and being extremely forbearing with me when I get irritable from the pressure of this research. Finally, my appreciation goes to the city of Melaka for being such a great place to do research.

WILSON WONG

June 2005

Table of Contents

CHAPTER 2 BACKGROUND INFORMATION

CHAPTER 3 REVIEW OF PROMINENT QUESTION ANSWERING SYSTEMS

List of Figures

List of Tables

List of Abbreviations

TREC	*T*ext *RE*trieval *C*onference, an annual conference and competition co-sponsored by the National Institute of Standards and Technology, and U.S. Department of Defense.
BBN	*B*olt, *B*eranek and *N*ewman, the last names of the three founders of BBN Technologies and the original name of the company.
SHRDLU	The name comes from the fact that the frequency order of letters in English is ETAOIN*SHRDLU*. As a result, the arrangement of the keys on Linotype typesetting machines was ETAOIN on the first column and *SHRDLU* the second.
QA	*Q*uestion *A*nswering
NLP/NLU	*N*atural *L*anguage *P*rocessing/*N*atural *L*anguage *U*nderstanding
IR	*I*nformation *R*etrieval
NaLURI	*Na*tural *L*anguage *U*nderstanding and *R*easoning for *I*ntelligence
MUSE	*MU*lti-*S*ource *E*ntity, an information extraction system to perform named entity recognition on diverse types of text with minimal adaptation.
FASTUS	*F*inite *S*tate *A*utomata-based Text *U*nderstanding *S*ystem, an information extraction system by Stanford Research Institute.
LISP	*LIS*t *P*rocessing, a functional programming language family originally developed as a practical computation model and later became the favored artificial intelligence research language.
HTML	*H*yper*T*ext *M*arkup *L*anguage
XI	*X* stands for cross-classification and *I* for inheritance

Related Publications

Wilson Wong, Shahrin Sahib & Ong-Sing Goh. *Evaluation of Response Quality for Heterogeneous Question Answering Systems*. Accepted to the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), University of Technology of Compiegne, France, 19-22 September 2005.

Wilson Wong, Shahrin Sahib & Ong-Sing Goh. NaLURI: Question Answering with Natural Language Understanding and Network-based Advanced Reasoning. Submitted for review to the International Conference on Intelligent Technologies (InTech 2005) in July 2005.

Wilson Wong, Shahrin Sahib & Ong-Sing Goh, "Question Answering Approaches in the 21st Century: A Survey", Submitted to the ACM Computing Survey for review in April 2005.

Wilson Wong, Shahrin Sahib & Ong-Sing Goh. *Response Quality Evaluation in Heterogeneous Question Answering System: A Black-box Approach.* Unpublished manuscript, Kolej Universiti Teknikal Kebangsaan Malaysia, 2005.

Wilson Wong, Halizah Basiron, Shahrin Sahib and Ong-Sing Goh, "Intelligent Responses through Network-based Answer Discovery with Advanced Reasoning", To appear in the Proceeding of the IASTED International Conference on Computational Intelligence (CI 2005), Alberta, Canada, 4-6 July.

Wilson Wong, Ong-Sing Goh, Mohamad-Ishak Desa and Shahrin Sahib, "Online Cyberlaw Knowledge Base Construction Using Semantic Network", Ed. M.H. Hamza. In Proceeding of the IASTED

International Conference on Applied Simulation & Modeling (ASM 2004), Rhodes, Greece, 28-30 June 2004. ISBN: 0-88986-401-2, pp. 347-352.

Wilson Wong, Ong-Sing Goh and Mokhtar Mohd-Yusof, "Syntax Preprocessing in Cyberlaw Web Knowledge Base Construction", Ed. M. Mohammadian. In Proceeding of the International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA 2004), Gold Coast, Australia, 12-14 July 2004. ISBN: 174-088-1893, pp: 174-184.

Abstract

The complexity of natural language and the open-domain nature of the World Wide Web have caused modern-day question answering systems to rely only on information retrieval techniques and shallow natural language processing tasks. This approach has brought about serious drawbacks namely restriction on the nature of question and response. This restriction constitutes the first problem addressed by this research. Through recent academic works, many researchers have begun to acknowledge the problem and agreed that the solution comes in the form of a new approach based on natural language understanding and reasoning in a knowledge-based environment. Due to the infancy stage of this new approach and practical consideration, the current practices vary greatly and are mostly based on only low-level natural language understanding, minimalist representation formalism and conventional reasoning approach without advanced features. As a result, not only were these systems found to be inadequate to solve the first problem but have also created the second problem, that is the limitation to scale across domains and to real-life natural language text. This research hypothesized that a practical approach in the form of a solution framework which combines full-discourse natural language understanding, powerful representation formalism capable of exploiting ontological information and reasoning approach with advanced features, will solve both the first and second problem without compromising practicality factors. The solution framework is implemented as a system called "*Natural Language Understanding and Reasoning for Intelligence*" (NaLURI). More importantly, two evaluations and their results are presented to demonstrate that the inclusion of more demanding features into a question answering system will not only allow for a wider range of questions and better response quality, but does not affect the response time, hence approving the hypothesis of this research.

Abstrak

Kerumitan dalam pengendalian bahasa tabii serta sifat-sifat domain terbuka dalam Jaringan Sedunia menyebabkan sistem soal jawab zaman moden tiada pilihan selain daripada bergantung sepenuhnya pada teknik carian maklumat berasaskan kata kunci serta pemprosesan bahasa tabii yang sempit. Pendekatan seperti ini menyebabkan kebolehan sistem untuk menjawab soalan terhad. Batasan ini merupakan masalah pertama yang cuba diselesaikan dalam penyelidikan ini. Melalui penulisan akademik yang terkini, para penyelidik mula mengakui kewujudan masalah tersebut dan bersetuju bahawa jalan penyelesaiannya terangkum dalam satu pendekatan baru yang berasaskan teknik-teknik pemahaman bahasa tabii dan taakulan yang berasaskan pengetahuan. Disebabkan oleh pelbagai pertimbangan dari segi isu-isu praktikal, amalan sedia ada kebanyakannya berasaskan hanya pemahaman bahasa tabii sempit, perwakilan formalisma yang minimum dan pendekatan taakulan konvensional tanpa ciri-ciri canggih. Akibatnya, sistem-sistem ini bukan sahaja didapati tidak sesuai untuk menyelesaikan masalah pertama dalam penyelidikan ini, malah amalan-amalan tersebut telah membawa kepada masalah kedua iaitu pembatasan dalam merentasi domain dan teks bahasa tabii yang sebenar. Penyelidikan ini mencadangkan satu pendekatan praktikal dalam bentuk rangka kerja yang akan menyelesaikan masalah pertama serta masalah kedua melalui pemahaman bahasa tabii dan wacana yang lengkap dengan perwakilan formalisme yang optimum seperti rangkaian semantik yang mampu mengeksploitasi maklumat ontologi untuk membolehkan pengenalan ciri-ciri canggih ke dalam pendekatan taakulan. Rangka kerja penyelesaian tersebut direalisasikan melalui "Natural Language Understanding and Reasoning for Intelligence" (NaLURI). Dua penilaian bagi menguji isu-isu praktikal juga dibuat bagi menunjukkan bahawa pengenalan ciri-ciri canggih ke dalam sistem soal jawab bukan sahaja akan meningkatkan kualiti soal-jawab, malah tidak akan menjejaskan masa tindak balas.

Chapter 1

Introduction

1.1 The Past, Present and Future of Question Answering

The common idea in question answering is to be able to provide responses to questions written in natural language (i.e. English) by finding the answer in some sources (e.g. web pages, plain texts, knowledge bases) or by generating explanations in the case of failures (i.e. which is only possible through intelligent approaches). Unlike information retrieval applications like web search engines, the goal is to find a specific answer (Lin *et al.,* 2003) rather than flooding the users with documents or even best-matching passages as most information retrieval systems currently do. With the increase in the number of online information seekers, the demand for automated question answering systems has risen accordingly. There are many ways of looking at question answering depending on the approaches towards the various dimensions (Hirschman & Gaizauskas, 2001). The different dimensions include question, response, technique, information source, domain and evaluation as depicted in Figure 1.1.

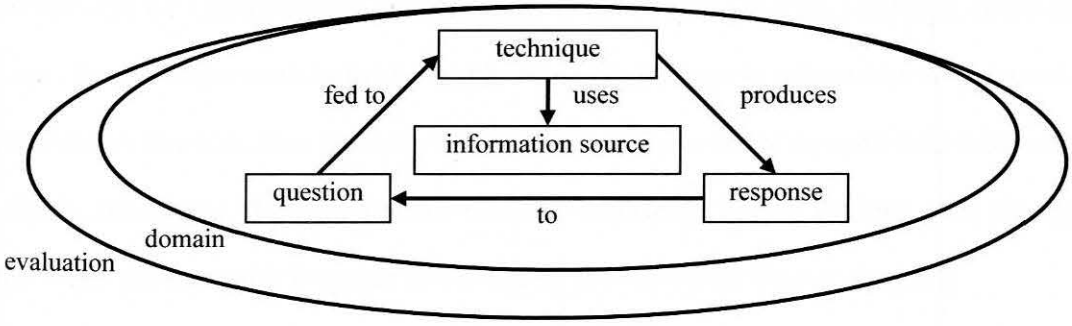

Figure 1.1: Dimensions of question answering

As different type of questions pose dissimilar level of challenges, the type of questions supported by a question answering system can be used to determine the strength of the system. Questions can be formulated in five ways (Moldovan *et al.*, 2002) namely factual questions (e.g. *"Where is Kuala Lumpur"*), questions requiring simple reasoning (e.g. *"Why did the accident happen?"*, synthesis-based questions (e.g. *"What are the daily activities of the victim a week before he was murdered?"*), dialogue-based questions (e.g. *"Who is the defendant in that case?"*) and finally, speculative questions (e.g. *"Is the idea of raising fuel price justified?"*).

Unlike questions, there are no definitions of what encompasses an exact response. Clearly an answer has to be correct to be of any use, but this still leaves a lot of scope for different systems to present the same answer in many different ways. Nevertheless, from the techniques employed for producing answers, one can almost predict the structure of the response. Systems that use unstructured texts as their source of answers for example, will usually return a short extract from the text as responses. The major question with such systems is how long the returned answer should be.

The following two dimensions, namely technique and information source, are the common aspects used to differentiate between the various types of question answering systems since the beginning of the question answering era. The technique used is usually highly related to the type of information source used. If the information source is free-text, then the technique will most likely be based on some information retrieval approach. On the other hand, if the information source is knowledge base or database, then the approach will be either logic-based or some language sanctioned by the knowledge base or database.

Domain is one of the dimensions that determines the focus or direction of a question answering system. Open-domain question answering practices techniques based on probabilistic measures and has a wider

range of information source. It is very likely that the techniques are more logic-based and well-founded with relatively limited sources for question answering that focuses on certain domains as compared to open-domain. A domain-oriented question answering system deals with questions under a specific domain and can be seen as a richer approach because natural language processing systems can exploit domain knowledge and ontologies. Advanced reasoning such as providing explanation for answers and generalizing questions is not possible in open-domain systems. Open-domain question answering systems need to deal with questions about nearly everything and it is very difficult to rely on ontological information due to the absence of wide and yet detailed world knowledge. On the other hand, these systems have much more data to exploit in the process of extracting the answers (Clarke *et al.,* 2001).

The last dimension, which is evaluation, can be rather subjective especially when dealing with different types of natural language systems in different domains. Surprisingly, the literatures on evaluation are relatively sparse given its state of importance and are mostly available in form of evaluating general natural language systems (King, 1996). It is easy to evaluate systems in which there is a clearly defined answer, however, for most natural language questions there is no single correct answer. Only the question answering systems based on shallow natural language processing and information retrieval that have the corpora and test questions readily available for example, can use recall and precision as evaluation criteria. The question answering track of TREC is a good example (Voorhees, 2003). The task of evaluating the system can be more subjective and difficult for other domain-oriented question answering.

1.1.1 Domain-Oriented Question Answering in Cyberlaw

Like many other question answering systems based on natural language understanding and reasoning, the choice of domain tend to be focused in certain areas. Some examples are the question answering system for biomedicine by Zweigenbaum (2003), question answering system for weather forecast by Chung *et al.*

(2004) and question answering system for the tourism domain by Benamara (2004). These domain-oriented question answering systems not only act as real-life example of the success of the natural language understanding and reasoning approach, but the systems itself proved to be a worthwhile attempt in providing intelligent assistant for domain experts.

While there is already a number of domain-oriented question answering systems out there based on a wide range of techniques in natural language understanding and reasoning covering unique domains, a system for the domain of Cyberlaw is yet to exist. This late emergence can be attributed to the fact that unlike other conventional domains which have been around for quite some time such as medicine, tourism and other branches of science and economics, Cyberlaw only surfaces during the boom of the various activities related to the World Wide Web like e-commerce, e-banking, etc. The description of an initial effort towards a question answering system for the Cyberlaw domain is presented by Wong *et al.* (2004a).

Activities involving the use of Internet and information technology have increased tremendously over the years. This is particularly true as more companies and countries are attempting to use technologies like e-commerce, e-marketing, e-government, telemedicine and many more to achieve better efficiency and a paperless environment. With the increasingly important role played by Information Technology and the Internet, security threats and human misconducts will follow suit, creating a whole new paradigm of online information on Cyberlaw (Zahri & Ahmad-Nasir, 2003). In Malaysia alone for example, the year 2004 sees a rise in cyber crime from 856 cases in January to 1393 cases in December (Anon., 2004). As more parts of our life become acquainted to technology, a unified and easily accessible source of knowledge on Cyberlaw will be a valuable asset to legal practitioners, legal students, academicians, enthusiasts and others alike.

Despite the hype surrounding the field of Cyberlaw, there is yet to be any definite description of the scope of Cyberlaw but nonetheless, Grossman (1999) offers a credible argument which describes Cyberlaw as a fusion between computer law, Internet law, e-commerce law, intellectual property law, traditional contract law, criminal law, litigation of technology related disputes and much more. In a nutshell, Cyberlaw concerns what technology does or can do. To illustrate this point further, the following examples should be considered. We can earn money with technology, thus we have e-commerce law and traditional contract law in the mix. We can find new ways to digitally lose privacy online and we may relate that to an Internet law issue. Technology can also give us new ways to commit crimes and infringe copyrights, so we have to include criminal law and intellectual property law.

1.1.2 The Early Days of Question Answering

Some of best-known question answering systems in the early days were designed to provide natural language front ends to databases. These systems operated under extremely limited domains. Some of the best-known were BASEBALL (Green *et al.*, 1963), LUNAR (Woods, 1973) and LIFER (Hendrix *et al.*, 1978). The BASEBALL system was designed to answer questions about baseball games which had been played in the American league over a single season, while LUNAR was designed to enable lunar geologists to conveniently access, compare and evaluate the chemical analysis data on lunar rock and soil composition that was accumulating as a result of the Apollo moon mission. Similarly, LIFER employs a front-end natural language interface to connect to databases allowing users to ask questions about United States' navy ships. The other two systems that had an equal share of fame during the 1970s are SHRDLU by Winograd (1972) and GUS (Bobrow *et al.*, 1977). GUS was designed to simulate a travel advisor and has access to a database containing limited information about airline flight times. SHRDLU, the better known between the two, was created for manipulating geometrical blocks in a confined world. The

difference between LUNAR and systems like SHRDLU and GUS is the latter's ability to carry out a dialogue.

Problems started to surface when some researchers during the early days attempted to apply their limited natural language interface to more general English text. This is mainly due to the expensive requirements for understanding and reasoning, and given the state-of-the-art technology during the early days, their approaches were only feasible in very limited domains (Hirschman & Gaizauskas, 2001). Moreover, the researchers in the field of natural language understanding during the early days were only starting to solve isolated problems in the lower level of linguistic analysis (Mueller, 1999). Due to the limiting factors in understanding natural language during the late 70s, a shift in the question answering approach began to take place. Rather than focusing on natural language understanding, researchers have opted for approaches that are known to be effective during that time, allowing them to move beyond domain restriction and exploit open-domain, natural language information (Fischer, 2003). These systems employ what is known to work best with free-text documents, namely information retrieval which typically relies on statistical methods to process the keywords in a query and calculate a relevance ranking. This ranking is used to search an open domain of texts to return a list of documents that possibly contain an answer. The process requires little or no linguistic knowledge because it relies primarily on word frequencies. However, many researches like Cardie *et al.* (2000) have examined and agreed that even weak linguistic knowledge can significantly improve the results of a question answering system. Thus, the marriage between shallow natural language processing and information retrieval for open-domain information marks the beginning of the modern-day question answering systems.

One of the first few systems that exhibit the modern-day question answering characteristics is MURAX by Kupiec (1993). MURAX employs an encyclopedia as the open-domain source and an information retrieval system for accessing it. Shallow linguistic analysis is performed using a part-of-speech tagger

and finite-state recognizers for matching syntactic patterns. FAQ Finder by Burke *et al.*, (1997) is also another type of modern-day question answering system that works on frequently-asked questions found in newsgroups using the SMART information retrieval systems (Buckley, 1985). In 1999, many more systems from this camp like YorkQA (Alfonseca *et al.*, 2001) began to appear when TREC-8 and its subsequent conferences provided large corpuses as the underlying source for developing and evaluating question answering systems.

As the demand for a better search and retrieval solution to the ever-growing World Wide Web increased, researchers began to look into the exploitation of information on the World Wide Web as the source for question answering. With the wide availability of web search engines, modern-day question answering systems using classical information retrieval were very quickly extended to the World Wide Web. Some of the well known systems that exploit the web search engines are like Webclopedia (Hermjakob, 2001), AnswerBus (Zheng, 2002b) and MULDER (Kwok *et al.*, 2001). This modern-day question answering approach has indeed lived up to its name and has flourished until the present day. Consequently, many of the current researches tackle the problem of question answering from the dimension where the technique is based on the marriage of shallow natural language processing and information retrieval, and the information source using either TREC corpora or the World Wide Web.

1.1.3 Limitations of Modern-Day Question Answering

Through a review of the existing question answering systems based on shallow natural language processing and information retrieval which is discussed in Chapter 3, it can be seen that the ubiquitous ways of accessing information on the World Wide Web is web search engines. This has provided many modern-day question answering systems with an easy way out. These modern-day question answering systems have become too reliant on the use of web search engines and the idea that endless source of