New Generation Computing, 26(2008)3-22
Ohmsha, Ltd. and Springer

Query Based Intelligent Web Interaction with Real World Knowledge

Ong Sing GOH, Chun Che FUNG and Kok Wai WONG
School of Information Technology
Murdoch University, South Street, Western Australia 6150
{os.goh,l.fung,k.wong}@murdoch.edu.au

Received 20 March 2007
Revised manuscript received 25 June 2007

Abstract This paper describes an integrated system based on open-domain and domain-specific knowledge for the purpose of providing query-based intelligent web interaction. It is understood that general purpose conversational agents are not able to answer questions on specific domain subject. On the other hand, domain specific systems lack the flexibility to handle common sense questions. To overcome the above limitations, this paper proposed an integrated system comprises of an artificial intelligent conversation software robot or chatterbot, called Artificial Intelligence Natural-language Identity (hereafter, AINI), and an Automated Knowledge Extraction Agent (AKEA) for the acquisition of real world knowledge from the Internet. The objective of AKEA is to retrieve real world knowledge or information from trustworthy websites. AINI is the mechanism used to manage the knowledge and to provide appropriate answer to the user. In this paper, we compare the performance of the proposed system against two popular search engines, two question answering systems and two other conversational systems.

Keywords: Automated Knowledge Extraction Agent (AKEA), Artificial Intelligent Natural Language Identity (AINI), Natural Language Processing (NLP), Artificial Intelligence (AI).

§1 Introduction

Traditional media such as television, radio, newspapers and magazines play an important role in reporting and providing the latest information on current events. However, nearly all of these media are linear and unidirectional in nature. In other words, they do not provide interactive communication nor can they answer any query from the users. Most of these traditional means of media also do not link the answers to the source or reference of the information. Hence, viewers or readers may require additional effort in locating or verifying the information.

On the other hand, the Internet, in particular, the World-Wide-Web (www), coupled with multimedia, browser and web service technologies, presents a powerful way of communication. If the system is designed properly, it could provide high level of interactivity between the users and the computer. This is clearly a major step in the advancement over the traditional means of communication. One of the possible applications of such system is to provide answers to queries on specific topics or knowledge domains.

Consider situations such as outbreak of diseases, natural disasters and terrorist attacks, they have caused much miseries, fear and confusion around the world. Examples of such crisis are the Severe Acute Respiratory Syndrome (SARS), bird flu, September 11, earthquakes and tsunamis. In such times, a lot of people will be eager to know as much information as possible. This is where the traditional media may be found inefficient as they are limited with constraints such as page number and air-time. They may also be unable to source all the information from around the world. In addition, people such as managers and decision-makers, frontline specialists or emergency services personnel, and concerned citizens who are directly or indirectly involved in the situations, require to be better informed of the situation and the development.

To this end, the Internet could have a major role to play as an essential communication channel in providing an intelligent interactive interface for the users. A form of a "global crisis communication system" could be used to provide a natural language query-and-answer system. This system should be capable of providing relevant responses to queries from the users in a conversational manner. In this paper, we describe part of a project which is currently under development. The aim of the project is to develop an intelligent conversation agent called AINI to answer domain specific questions as well as open-domain (or common sense) questions. In this report, we have used the subject – bird flu as the domain knowledge of interest to demonstrate the feasibility of the developed system. The key contribution described in this paper is the integration of the common sense knowledge and domain specific knowledge in the form of a "knowledge matrix." The system is based on a layered and modular design, and the answers for the queries are searched from these modules. The proposed system is based on an intelligent conversation agent called AINI and an AKEA. The objective of AKEA is to retrieve real world knowledge or information from trustworthy websites whereas AINI is the mechanism used to manage the knowledge and to provide appropriate answer to the user. Descriptions of the system are given in the subsequent sections.

The prototype system is compared to two popular search engines – Google and Yahoo!, two question answering systems – START and AskJeeves and two conversational systems – ELIZA and ALICE. While it is understood that the purpose of those systems are different, the objective of the comparison is to demonstrate the need of the proposed system described in this paper. This also shows that the proposed system is capable to provide appropriate answers by integrating open domain and domain specific knowledge into one conversation agent architecture.

§2 AINI's Conversational Agent Architecture

The AINI architecture is shown in Fig.1. The architecture has been reported in previous publications by the authors.[19,20] Basically, the AINI engine comprises a number of knowledge modules in the Data Layer. It has the ability to communicate with three layers: the Application Layer, the Data Layer and the Client Layer. The Client Layer is capable to communicate with the user via different channels such as Web browser, Mobile Browser, WAP Browser and GSM Interface. It can carry on multiple independent conversations at the same time. AINI's knowledge bases are located within the Data Layer. There are a number of modules supporting the Application Layer which governs the manipulation and searching of the answers. The modules use plug-in principles that can quickly be augmented with domain knowledge for specific purposes.

Originally, this research project involves the establishment of an embodied conversational agent (ECA) based on an AINI architecture.[20] The prototype system is designed specifically for the web and mobile technology as shown in the Client Layer in Figure 1. The complete software agent can be considered as a multi-domain knowledge system with multimodal human-computer communication interface. The query and answer between the user and the computer are communicated via the common protocol TCP/IP. AINI is designed to engage the user with focus on the chosen subject topic. In this particular application, the topic is on the possible pandemic virus, H5N1. AINI communicates with the user in natural language via typed messages. The system is also capable to reply in text-prompts or Text-to-Speech Synthesis together with appropriate facial-expressions on the displayed object which can be an animated avatar or a human face.

As illustrated in Fig.1, AINI employs an Internet three-tier, thin-client architecture that may be configured to work with any web application. It comprises a client layer, an application layer and a data server layer. The hybrid architecture provides features of multimodal interface (Client Layer), multilevel natural language query (Application Layer) and multiple knowledge bases (Data Layer). The process of communication and answering is as follows. Given a question, AINI first performs a question analysis by extracting pertinent information to be used in query formulation, such as the Noun Phrases (NPs) and Verb Phrases (VPs) using the MINIPAR parser.[26] MINIPAR is a broad-coverage parser for the English language. An evaluation with the SUSANNE corpus shows that MINIPAR achieves about 88% precision and 80% recall with respect to dependency relationships. In our experiment by using corpus extracted by Automated Knowledge Extraction Agent (AKEA),[18] MINIPAR parser is capable to parses nearly 500 words per second on a Dell Precision PWS380 Server 3GH with 1GB memory.

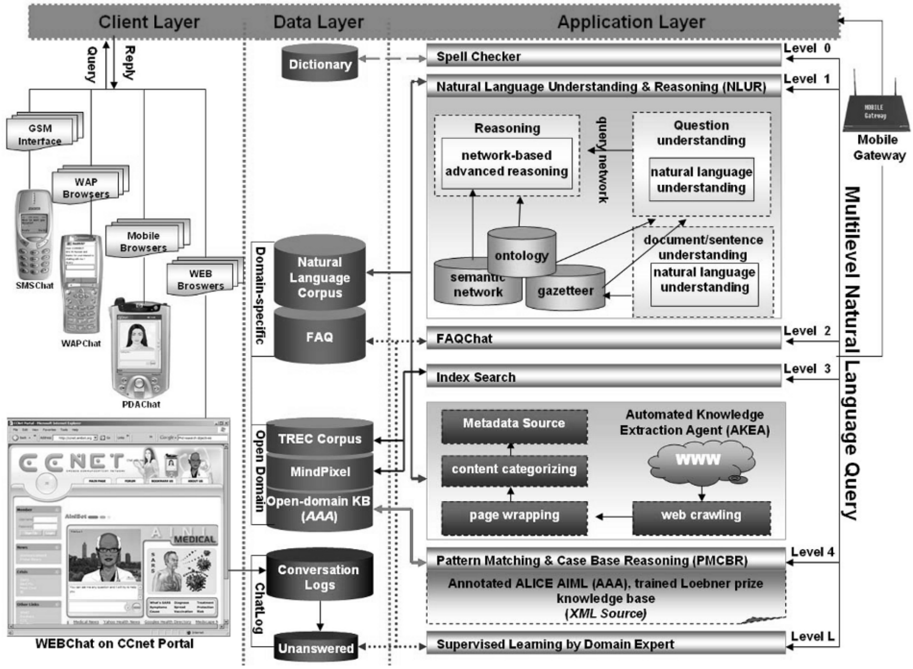


Fig. 1 AINI's Conversational Agent Architecture

2.1 Client Layer

The user interface resides in the thin-client layer and it supports web-based and mobile service interface. For the web-based interface, the system is based on Multimodal Agent Markup Language (MAML) interpreter in order to handle the user interface. MAML is a prototype multimodal markup language based on XML that enables animated presentation agents or avatars. It involves a virtual lifelike talking 3D agent character designed to carry out a more natural conversation with the user.

For mobile devices, due to their small screens, there are limitations on the amount of information that can be presented at one time. Reading large amounts of information from such devices can require excessive amount of scrolling and concentration. To reduce distraction, interactions, and potential information overload, a better way of presenting information might be through multilevel or hierarchical mechanisms.[9] A chat mode interface will suit the multilevel mechanism nicely and this will be a better solution for mobile service. In addition, current wireless network service vendors are now providing a wide bandwidth telephone network, known as 3G communication.[23] This development has led to the increasing adoption of smartphone as a client in the traditional distributed systems. On such system, the proposed system will require the Mobile Flash Player.[2] For WAP services, the application was embedded WAP browsers from vendors such as Openwave*[1] and Nokia.*[2]

2.2 Application Layer

All communication with AINI takes place through typed text messages and they are processed based on natural language understanding and reasoning. AINI's engine implements its decision making network based on the information it encounters in the six levels of knowledge modules. The input and output of each module is an XML-encoded data structure that keeps track of the current computational state. The knowledge modules can be considered as transformations over this XML data structure. The system accepts questions or queries from the users and it processes the queries based on the information contained in AINI's knowledge bases. The system is implemented by open-source architecture employing a Kannel Mobile gateway, PHP, Perl scripting language, Apache Server and knowledge base stored in a MySQL server. In this server, it handles the process of the queries. Here, one or more application servers are configured to compute the dialogue logic through the multilevel natural language query algorithm. In this layer, it is based on a goal-driven or top-down natural language query (NL-Query) approach which is similar to the way that humans process their language. As indicated by literature in the field of Natural Language Processing (NLP), the top-down approach is by far the best approach. As shown in Fig. 1, the top-down NL-query approach consists of six levels of queries, namely Spell Checker (Level 0), Natural Language Understanding and Reasoning (NLUR) (Level 1), FAQChat (Level 2), Metadata Index Search (Level 3), Pattern Matching and Case-based Reasoning (PMCBR) (Level 4) and Supervised Learning (Level 5). All these have been discussed in reference.[17)]

2.3 Data Layer

The data layer serves as storage for data and knowledge required by the system. This is where the AINI bot's conversational knowledge bases are stored. In the proposed approach, the architecture can be considered as a Domain Matrix Knowledge Model[16)] for the support of the conversational system. It is well understood that true intelligent action requires large quantities of knowledge. Such reservoir of knowledge is harvested from the Internet and deployed in the domain matrix knowledge bases architecture. This form the basis for the construction of large-scale knowledge bases to be used as the engine for the intelligent conversation systems.

These databases established so far are Dictionary, Domain-Specific knowledge bases, Open Domain knowledge bases and Chatlog (conversation logs and unanswered question logs). As mentioned previously, the first step taken by AINI is to perform a spell check to eliminate possible wrong spellings. The spell checker is located at the Application Layer and the dictionary contains the database of words. The spell checker used in this study is called ***ispell.*** Ispell is an open-source program that is used to correct spelling and typographical

*[1] http://developer.openwave.com

*[2] http://www.nokiausa.com/support/software

errors in a file. Ispell was first run on TOPS-20 systems at the MIT-AI lab. Many people have contributed dictionaries to allow ispell to be used with their native languages. At least 50 dictionaries have already been established. As far as this project is concerned, only English is used. However, it can be anticipated that AINI has the potential to be used as a true intelligent web interaction with users who are using different languages. Combining the dictionary and the spell checker in the application layer, this combination can be considered as the Level 0 knowledge.

In terms of the domain-specific knowledge databases, it can be observed in Fig. 1 that the domain specific knowledge modules are made up of a Natural Language Corpus and a FAQ (Frequent Asked Questions) module. The Natural Language Understanding & Reasoning (NLUR) module in the application layer is used to gather data to be stored in the Natural Language Corpus. In addition, the AKEA is used to extract information from the web and in particular trusted web sites. The architecture of AKEA has been reported previously and the details can be found in reference.[18] A description of its operation will also be given in the subsequent section. This forms the Level 1 knowledge.

The FAQ (Frequent Asked Questions) module is connected to the FAQChat module in the Application Layer. It is used to provide information based on the FAQ sections from websites and from the actual chat logs. In other words, the conversations and answers given by the user to the system are also used to enhance the information in the FAQ module. This forms the Level 2 knowledge. In this study, the bulk of the information was collected during the height of the SARS epidemic in 2003[19] and the Bird Flu pandemic crisis in 2005.[21] However, in this paper, the domain specific knowledge is only focused on Bird Flu.

The Open-domain knowledge modules in the data layer are based on TREC Corpus, Minipixel and the Open-domain Knowledge Base (AAA). They are further supported by the Conversation Logs which store the previous conversation for extraction of answers from the unanswered questions. It is assumed that such answers could be used in future interaction with the users. Details of these open-domain knowledge modules are described in the Section 3.1 and they form the subsequent levels of knowledge in the Application Layer.

§3 Domain Knowledge Matrix Model

In this paper, Bird Flu pandemic is the domain-specific knowledge used as an example to illustrate the application of this proposal. Research and information on H5N1 pandemic have become increasingly important as the pandemic may have dire global implications. Wall Street Journal Online[7] predicted that this pandemic could be worse than the one in 1918 which killed at least 20 million people. In addition, the World Health Organization estimates the H5N1 virus could infect up to 30 percent of the world's population. Shigeru Omi, a WHO official, also warned that an estimation of 2 to 7 million deaths are "conservative" and that the maximum figure could be as high as 50 million.[29]

The AINI's domain knowledge matrix model incorporates several knowledge subjects. This is analogous to the consultation of expertise knowledge from multiple experts. For example, a *sales* domain knowledge should contain expertise or knowledge on how to improve sales. However, as a sales person, he or she is expected to have a wide range of common sense knowledge and the ability to engage the potential customer. Hence, the intelligent system should also incorporate open-domain knowledge to handle general or generic questions. In here, the open-domain knowledge is not necessary the common sense knowledge which is assumed to be possessed by everyone. The open-domain refers to the accessibility of the knowledge. Hence, they are still categorized according to knowledge domain. By including multiple domain knowledge bases with AINI's single domain knowledge, the proposed AINI will be able to hold "meaningful" conversations with the users. In this proposed system, the open-domain and domain-specific knowledge are pre-defined in the Data Layer. These modules are used to support the various knowledge levels at the Application Layer. Depending on the user's input, the agent will respond or switch from one Level to another. While the system is capable to communicate with the user beyond the domain knowledge, there are cases that the system will exhaust its capability to answer the queries. In such case, the system will attempt to divert the focus back to the current topic of interest by responding with some predefined random statements. The purpose is to direct the users' back to the system's domain-specific state. Hence AINI will attempt to "cycle" between the 6 levels of information processing within the Application Layer supported by the various knowledge modules in the Data Layer.

A way to view the proposed Domain Knowledge Matrix Model is given in Fig. 2. In this approach, the knowledge base of the AINI can be considered as a collection of specific conversation domain units. Each unit handles a specific body of knowledge used during the conversation between AINI and the user. The knowledge can be seen as arranged in the vertical columns making up the open domain or domain-specific knowledge. In addition, specific subjects are shown in the horizontal rows. For example, in the open domain knowledge, the subject units will cover topics such as personality, business, biology, computer, etc. In this report, our focus is on the medical subject and in particular, the bird flu pandemic, therefore there are additional modules being incorporated.

The domain knowledge model plays a major role in conversational systems. Such systems normally are comprised with two subcategories: the *traditional* or *narrow* domain, and, the *open domain*. In the traditional domain, systems attempt conversational fluency based on limited domains of expertise. ELIZA[34)] for example simulates a Rogerian psychotherapist, and this implementation is commonly known as DOCTOR. The Rogerian psychotherapist knowledge base attempts to engage the patient by repeating the patient's statements, and by encouraging the patient to continue talking. Terry Winograd's SHRDLU[37)] is another program simulating a robot which is able to interact within a simple world which consists of colored building blocks and boxes on a flat surface. SHRDLU has knowledge about its world and it can answer questions in natural language.

Three decades have passed since ELIZA was created. Computers have

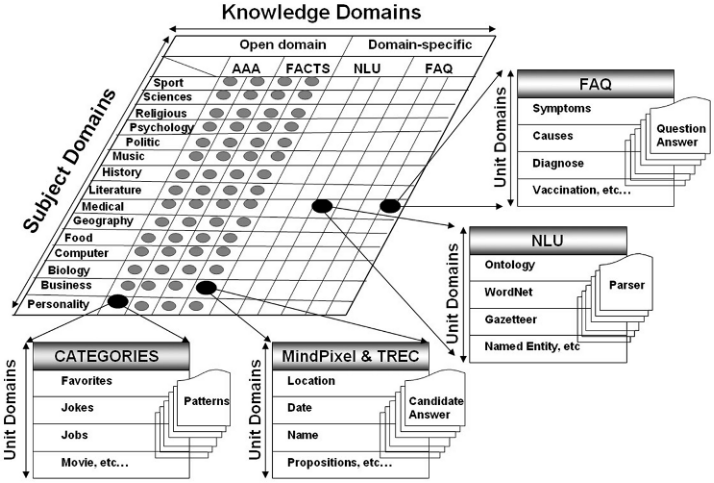


Fig. 2 AINI Domain Knowledge Matrix Model

become significantly more powerful, while storage space and memory size have increased exponentially. These advances have given researchers the opportunity to provide embedded human-like knowledge bases for conversational systems. It is assumed that the human brain contains a world of knowledge, but it has limitations on memory retrieval. Hence, knowledge needs to be specified. In the proposed system, the domain knowledge matrix model is designed along this line. The system uses custom domain-oriented knowledge bases and existing knowledge bases from online documents and training corpora. It contains a spectrum of possible solutions from queries on specific domains to general conversation questions. It can also be considered as a XML-like metadata model. This approach does not attempt to predict possible inputs from the user. Instead, the system aims to handle conversations within a specified domain or focus on 'domain-specific' conversations.

In this project, the novel contribution is the development of the "*domain knowledge plug-in components*." With this arrangement, the domain-specific knowledge could become portable, scalable and incorporated easily with other applications. This approach will allow future improvements to encourage collaborative contribution to the specific knowledge domain. The proposed system will enable the development of a wide range of information sources.

3.1 Open-Domain Knowledge Bases

Open-domain conversational systems need to deal with questions about nearly any topic. It is very difficult to rely on ontological information due to the absence of wide and yet detailed banks of world knowledge. On the other hand, these systems have much more information and data to be used in the process of answering the queries then any domain specific systems. In AINI's conversation system, information from the large-scale mass collaboration Mindpixel[30)] and training data sets from the Text Retrieval Conference's (TREC) training corpus[32)] are used. AINI also uses ALICE Annotated AIML (AAA),[5)] the Loebner Prize winner[28)] and the Chatterbox Challenge Winner[10)] hand crafted knowledge bases. These are illustrated in the Data Layer in Figure 1 and under the Open-domain columns in Fig. 2.

Mindpixel is a common sense knowledge component and it is similar to OpenMind[*3] and Cyc.[*4] The system accepts public contributions. However, Cyc model and OpenMind had bottlenecks which prevent truly large-scale collaboration[33)]. The first is the fact that knowledge does not grow by itself. Every new rule or axiom has to be entered manually and the process takes a lot of patience and time. Furthermore, information has to be input with the CycL programming language and to follow the rules of the system. The second drawback is the complexity of Cyc. It will need months to install and implement a system that is based on the knowledge base of Cyc. On the other hand, Mindpixel started collecting their propositions privately via email in 1994 and then evolved to online mass collaboration. To date, the project's user base of nearly fifty thousand people has contributed more than one million propositions and recorded almost ten million individual propositional response measurements. AINIs use only 10% of the Mindpixel propositions. In practice, 10% of the training corpus is held back from training to act as a generalization test to ensure the system does not simply memorize the corpus. Passing this generalization test would be the basis for claiming that the system is able to replicate human-level intelligence in a machine. Although a lot of knowledge has been collected, it is recognized that the system is still less than the uncountable "pieces" of common sense knowledge that are estimated to be involved with human intelligence.[31)]

A second common sense knowledge component deployed by AINI is a training corpus from TREC as shown in Table 1. TREC, organized each year by the National Institute of Standards and Technology (NIST), has offered a specific track to evaluate large-scale open domain question-answering (QA) systems since 1999. Finding textual answers to open domain questions in large text collections is a difficult problem. In our system, we only extracted factoid questions to be incorporated in the AINI's engine. Our concern is the types of questions that could be answered in more than one way. We have tried to avoid such questions. In conversational systems, factoid questions should have only single factual answers.[3, 12, 13, 36)] These are considered as a good stimulus-response type of knowledge unit. Examples of such questions are, "*Who is the author of the book, The Iron Lady: A Biography's of Margaret Thatcher?*" "*What was the name of the first Russian astronaut to do a Spacewalk?*" or "*When was the telegraph invented?*" TREC's corpus has a considerably lower rate of answer redundancy than the web and thus, it is easier to answer a question by simply extracting the answers from the matching text. To gather this data, we automatically classified questions in the TREC 8 through TREC 10 test sets by their 'wh'-word and then manually distinguished factoid questions, which represented

[*3] www.openmind.org

[*4] www.cyc.com

Table 1 Number of Factoid Questions from TREC 8, 9 and 11

TREC	Factoid Question	Text Research Collection
8	196	• Financial Times Limited (1991, 1992, 1993, 1994) • the Congressional Record of the 103$^{\mathrm{rd}}$ Congress (1993), and the Federal Register (1994) • Foreign Broadcast Information Service (1996) and the Los Angeles Times (1989, 1990).
9	692	Set of newspaper/newswire documents which includes: • AP newswire • Wall Street Journal • San Jose Mercury News • Financial Times • Los Angeles Times • Foreign Broadcast Information Service
11	109	• MSNSearch logs donated by Microsoft • AskJeeves logs donated by Ask Jeeves.

around half of the initial corpus as shown in Table 1.

The third knowledge base in the AINI's open domain knowledge model is obtained from hand-crafted Annotated ALICE AIML (AAA),[*5] a Loebner Prize winning[28)] conversation system knowledge base. AAA is a free and open-source software package based on XML specifications. It is a set of Artificial Intelligence Markup Language (AIML) scripts and this is the backbone of the award winning conversation system. AAA is specifically reorganized to facilitate conversational system developers to clone the 'brain' of the conversation system and to enable the creation of customized conversation agent personalities. The approach has reduced the need to invest huge efforts in editing the original AAA content. AAA's knowledge bases covered a wide range of subject domains based on the conversation agent's "personality." Example subjects include AI, games, emotion, economics, film, books, sport, science, epistemology and metaphysics. These subjects are shown in Fig.2 as part of the Domain Knowledge Matrix Model.

In order to illustrate the ability of ALICE in handling common sense or open domain queries, ALICE has won the 2000, 2001 and 2004 Loebner Prize for being the most lifelike machine. The competition is based on the Turing Test[1)] which aims to determine whether the responses from a computer can convince a human into thinking that the computer is a real person. In the competition, ALICE used a library of over 30,000 stimulus-response pairs written in AIML to answer the queries. The development of ALICE is based on the fact that the distribution of the sentences in conversations tends to follow Zipf's Law.[25)] It is indicated that the number of "first words" is only limited to about two thousand. The frequency of any word is roughly inversely proportional to its rank in the frequency table. The most frequently used word will occur approximately twice as often as the second most frequent word. It in turn occurs twice as often as the third most frequent word, and so forth. Questions starting with "WHAT IS" tend to have Zipf-like distributions. This type of analysis can now be accomplished in a few milliseconds of computer time. While the possibilities

[*5] www.alicebot.org/aiml/aaa/

of what can be said are infinite, the range of what is *actually* said in conversation in most cases is surprisingly small. Specifically, 1800 words cover 95% of all the first words input. It is this principle that AINI is operating on which enables it to be able to respond in an efficient and mostly accurate manner.

3.2 Domain-Specific Knowledge Bases

At present, the World-Wide Web provides a distributed hypermedia interface to a vast amount of information available online. For instance, Google[14)] currently has a training corpus of more than one trillion words (1,024,908,267,229) from public web pages. This is valuable for many types of research. The Web is a potentially unlimited source of knowledge repository; however, commercial search engines may not be the best way to gather answers from queries due to the overwhelming number of results from a search.

Before the rise of domain-oriented conversational systems using on natural language understanding and reasoning, evaluation was never a problem, as information retrieval-based metrics were readily available for use. However, when conversation systems begin to become more domain specific, evaluation becomes a real issue.[15)] This is especially true when Natural Language Processing (NLP) is required to cater for a wider variety of questions and, at the same time, required to achieve high quality responses.

As shown in Fig. 1 and 2, AINI's domain-specific knowledge bases consist of Natural Language Corpus and Frequently Asked Questions (FAQ). Both components are extracted from the online documents using the AKEA as described in reference.[18)] Another significant aspect of this paper is the objective of AINI to deliver essential information from trusted sources while capable of interacting with the users. A discussion on the selection of the trusted websites is given below.

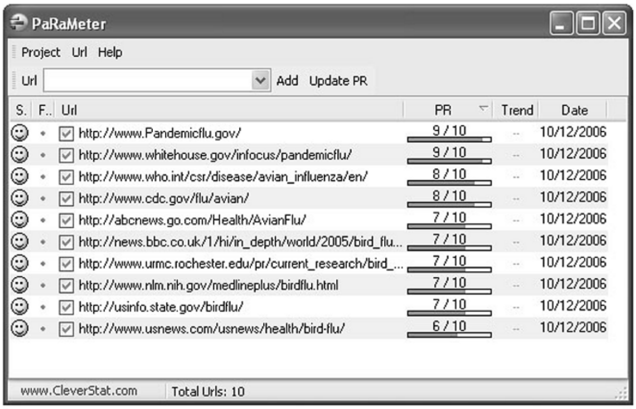
[1] Selecting trusted websites

As the Web that we know today becomes increasingly chaotic, overpowering and untrustworthy, selection of trusted Web pages is becoming an important factor contributing to its long-term survival as a useful global information repository. In our experiment, the selection of the trusted websites is based on PageRank$^{\mathrm{TM}}$.[8)] PageRank$^{\mathrm{TM}}$is a system for ranking web pages developed by Larry Page and Sergey Brin at Stanford University. PageRank$^{\mathrm{TM}}$relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value and high-quality sites always receive a higher PageRank$^{\mathrm{TM}}$.

In our study, the selection of trustworthy websites started with the initial six seed words: *bird, flu, avian, influenza, pandemic* and $H5N1$. These seeds are supposed to be representative of the domain under investigation. The seed terms are randomly combined and each combination is used in Google API[*6] and BootCat Tool[6)] for bootstrapping corpora and terms from the web. We used the seeds to perform a set of Google searches, asking Google to return

[*6] http://www.google.com/apis

a maximum of 20 URLs per query and then we collected the corpus. After visual inspection of the corpus, we used the top 40 seeds extracted from token frequencies for the second run. Finally, we retrieved 1,428 URLs out of 1,500 URLs related to the domain being investigated. The reduction in number is due to the duplicated and broken link URLs being removed. Based on the 1,428 URLs, we sent a query to Google's PageRank$^{\mathrm{TM}}$directory using PaRaMeter Tool[11)] to determine their rankings. Figure 3 shows the results of the top 10 site based on the PageRank$^{\mathrm{TM}}$scale. The PageRank$^{\mathrm{TM}}$scale goes from 1 to 10. A less important site is the one with a PageRank (PR) of 1. The most referenced and supposedly important sites are those with a PR of between 7 and 10.



S.	F..	Url	PR	Trend	Date
☺	•	☑ http://www.Pandemicflu.gov/	9 / 10	--	10/12/2006
☺	•	☑ http://www.whitehouse.gov/infocus/pandemicflu/	9 / 10	--	10/12/2006
☺	•	☑ http://www.who.int/csr/disease/avian_influenza/en/	8 / 10	--	10/12/2006
☺	•	☑ http://www.cdc.gov/flu/avian/	8 / 10	--	10/12/2006
☺	•	☑ http://abcnews.go.com/Health/AvianFlu/	7 / 10	--	10/12/2006
☺	•	☑ http://news.bbc.co.uk/1/hi/in_depth/world/2005/bird_flu...	7 / 10	--	10/12/2006
☺	•	☑ http://www.urmc.rochester.edu/pr/current_research/bird_...	7 / 10	--	10/12/2006
☺	•	☑ http://www.nlm.nih.gov/medlineplus/birdflu.html	7 / 10	--	10/12/2006
☺	•	☑ http://usinfo.state.gov/birdflu/	7 / 10	--	10/12/2006
☺	•	☑ http://www.usnews.com/usnews/health/bird-flu/	6 / 10	--	10/12/2006

www.CleverStat.com Total Urls: 10

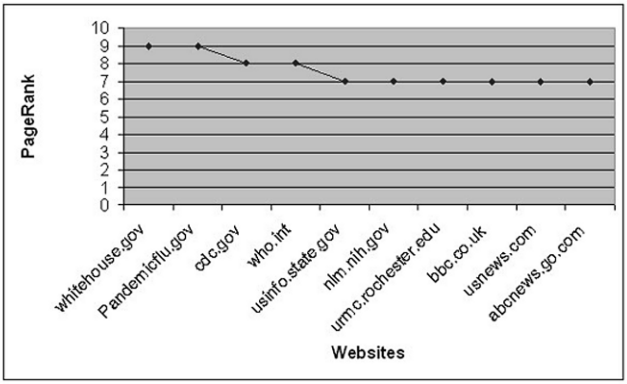


Fig. 3 Top 10 Pagerank$^{\mathrm{TM}}$Scale for Bird Flu Domain

The final set of URLs was further culled to include only selected sites which are attributed to a regulated authority (such as a governmental or educational institution) that controls the content of the sites. Once the seed set is determined, each URL's page is further examined and rated as either reliable or reputable. This selection is reviewed, rated and tested for connectivity with the trusted seed pages. From this exercise, *whitehouse.gov, pandemicflu.org, cdc.gov* and *who.int* were selected due to their PageRank$^{\mathrm{TM}}$scale scores being above 7. The most important factors in determining the "reliable authority" of a site is based on its history and the number of back-links to the governmental and international organization links. The more established and relevantly linked the more it will be considered as "stronger" or "more reliable." This effectively gives the linked site a measure of "trust" and "credential." The selected URLs are then used as the source knowledge base for AKEA to extract the contents on bird flu so as to build AINI's domain-specific knowledge base. A discussion of the extracted information is given in the following section.

[2] Extracted online documents

In AINI's Domain Knowledge Matrix Model, the unit domains in the Natural Language Corpus component consist of knowledge and information harvested from or expressed in ontologies, gazetteers, named entities and WordNet. These have been implemented as domain-dependent modular components. The named entity module identifies named locations, named persons, named organization, dates, times and key measures in text formats. The information is obtained by AKEA. For example, information for diseases is based on symptoms, causes, diagnoses, vaccination locations, persons and organizations. In order to identify these entities, our system uses rules to specify the named entities' structure in terms of text tokens and information from the source such as tagger, morphosyntactic analyzer and knowledge bases of names, clue words and abbreviations.

The web knowledge base is then continuously updated with facts extracted from online *pandemic news* using information extraction (IE) by AKEA. IE is the task of extracting relevant fragments of text from larger documents and to allow the fragments to be processed further in an automated manner. An example of an application of AKEA is to prepare an answer for a user's query. The ontology and gazetteer have been implemented as domain-dependent modular components which will allow future improvements in the domain knowledge.

In AINI's FAQ component, the unit domain consists of information concerning diseases, symptoms, causes, diagnoses, vaccinations, etc. The selection of FAQ trusted Web pages has been carried out using PageRank$^{\text{TM}}$as discussed above. But at this stage, each of the selected websites was evaluated in order to find the more suitable and reliable FAQ pages. From this experiment, the *answers.pandemicflu.gov* and *who.int/csr/disease/avian_influenza/avian_faq* pages have been selected as the source of information for AKEA to build AINI's FAQ knowledge base.

Based on the proposed approach, the quality of the results returned from AINI's engine using the FAQ knowledge base are either similar to or better than those generated by search engines such as Google. AINI's SQL engine uses the most significant words as keywords or phrase. It attempts to find the longest pattern to match without using any linguistic tools or NLP analysis. In this component, AINI does not need a linguistic knowledge unit and relies on just an SQL query. All questions and answers can be extracted from the database which was built by AKEA after applying a filtering process to remove unnecessary tags. Example results from the system are illustrated in the next section.

As shown in Table 2, AINI's open domain knowledge base currently has more than 150,000 entries in the common sense stimulus-response categories. Out of these, 100,000 came from Mindpixel, 997 factoid questions from the TREC training corpus and 45,318 categories from AAA knowledge bases. On the domain-specific knowledge base, AINI has more then 1,000 online documents extracted by AKEA. This makes up 10,000 stimulus-response items in total. AINI also has 158 FAQ pairs of questions and answers which have been updated using AKEA. In addition, AINI has also collected more than 382,623 utterances in conversations with online users since 2005. These utterances will be integrated into AINI's knowledge bases through supervised learning by domain experts. At present, AINI has learnt 50,000 categories from conversations with online users. All of this combined knowledge has made up the total of 206,473 stimulus response categories in AINI's knowledge bases. In comparison to other system, the original conversational programs such as ELIZA,[35] written by Professor Joseph Weizenbaum of MIT, have only 200 stimulus response categories. ALICE Silver Edition was ranked the "most human" computer, and has about 120,000 categories which include 80,000 taken from Mindpixel.

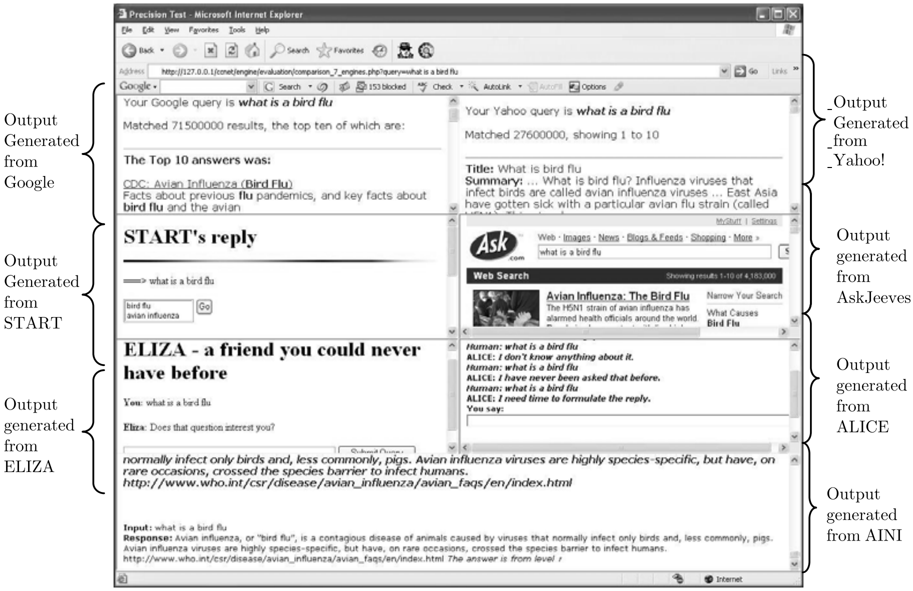
Table 2 AINI's Knowledge Bases

Domain Knowledge	Sources	Categories	%
Domain-Specific	NL Corpus	10,000	4.8%
	FAQ	158	0.1%
Open-Domain	Mindpixel	100,000	48.4%
	TREC Corpus	997	0.5%
	AAA	45,318	21.9%
Supervised Learning	Conversation Logs	50,000	24.2%
	TOTAL	**206,473**	

§4 Experimental Setup

In this experiment, three types of systems are compared. They are *search engine*, *question answering system* and *conversational system*. For each system, we compared two different engines against AINI. The two search engines compared are Google and Yahoo. For the question answering engines, AskJeeves and START are used. They are supposed to use natural language processing for their queries. For the conversational engines, ELIZA and ALICE are selected. In particular, ALICE was ranked as the "most human computer" in the Turing Test competition.[28]

Google is a well known search engine which determines relevancy of information primarily on their PageRank algorithm.[24] In our experiment, we developed a search engine interface using Google SOAP Search API service[22] and Yahoo!.[38] For the Question Answering system, the idea behind Ask Jeeves and START is to allow users to get answers for questions posed in natural language. ASK Jeeves is the first commercially question answering system available on the Internet. START[27] is the world's first Web-based question answering system which commenced operation since December, 1993. ELIZA is a well known program in the discipline of Artificial Intelligence and it is also the oldest system


Precision Test - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://127.0.0.1/conet/engine/evaluation/comparison_7_engines.php?query=what is a bird flu
Go Links
Google Search 153 blocked Check AutoLink AutoFill Options
Internet

Output Generated from Google

Your Google query is ***what is a bird flu***

Matched 71500000 results, the top ten of which are:

The Top 10 answers was:

CDC: Avian Influenza (**Bird Flu**)
Facts about previous **flu** pandemics, and key facts about **bird flu** and the avian

Output Generated from Yahoo!

Your Yahoo query is ***what is a bird flu***

Matched 27600000, showing 1 to 10

Title: What is bird flu
Summary: ... What is bird flu? Influenza viruses that infect birds are called avian influenza viruses ... East Asia have gotten sick with a particular avian flu strain (called

Output Generated from START

START's reply

===> what is a bird flu

bird flu
avian influenza

Go

Output generated from AskJeeves

MyStuff | Settings

Web · Images · News · Blogs & Feeds · Shopping · More »

what is a bird flu

Web Search Showing results 1-10 of 4,183,000

Avian Influenza: The Bird Flu
The H5N1 strain of avian influenza has alarmed health officials around the world.

Narrow Your Search
What Causes **Bird Flu**

Output generated from ELIZA

ELIZA - a friend you could never have before

You: what is a bird flu

Eliza: Does that question interest you?

Submit Query

Output generated from ALICE

Human: ***what is a bird flu***
ALICE: ***I don't know anything about it.***
Human: ***what is a bird flu***
ALICE: ***I have never been asked that before.***
Human: ***what is a bird flu***
ALICE: ***I need time to formulate the reply.***
You say:

Output generated from AINI

normally infect only birds and, less commonly, pigs. Avian influenza viruses are highly species-specific, but have, on rare occasions, crossed the species barrier to infect humans.
http://www.who.int/csr/disease/avian_influenza/avian_faqs/en/index.html

Input: what is a bird flu
Response: Avian influenza, or "bird flu", is a contagious disease of animals caused by viruses that normally infect only birds and, less commonly, pigs. Avian influenza viruses are highly species-specific, but have, on rare occasions, crossed the species barrier to infect humans.
http://www.who.int/csr/disease/avian_influenza/avian_faqs/en/index.html *The answer is from level 1*

Fig. 4 Experimental Design Interface

of similar type. ALICE[4] is a general conversation system based on the Annotated ALICE Artificial Intelligence Markup Language. The knowledge base rule set consists of approximately 46,424 categories. In a way, AINI can also be considered as an enhancement of ALICE with the inclusion of the pandemic domain-specific knowledge base which was extracted by AKEA, and it also has parsing capabilities based on a full Natural Language Understanding engine for multilevel natural language query.[17] In the present study, appropriate or accurate domain responses are expected to be important. The control question set was submitted to the seven URLs where the seven systems were located. The responses of the queries are then collected and displayed as shown in Fig. 4.

§5 Experiment Results

In the course of the present study, a question was raised on how to compare the performance of the proposed system. It was decided that the performance of the system would be based on the keywords instead of complex phrases. For example, if the question "*What is bird flu*?" is asked, responses from the search engines, question-answering engines and conversational engines are summarized in Table 3, 4 and 5 respectively. The responses are noted as follows:

- **Search Engines** – The systems identify the key words and return documents that are relevant to the query. The results could be tens of thousand or even millions of hits as shown in Table 3.

- **Natural Language Question-answering Systems** – They generate responses using NLP and return fewer possible answers. The answer is

Table 3 Responses from Popular Search Engines – Google and Yahoo

SEARCH ENGINES	
ENGINE	**RESPONSES**
Google	Reply 7.15 million pages. First hit was " Facts about previous flu pandemics, and key facts about bird flu and the avian influenza A (H5N1) virus.."
Yahoo	Reply 27.6 million pages. First hit was " Influenza viruses that infect birds are called avian influenza viruses ... East Asia have gotten sick with a particular avian flu strain (called H5N1). This virus has ..."

Table 4 Responses from Popular Question Answering Engines

QUESTION ANSWERING ENGINES	
ENGINE	**RESPONSES**
AskJeeves	Reply 4.183 million pages. First hit was " The H5N1 strain of avian influenza has alarmed health officials around the world. People in close contact with live birds have contracted the disease, raising concerns that it may change into a highly contagious form that can be transmitted from human to human. ."
START	Main Entry: **bird flu** Function: *noun*: **AVIAN INFLUENZA** **Source**: Merriam-Webster Dictionary

Table 5 Respohnses from Conversational Engines

CONVERSATIONAL ENGINES	
ENGINE	**RESPONSES**
ELIZA	Does that question interest you?
ALICE	I haven't heard of bird flu.
AINI	Avian influenza, or "bird flu", is a contagious disease of animals caused by viruses that normally infect only birds and, less commonly, pigs. Avian influenza viruses are highly species-specific, but have, on rare occasions, crossed the species barrier to infect humans. http://www.who.int/csr/disease/avian_influenza/avian_faqs/en/index.html

possibly within the context of the document. The sample response is shown in Table 4.

- **Conversation Engines -** provide quick response from the stimulus. Simple pattern matching and substitution are used based on a surprisingly small number of pre-defined rules. The example is shown in Table 5.

§6 Discussion

In this example, ELIZA responded with "*Does that question interest you?*" It is observed that ELIZA tries to ask another question, instead of giving an answer. The objective is to encourage the user to continue with the conversation. On the other hand, ALICE attempts to convince the user by generating randomly answer from AIML knowledge base. ALICE's response does not need a grammatical parser as her knowledge base contains the pattern "WHAT IS

BIRD FLU?" and the witty reply is "XFIND *" with an AIML categories. By using "XFIND *" pattern, ALICE will randomly generated responses such as *"Is there only one," "let me think about it," "Have you tried a web search," "I haven't heard of bird flu,"* etc. The pattern matching language used in ALICE permits only one wild-card ('*') match character per pattern. Therefore, ALICE responds with a variety of inputs from the users. ALICE does not concern whether it really "understands" the input. It aims to provide a coherent response to the client in order to convey the impression that the system understands the client's intention. For the Eliza and ALICE systems, they are not able to handle questions that demand specific answers. They are simply not designed for such purpose. The three possible ways to handle these types of questions are:

(a) Analyze the problems with NLP and then provide an appropriate answer,
(b) Rely on human to review the conversation logs and continually improves the knowledge base, or,
(c) Treat the query as impossible and then choose a pre-defined random answer.

For the AINI chatterbot, the response was *"Avian influenza, or "bird flu," is a contagious disease of animals caused by viruses... ".* The answer was generated from the domain-specific knowledge base using Natural Language Understanding parsing from Level 1. In this query, the answers were discovered by AINI from the trusted sites such as WHO. In addition, the response is based on the natural language understanding and reasoning. The reasoning mechanism of the AINI is based on answer discovery in layer-oriented knowledge base. Currently ongoing work includes quantitative measurement and assessment of the results and performance of these different systems. Although the systems used in this study were built with different objectives in mind, the purpose of this study is to show that there is a need of the proposed system to handle domain specific applications. At the same time, this study also shown that the proposed conversation agent architecture can achieve the expected objectives.

§7 Conclusion

In this paper, we have reported the use of an AINI conversation agent architecture to develop a domain specific intelligent web interaction. AINI consists of a knowledge acquisition tool called AKEA for the gathering of conversation and domain-related knowledge. The proposed system has the potential to be used in a domain specific application area. The study has demonstrated with one particular area of domain expertise - Bird flu. In this paper, we only worked on selected pandemic crisis websites where we performed knowledge extraction through AKEA for the domain-specific knowledge databases on the server. It is believed that the approach would be useful and applicable for other domains. We have also compared and tested the flexibility of AINI against other popular search engines, question answering systems and conversation systems. Furthermore, we have found that domain-specific knowledge base has higher response than the corresponding conversational-style responses. Further work will be done on expanding the sources of knowledge and to provide quantitative measurements of the quality of responses from the AINI.

Acknowledgements

We would like to thank the anonymous referees for the constructive and helpful suggestions. This research project is funded by Murdoch University Research Excellence Grant Scheme (REGS), 2006.

References

1) Turing, A.M., "Computing Machinery and Intelligence," *MIND the Journal of the Mind Association, LIX*, 236, pp. 433-460, 1950.
2) Adobe, "Mobile and Devices," www.macromedia.com/mobile/, 2006.
3) Agichtein, E., Cucerzan, S. and Brill, E., "Analysis of factoid questions for effective relation extraction," in *Proc. 28th Annual Int'l ACM SIGIR Conf. on Research and development in information retrieval*, Salvador, Brazil, 2005.
4) Alice, "Artificial Linguistic Internet Computer Entity," http://www.alicebot.org, 2005.
5) Alicebot, "The Annotated A.L.I.C.E. AIML," Alicebot.org. http://www.alicebot.org/aiml/aaa/, 2006.
6) Baroni, M. and Bernardini, S., "BootCaT: Bootstrapping corpora and terms from the web," in *Proc. Fourth Language Resources and Evaluation Conf., 2004*.
7) Bialik, C., "Just How Deadly Is Bird Flu? It Depends on Whom You Ask," *The Wall Street Journal Online*, http://online.wsj.com/public/article/SB110512998255120225.html?mod=todays_free_feature, January 13, 2005.
8) Bianchini, M., Gori, M. and Scarselli, F., "Inside PageRank. ACM Transactions on Internet Technology," *ACM Transactions on Internet Technology, 5-1*, 2005.
9) Brewster, S., "Overcoming the Lack of Screen Spaces on Mobile Computers," *Presonal and Ubiguitous Computing, 6, 3*, pp. 188-205, Springer, London, 2002.
10) Chatterboxchallenge, "ALICE Winner of Chatterbox Challenge 2004," http://www.chatterboxchallenge.com/, 2006.
11) CleverStat, "PaRaMeter," http://www.cleverstat.com, 2006.
12) Collins-Thompson, K., Terra, E., Callan, J., and Clarke, C. "The effect of document retrieval quality on factoid question-answering performance," in *Proc. IGIR 2004*, Sheffield, UK, 2004.
13) Cucerzan, S. and Agichtein, E., "Factoid Question Answering over Unstructured and Structured Web Content," Microsoft http://research.microsoft.com/users/silviu/Papers/trec05.pdf, 2005.
14) Franz, A. and Brants, T., "All our N-gram are Belong to You," *Google Machine Translation Team* http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html, 2006.
15) Goh, O.S., Ardil, C., Wong, W. and Fung, C.C., "A Black-box Approach for Response Quality Evaluation Conversational Agent System," *Int'l Journal of Computational Intelligence, 3-3*, pp. 195-203, 2006.

16) Goh, O.S., Depickere, A., Fung, C.C. and Wong, K.W. "Domain Matrix Knowledge Model for Embodied Conversation Agents," in *Proc. 5th Int'l Conf. on Research, Innovation & Vision for the Future (RIVF'07), Hanoi, Vietnam*, IEEE Press, 2007.

17) Goh, O.S., Depickere, A., Fung, C.C., and Wong, K.W., "Top-down Natural Language Query Approach for Embodied Conversational Agent," in *Proc. Int'l MultiConf. of Engineers and Computer Scientists 2006, Hong Kong*, International Association of Engineers, 2006.

18) Goh, O.S. and Fung, C.C., "Automated Knowledge Extraction from Internet for a Crisis Communication Portal," in *First Int'l Conf. on Natural Computation, Changsha, China, LNCS*, pp. 1226-1235, 2005.

19) Goh, O.S., Fung, C.C., Depickere, A., Wong, K.W. and Wilson, W., "Domain Knowledge Model for Embodied Conversation Agent," in *Proc. 3rd Int'l Conf. on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2005), Singapore*, 2005.

20) Goh, O.S., Fung, C.C. and Lee, M.P., "Intelligent Agents for an Internet-based Global Crisis Communication System," *Journal of Technology Management and Entrepreneurship, 2-1*, pp. 65-78, 2005.

21) Goh, O.S., Fung, C.C., Wong, K.W. and Depickere, A., "Embodied Conversational Agents for H5N1 Pandemic Crisis," *Journal of Advanced Computational Intelligence and Intelligent Informatics, 11-2*, 2007.

22) Google, "Google SOAP Search API (beta)," Google,
http://www.google.com/apis/, 2006.

23) GSMWorld, "3GSM Statistics,"
http://www.gsmworld.com/technology/3g/statistics.shtml, 2006.

24) Journal Search Engine, "Google and Yahoo Search Engine Technology Comparison," www.searchenginejournal.com
http://www.searchenginejournal.com/?p=2267, 2006.

25) Li, W., "Zipf's Law," Feinstein Institute for Medical Research,
http://www.nslij-genetics.org/wli/zipf/, 2006.

26) Lin, D., "Dependency-based Evaluation of MINIPAR," in *Proc. Workshop on the Evaluation of Parsing Systems. Granada, Spain*, 1998.

27) Lin, J. and Katz, B., "Building a Reusable Test Collection for Question Answering," *Journal of the American Society of Information Science and Technology, 57-7*, pp. 851-861, 2006.

28) Loebner, H., "Loebner Prize,"
http://www.loebner.net/Prizef/loebner-prize.html, 2006.

29) Lovgren, S., "Is Asian Bird Flu the Next Pandemic?" *National Geographic News*,
http://news.nationalgeographic.com/news/2004/12/1207_041207_birdflu.html, December 7, 2004.

30) MindPixel, "GAC-80K Mindpixel,"
http://www.mindpixel.com/chris/gac80k-06-july-2005.html, 2006.

31) Mueller, E., "Common Sense in Humans,"
http://www.signiform.com/erik/pubs/cshumans.htm, 2001.

32) NIST, "Text REtrieval Conference (TREC),"
http://trec.nist.gov/, 2006.

33) Richardson, M. and Domingos, P., "Building large Knowledge Bases by Mass Collaboration," in *Proc. K-CAP'03. Sanibel Island, Florida, USA*, ACM Press, 2003.

34) Weizenbaum, J., *Computer Power and Human Reason*, W.H. Freeman and Company, 1976.

35) Weizenbaum, J., "ELIZA - A computer program for the study of natural language communication between man and machine," *Communications of the ACM, 9-1*, pp. 36-45, 1966.

36) Whittaker, E.W.D., Hamonic, J., Yang, D., Klingberg, T. and Furui, S., "Monolingual Web-based Factoid Question Answering in Chinese, Swedish, English and Japanese," in *Proc. EACL 2006 Workshop on Multilingual Question Answering - MLQA06*, 2006.

37) Winograd, T., *Understanding Natural Language*, Academic Press, 1972.

38) Yahoo!, "Yahoo! Search Web Services," Yahoo! http://developer.yahoo.com/search/, 2006.

Ong Sing GOH: He was Associate Professor of Faculty of Information Technology and Communication, University Technical Malaysia Melaka. His research interest is in the development of intelligent agent and conversational agents to facilitate graceful human-computer interactions. He is a Member of IEEE, Senior Member, Malaysia Senior Scientists' Association (MSSA)

Chun Che FUNG: He received the B.Sc.(1st Class Hon.) and M.Eng degrees from the University of Wales, Cardiff, and the Ph.D degree from the University of Western Australia. He is currently an Associate Professor at Murdoch University, Australia. His research interests include computational intelligence techniques and their applications.

Kok Wai WONG: He received B.Eng (Hons.) and Ph.D. from Curtin University of Technology in 1994 and 2000 respectively. He is currently working as an Associate Professor at Murdoch University in Western Australia. His current research interests include intelligent data analysis, digital media technology, and data mining.