Goh, O.S., Fung, C.C. and Depickere, A. (2008) ***Domain knowledge query conversation bots in instant messaging (IM).*** **Knowledge-Based Systems, 21 (7). pp. 681-691.**

Domain knowledge query conversation bots in instant messaging (IM)

Ong Sing Goh, Chun Che Fung and Arnold Depickere

School of Information Technology, Murdoch University, Murdoch, WA 6150, Australia

Abstract

In this paper, we examine the use of knowledge query technology as applied to conversation bots in the instant messaging environment. Hence, we designed an artificial intelligent conversation robot or bots called Artificial Intelligence Natural language Identity (hereafter, AINI) to mimic human conversation. Our goal is to introduce a Domain Matrix Knowledge Model and an Automated Knowledge Extraction Agent (AKEA) to create AINI's knowledge bases, and in turn provide intelligent query mechanisms. We report an evaluation on the collection and analysis of a corpus containing over 3280 utterances in a series of real instant messages exchanged between the AINI conversation bot and 65 online "buddies". About 1721 utterances were produced by AINI, 88.03% were from open-domain knowledge, 2.15% from domain-specific knowledge base and 9.82% were inappropriate and amusing responses. These results show that domain knowledge plays significant roles in conversations between two or more human users and in human–machine conversation.

Keywords: Conversation bots; Artificial Intelligent Natural language identity (AINI); Artificial intelligence (AI); Instant messaging (IM)

1. Introduction

Instant messaging has exploded in popularity in the past few years as a supplement to emails. Teens in particular have joined instant messaging networks *en masse*, prizing the ability to engage in flurries of instantaneous conversation with a selected list of online "buddies". Popular chatting or instant messaging (IM) systems, such as America Online's Instant Messenger, Microsoft's MSN Messenger, ICQ, and Internet Relay Chat (IRC) have changed the way we communicate with friends, acquaintances, and colleagues. Once limited to desktops, popular instant messaging systems are finding their way onto handheld devices and cell phones, allowing users to chat from virtually anywhere. Nowadays, IM is found on just about every private PC connected to the Internet as well as on many corporate desktops. This technology makes communication even easier than emails or phone calls. Just type a simple message, hit enter, and you and your buddy are up and chatting. This technology is growing rapidly as a dominant form of communication. Research by Pew Internet & American Life [48]reveals that 53 million adults trade instant messages and of those, 24% swap IMs more frequently than email.

This popularity created an enthusiasm among the IM proprietary, including Microsoft, to integrate conversation robots (or 'bots') within their MSN Messenger system. Microsoft challenged developers worldwide to create conversation bots for MSN®Messenger and Windows Live™ in the "Invasion of the Robots Contest".[1] In this paper, we propose a Domain Matrix Knowledge Model for conversation bots in IM. True intelligent action requires large quantities of knowledge. The inability to acquire appropriate and sufficient knowledge has long been the major constraint preventing the rapid adoption of conversational systems. Using manual approaches to input handcrafted knowledge is time consuming, costly and impractical. In our proposed approach, we aim to harvest the vast reservoir of knowledge from the internet by deploying the Domain Matrix Knowledge model. This will assist the construction of large-scale knowledge bases to be used as the engine for the future intelligent conversation bots for IM.

More precisely, in this paper we test our hypothesis using the information domain of pandemic bird flu, taking advantage of the wealth of documents on the subject from the World-Wide Web (WWW). This leads to the assumption that integration of domain-specific and open-domain knowledge bases in the Domain Matrix Knowledge Model is more promising and could lead to better overall query results.

2. Related works

Several works have been published that refer in general to the use of IM as a new media of communication between human users. However, few are working on human–machine conversation in MSN Messenger. Recent reports have stated that Internet chat, including IM, is being monitored by US officials in exchanges where participants appear to be planning terrorist attacks [50], and also where there are concerns for the security of younger users who could become victims of criminals [17,45]. On the social impact, some papers strongly criticize this new form of communication [35] while others suggest that its not a matter for society's approval or otherwise, but of accepting that IM is here to stay, and that digital communications technologies evolve and improve constantly and quickly [11,19,7]. There are also papers referring to research on design and usability for the public in general [35,41,47] and IM usage in workplace and corporate contexts has recently soared [13,29,43,40]. With regard to the linguistic aspects of IM usage, research has been undertaken in Spain [19], United Kingdom [35], United State [16], Sweden [28]and Portugal [17]. However, research on conversation bots has only gained popularity in recent years such as in the AOL Instant Messenger [30]. The first initiative to develop conversation bots was launched by Microsoft[2] in the "Robot Invaders Contest 2006",[3] which seeks the best new conversational robot ideas for MSN Messenger. Entries will be judged on user interaction, usefulness, creativity, and use of multiple Windows Live services. This contest was supported by bot developers such as BuddyScript,[4] Incesoft Bot[5] and Akonix,[6] who unleashed their Software Development Kits (SDKs) that each work like a "bot" within the three major IM systems – America Online's AIM, Microsoft's MSN Messenger and Yahoo! Messenger. For instance, BuddyScript's services [39] are built on a natural language search called a "buddy script" that is programmed to tap into content databases. The script can be custom-built for corporate clients. However, the majority of the conversation bot's answers are based on a computer's recognition of just a few keywords, and pre-scripting which is only able to recognize phrases and sentence patterns.

In this paper, we are also examining the impacts of the IM users chatting with conversation bots added into MSN Messenger and Windows Live Messenger. We will investigate the conversation bots' domain knowledge query which can perform a wide variety of useful and fun tasks such as chatting. Emphasis will also be placed upon the ways in which a conversation bot can access customer specific information and resolve more complex issues from "her" knowledge bases. Therefore, this research continues to examine these implications by deploying newly developed conversation bots called Artificial Intelligent Natural language Identity (AINI).

3. AINI architecture

Soft computing techniques such as FSL (Fuzzy Set Logic) have been used quite effectively for realizing human-friendly systems which is a very powerful tool for encoding human knowledge formed from perception-based data and for transforming it into machine knowledge for feedback and inference [10]. The aim of the AINI framework is to allow developers to construct applications by using different anthropomorphic interface, domain knowledge and natural language query As reported in literature [25,26], the AINI's architecture can be scaled up to port to any new application in the absence of a principal approach. AINI's engines are portable, have the ability to communicate naturally and are able to carry on multiple independent conversations at the same time. AINI's knowledge bases and conversational engines use plug-in principles which can quickly be augmented with specific knowledge and are portable in nature for specific purposes.

This research project involves the establishment of an AINI conversational bot system within the MSN Messenger communication framework. The objective is to use AINI's conversation bots as the basic architecture to simulate a human partner in IM. Our real-time prototype relies on a distributed agent architecture designed specifically for Desktop, web, Mobile devices and Personal Digital Assistant (PDA) [24], all of which use human–computer communication systems. All software agents, such as the conversation engine, knowledge model and natural language query, communicate with one another via TCP/IP. This is a combination of natural language processing and multimodal communication. A human user can communicate with the developed system using typed natural language conversation.

An AINI conversation bot can be seen as a 'digital spirit', capable of occupying and controlling a physical entity such as robot [49], or an embodied container, like an avatar in our conversational agent[22]. AINI is a conversation bot designed by the authors that is capable of having a meaningful conversation with human users who interact with "her".

For the purposes of this research, the application area chosen for designing the conversation bot is primarily grounded in an ability to communicate based upon scripting and/or artificial intelligence programming in instant messaging. While this goal lies some distance down the path of future research, we present an architecture that reaches towards it; at the same time aiming for the possibility of practical applications in the nearer future. AINI adopts a hybrid architecture that combines the utility of multi-domain knowledge bases, multimodal interfaces and multilevel natural language query software. Given a question, AINI first performs 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 [34]. 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) [23], MINIPAR parser is capable to parses nearly 500 words per second on a Dell Precision PWS380 Server 3GH with 1 GB memory.

AINI employs an Internet three-tier, thin-client architecture that may be configured to work with any web application. It comprises of a data server layer, application layer and client layer. This Internet specific architecture offers a flexible solution to the unique implementation requirements of the AINI system.

4. AINI and MSN Messenger protocol

The architecture of MSN Messenger is very complicated compared to other instant messaging services such as AIM and Yahoo!, since it relies on five different types of servers to handle the communication and operation of its service.[7] MSN Messenger uses the Mobile Status Notification Protocol (MSNP) for communication. AINI uses MSN protocol to communicate with MSN Messenger servers. AINI utilizes the NET Passport to sign into the MSN Messenger service by using the ainibot@hotmail.com passport. The MSN Messenger sign-in session is based on a challenge-response mechanism to authenticate user credentials. The communication with the Passport server is conducted over the HTTPS (Hypertext Transfer Protocol over Secure Sockets Layer) protocol, ensuring that the sign-in information is encrypted. The client sends the challenge string, Passport username, and password to the Passport URL. If the credentials for signing in are confirmed, the Passport server issues a ticket, which is passed back to the notification server to complete the authentication procedure. Fig. 1 details the entire authentication procedure for AINI and MSN Messenger. Once both users connect to the same switchboard server, the messaging session commences.

5. AINI and MSN Messenger interface

We have outlined the conceptual and practical basis for the development of conversation bots for DesktopChat, WebChat and MobileChat which are supported by MSN Messenger protocol, as shown in Fig. 2. This will pave the way for a human–computer interface based on human natural language technologies. Handheld devices provide an ideal platform for art and entertainment applications considering the growing number of mobile phone users world-wide. This will improve techniques for displaying content, interaction, conversation and the emergence of wireless and shared interaction among networked users. Such applications can capture the users' attention through natural language conversational system features and the creation of personal attachment.

MSN Messenger for Desktop, or DesktopChat, was a free instant messaging client that was developed and distributed by Microsoft Windows since 1999. MSN Messenger was renamed to Windows Live Messenger in 2006. The WebChat sessions allow the users to interact in real-time with the AINI software robot at the website via a browser through MSN Web Messenger. It is possible for virtually any computer with an Internet connection, Windows XP and Internet Explorer to connect to the Messenger Service by using MSN Web Messenger. MobileChat uses a mobile chatting module, and is implemented in a series of logical phases which includes mobile-to-internet → internet-to-bots → bots-to-mobile chats. Mobile chat is an alternative way in which users can chat with AINI using GPRS, WI-FI and 3G services.

6. Domain Matrix Knowledge Model

Academic investigators and commercial developers have drawn into diverse research areas including natural language understanding (NLU), database systems, human factors and knowledge-based systems to make their systems more intuitive, user friendly and intelligent. Robert Planta and Stephen Murrell [44] investigate the development of a natural language (NL) interface for mixed initiative dialogues within a constrained domain and demonstrates the applicability of the functional approach to Natural language system. In our paper, a significant difference between our proposed conversation bots is our domain matrix knowledge model as shown in Fig. 3. The domain knowledge model was aimed to play a major role in AINI systems. This domain model can be considered as a domain-dependent modular component and allows future improvements to encourage openness and collaborative contribution to the specific knowledge domain. In our approach, we define the knowledge base of our conversation bots as a collection of specific conversation domain units. Each unit handles a specific body of knowledge used during the conversation between the AINI and the IM buddies. For example, in the open-domain knowledge, the subject domains will cover subjects such as personality, business, biology, computer, etc. In this report, our focus is on the medical subject and in particular, the bird flu pandemic.

The domain knowledge model plays a major role in conversational bot systems. Systems that rely on specially crafted knowledge bases are normally composed of 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 [52] for example simulated 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 [55] is another program simulating a robot which is able to interact within a simple world which consists of coloured building blocks and boxes on a flat surface. SHRDLU has knowledge about his world and it can answer questions about it in natural language. However, it is anticipated that SHRDLU will not be competitive enough to win the Turing test because of its limited domain of discourse.

Three decades have passed since ELIZA was created. Computers have 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 conversation bots. We follow the assumption that the human brain contains a world of knowledge, but it has limitations on memory retrieval. Hence, knowledge needs to be specified. In our system, the Domain Matrix Knowledge Model is designed along this line. Our 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. In our study, the domain model is the taxonomy of knowledge related to the specific topic. It can also be considered as XML-like metadata model. This approach will reduce the need to predict every possible input from the user. Instead, this approach allows the manager of the system to devote more effort to working out how to handle conversations within a specified domain; or 'domain-specific' conversations.

In this project, the novel contribution is the development of the "*domain knowledge plug-in components*". By this arrangement, the domain-specific knowledge becomes portable, scalable and easily incorporated in other applications. In short, this domain model can be considered as a domain-dependent modular component. This approach will allow future improvements to encourage openness and collaborative contribution to the specific knowledge domain. An Open-Domain will enable the development of a wide range of information sources. For a system that focuses on certain domains, it is more likely that the techniques are more restrictive and logic-based. There will be relatively limited available sources as compared to an Open-Domain system. In the Open-Domain system, candidate answers are ranked according to individual features such as how well the answer matches the question. A domain-oriented conversational system that deals with questions within a domain-specific environment will be seen as a richer approach. This is because natural language processing systems can exploit domain knowledge and ontologies. However, advanced reasoning, such as providing explanations for answers, generalizing questions, etc., is not possible in Open-Domain systems. Establishing this metamorphosis in conversation bots is something that's not just good for knowledge customization, but also for 'buddies' who talk with the bots.

6.1. 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. In AINI's conversation system, we deployed the large-scale mass collaboration Mindpixel [37] and training data sets from the Text Retrieval Conference's (TREC) training corpus [42]. It also uses ALICE Annotated AIML (AAA) [3], the Loebner Prize winner [36] and the Chatterbox Challenge Winner [12].

Mindpixel is a common sense component and it is similar to OpenMind.[8] and Cyc[9] The system accepts public contributions. However, Cyc model and OpenMind had a bottleneck which prevent truly large-scale collaboration [46]. 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 10 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 generalisation test to ensure the system did 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 hundreds-of-millions to billions of "pieces of knowledge" that are estimated to be involved with human intelligence [38].

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 conversation bots, factoid questions should have only single factual answers [2,15,18,54]. This will be 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 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), a Loebner Prize winning [36] ALICE conversation bot knowledge base. AAA is a free open-source software package based on XML specifications. It is a set of Artificial Intelligence Markup Language (AIML) scripts comprising the award winning conversation bots. The AAA is specifically reorganised to make it easier for conversational system developers to clone the conversation bot's 'brain' and to create custom conversation agent personalities, without having to invest huge efforts in editing the original AAA content. AAA's knowledge bases covered a wide range of subject domains based on the bot's personality. Example subjects include AI, games, emotion, economics, film, books, sport, science, epistemology and metaphysics.

ALICE won the 2000, 2001 and 2004 Loebner Prize for being the most lifelike machines. The competition uses the Turing Test [1], named after British mathematician Alan Turing, to determine if 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 Artificial Intelligence Markup Language (AIML). The development of ALICE is based on the fact that the distribution of sentences in conversations tend to follow Zipf's Law [33]. 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, which used to require many hours of Dr. Zipf's labor, is now accomplished in a few milliseconds of computer time. While the possibilities 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.

6.2. 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 [20] currently has a training corpus of more than one trillion words (1,024,908,267,229) from public web pages. This is valuable for our type of research. The web is a potentially unlimited source of knowledge data; however, commercial search engines are not the best way to gather answers from queries due to the overwhelming number of results from a search.

Before the rise of domain-oriented conversation bots based on natural language understanding and reasoning, evaluation was never a problem, as information retrieval-based matrices were readily available for use. However, when conversation bots begin to become more domain-specific, evaluation becomes a real issue [22]. This is especially true when Natural Language Processing (NLP) is required to cater for a wider variety of questions and, at the same time, to achieve high-quality responses.

As shown in Fig. 3, AINI's domain-specific knowledge base consists of Natural Language Corpus and Frequently Asked Questions (FAQ). Both components are extracted from the online documents using an Automated Knowledge Extraction Agent (AKEA) [23]. AKEA was aimed at restricted knowledge base, in particular pandemic domains and exploiting linguistic knowledge from the documents and natural language knowledge about a specific domain. It was aimed at providing up-to-date information to its users via AINI. Another objective of AINI is to deliver essential information from trusted sources and it should be capable of interacting with its users. The idea is to rely on a human-like communication approach, thereby providing a sense of familiarity and ease.

6.2.1. 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 [51]. This is to avoid rumours, hoaxes and misinformation. In our experiment, the selection of the trusted websites is based on PageRank^{TM}[9]. PageRank^{TM} is a system for ranking web pages developed by Larry Page and Sergey Brin at Stanford University. PageRank^{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^{TM}.

In our study, the selection of trustworthy websites starts with the initial 6 seed words: *bird, flu, avian, influenza, pandemic*and *H*5*N*1. 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[10] and BootCat Tool [8] for bootstrapping corpora and terms from the web. We use the seeds to perform a set of Google searches, asking Google to return a maximum of 20 URLs per query and get the first corpus. After visual inspection of the corpus, we used the top 40 seeds extracted from token frequencies for the second run. Finally, we retrieved 1428 URLs out of 1500 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 1428 URLs, we sent a query to Google's PageRank^{TM} directory using PaRaMeter Tool [14] to determine their rankings. Fig. 4shows the results of the top 10 site based on the PageRank^{TM}scale. The PageRank^{TM} scale goes from 1 to 10. A less important site is one with a PageRank (PR) of 1. The most referenced and supposedly important sites are those with a PR of between 7 and 10.

The final set of URLs was further culled to include only selected sites 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^{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 will be considered as "stronger" or "more reliable". This effectively gives the linked site "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.

6.2.2. Extracted online documents

In AINI's domain knowledge Matrix, 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 form. The information is obtained by AKEA. We have also developed a new system for diseases based on symptoms, causes, diagnoses, vaccination locations, persons and organizations. In order to identify these entities, our system uses rules in which we specify the named entities' structure in term of text tokens and what can be found out about them from resources 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 automated ways. An example of an application of AKEA is to prepare an answer for a user's query. The ontology and gazetteer has been implemented as domain-dependent modular components which will allow future improvements and to maintain openness 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^{TM} as discussed above. But at this stage, each of the selected websites was evaluated in order to find the most 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 complete database which was built by AKEA after applying a filtering process to remove unnecessary tags. Results from our work are discussed in the next section.

As shown in Table 2,[11] AINI's open-domain knowledge currently has more than 150,000 entries in the commonsense 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, AINI's had more then 1000 online documents extracted using AKEA. This makes up 10,000 stimulus–response categories 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 updated to 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. The original and simple conversational programs such as ELIZA [53], 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.

7. Experimental setting

This paper examines domain knowledge query of conversational logs collected from the conversational system between AINI and MSN Messenger online users. The study is based on a corpus of instant messaging texts using MSN Messenger on the DesktopChat, WebChat and MobileChat, which was collected from May 15 to September 15, 2006 during an Invasion of the Robots Contest.

7.1. Participants and corpus

The experiment’s portal[12] is open to the public from all over the world who can access this portal and freely participate in the study. This portal allows the online users to add AINI’s contact to their “buddy-list”, by allowing them to easily send and receive short textual messages. When a participant opens a message window to a buddy for the first time (and that buddy was online), an alert is sent to the buddy notifying them of their participation in the study. Participation is voluntary. In the conversation log files, the nickname, MSN account, date and time of the dialog, as well as the spoken texts (inputs and outputs), are recorded. During a conversation, we created a unique ID for each buddy and stored the ID of the buddy instead of the buddy-account itself. These measures were employed to protect the privacy and confidentiality of the participants.

Previous research has shown significant differences in IM communication resulting from the frequency of communication [5,31]. In this study, we use word frequency for our analysis of the corpus. We processed 29,447 words of running text and there are 2541 unique words, 129,760 characters, 4251 sentence count are recorded. From this data, we collected a total of approximately 63 h of recorded data, observing over 3280 outgoing and incoming instant messages exchanged with over 65 buddies (only three of them use MSN Mobile). The average sentence length of an IM transmission was 6.90 words, with approximately 13% of all transmissions being a single word in length. Table 3 provides a summary of data collected.

The participant gets to know AINI at the MSN from the advertisement from eight famous BBS (bulletin board systems), which include blog websites and the AINI portal. We gathered usage data via automatic logging on the server, which included logins, logouts, joining, as well as chat messages.

7.2. Chatlog system

We have developed a Chatlog System used MySQL which stores user messages to secondary storage. It provides real-time archiving to IM clients that captures chat messages so they can be searched by keyword or user ID, and allows topic-based retrieval of chat sessions. These chat messages are essentially plain text messages that are quite small in comparison with images, video, or even documents. These plain text messages, also known as instant messages, are the regular messages sent between principals on MSN Messenger. Sending a plain text message instructs clients receiving the message to display it on-screen, optionally with some simple formatting.

The fastest typists can enter well over 150 words per minute. Many jobs require keyboard speeds of 60–70 words per minute [6]. However, the actual typing speeds for human users during chatting are much slower. By using time-stamping, we calculated both transmission rates from the corpus. On average, there were 3.5 transmissions per minute. Given that the mean transmission length was AINI 6.46 words and human 7.85 words, AINI averaged 22.61 words per minute and IM human 27.47 words per minute. For both transmissions between AINI and IM human users, the average message exchanged speed was approximately 25 words per minute. With this speed, even if a user chatted average 4 h a day, it would only require approximately 100 kB of storage per day. An entire year's worth of chat would use 36 MB of storage which can be easily handled with today's storage technologies.

8. Results and discussion

In this section, some of the interaction features of the recorded chat are discussed and domain knowledge query will be our focus analysis. Studies of text chat have tended to focus on the interaction problems caused by the properties of text chat. This research seeks to examine the effect of domain knowledge on the relationship between human users and conversation bots in the context of MSN Messenger. To be more specific, in this study, we defined how domain knowledge plays important roles in the textual communication via the Internet between at least two "participants"; of which at least one is a human user communicating with the AINI conversation bot. It begins by identifying general features of the texts collected.

8.1. Conversation logs

In our study, the experiment on the Domain Matrix Knowledge Model has shown interesting results from the natural conversation bots. The key assumption is that important queries do not necessarily turn up all the answers that can be found in a single domain, but that appropriate answers may also come from different domains.

The excerpt shown in Fig. 5 is from a typical single session IM conversation exchanged between AINI and one of "her" buddies with ID U0025. This session illustrates the nature of the IM communication. Each new session will start with AINI being given a random greeting (message #1) eg. "Hi there:)", "How are you today?", "Hey:), "nice to meet u". How I can call u?", etc. These greetings will indirectly get some information about the user's identity, such as their name or gender (utterance #2). In this session, we managed to identify U0025 user as "Hommer" and a "male". Although "Hommer" used an emoticon to represent the first initial of his name "(H)", which refers also to "hot smile = F[x]"; however, AINI defined him as "Hommer". Following some dialogue (utterance #3), AINI gave a greeting to "Hommer" which came from open-domain knowledge. Since AINI's knowledge is not equipped with full IM features such as acronyms, abbreviations and emoticons, utterance #5 shows that AINI failed to recognize the user input "sh F[x]", which refers to "Same here" and "(smile)" emoticon. This drawback caused AINI to return to the random answer domain.

The next message (utterance #6–15) shows user U0025 trying to challenge AINI by putting forward questions such as "Who was the first American in space?" and "what is bird flu?". This challenge could definitely be answered by AINI, because "she" is fitted with TREC factoids, questions and answers on the subject of the bird flu pandemic from domain-specific. In the final session (utterance #15–21), both participants ended with greetings, and AINI replied with the simple abbreviation "*TTYL*" (*Talk to you later*) and the intonation "Gee". These transmissions constitute a single session and also include the use of shorthand, acronym, abbreviations, loose grammar and minimal punctuation in IM as has been shown to be standard by the previously cited research [4,5,17,32].

8.2. Topicality

After a thorough examination of the logs of over 3280 utterances, we found that human–machine dialogues have discussed almost every aspect of everyday life. These topics include emotion, love, sex, computers, entertainment, sport, etc. As shown in Fig. 6, in every category, several detailed topics have been discussed. Almost 39.4% of the IM exchanges have discussed emotional problems; these include friendship, sex and love. This finding is remarkable as the AINI conversation bots are trained to simulate a human partner in IM. This was because the IM users, who are mostly young people, wanted to tell AINI some private problems and experiences. Even in the dialogues, some of them praised AINI, invited “her” on a date, and some of them disclosed their personal problems. However, as expected, about 17.7% invited AINI to talk about robot technology and some even tried to test AINI’s intelligence by arguing with “her”, and some of them tried to cheat. This may be explained by the fact that almost 50% of the human IM users came to “know” AINI from the Invasion of the Robots Contest websites and blogs. Within the group, few are conversational bot developers or programmers who participate in the contest. There are 53 bot programmers competing in this contest. In addition, some of them realised that they were talking with a robot or a computer program after a short period of chatting with AINI.

8.3. Domain knowledge oriented

As discussed earlier, AINI’s domain knowledge model incorporates several knowledge domains, thus merging the expertise of one or more experts. A ‘sales’ knowledge domain for instance, would contain expertise on improving sales, but it would also incorporate open-domain knowledge. Multiple domain knowledge, merged into AINI’s single domain knowledge would give the users the best conversation as show in Fig. 7.

From the conversation logs with 1721 utterances, we found that AINI used 88.03% of the knowledge from open-domain knowledge and only 2.15% from domain-specific knowledge bases. In this experiment, we did not restrict the conversation to the domain-specific on the SARS epidemic [21] and Bird Flu Pandemic [27]. However, in this research, AINI’s domain knowledge was equipped with crisis communication knowledge bases which were included in the Natural Language Corpus and FAQ extracted from online documents using Automated Knowledge Extracted Agent (AKEA) [23]. The notions of range and dispersion are well illustrated by two words with roughly equal overall frequency: *SARS* and *Bird Flu*. These words appear to be rather specialised terms, used in a restricted number of conversations in contrast with other words, which are more widely used. AINI’s “buddies” get to know availability of these domain knowledge bases from AINI’s Crisis Communication Network portal (CCNet) available online at http://ainibot.murdoch.edu.au/ccnet. The users took opportunities to ask questions related to this crisis communication. Therefore, about 80% of the 37 questions are related to *diagnose treatment, symptoms, spread, protection, cause, vaccination* and *risk* of the SARS epidemic or Bird Flu pandemic.

On the other hand, AINI holds 85.7% of its conversation from the AAA’s knowledge bases. This did not surprise us, because in the AAA’s knowledge bases, the topics cover almost everything; including emotion, sex, literature, music, religion, science, sports, etc. More than 45,318 AAA stimulus–response categories are stored in AINI’s knowledge base. Each category contains a stimulus–response (also called input-pattern) and an output-template. More of AINI’s stimulus–response categories come from the commonsense, TREC and Mindpixel corpora. Although commonsense stimulus–response categories hold the majority of AINI’s knowledge bases (49%), only 2.32% of the total responses are related to the commonsense questions. Despite commonsense questions playing a major role in formal conversation, AINI’s “buddies” are normally more interested in issues of daily life or personal interest, instead of the factoid questions which are provided in TREC and the Mindpixel corpus.

AINI’s query engine works based on the natural language query: if a matching category is found in the knowledge bases, it will be retrieved and be transformed to the output. If no matching category is found, AINI’s query engine will send the request to the Unanswered Domain knowledge base shown in Fig. 3. In this case, AINI will generate a random answer. These replies were irresponsible, inappropriate, amusing and thoughtless responses and comprised 9.82% of the total output of the IM conversational bot. Obviously, these expressions are irrelevant and unrelated and make AINI’s “buddies” feel irritated and confronted by AINI. These expressions occur because of the differences in manners of speech and speech acts (e.g. declarative, interrogative or imperative or exclamatory). As discussed earlier, human IM users have a tendency to use shorthand, acronyms, abbreviations and emoticons. Unfortunately, AINI was not trained to understand such expressions in the short period of time in which this study was conducted. However, AINI is capable of learning from the users and domain experts. The unanswered questions will be maintained separately by a domain expert or 'botmaster' who will keep AINI's knowledge bases updated regularly. Our domain model has designed such a way to make sure, in subsequent sessions of conversation, AINI will 'understand', and should be able to participate in a meaningful conversation.

All these perspectives should be considered in the design of conversation bots. Conversation bots should not only work as specific purpose bots with rich special knowledge, but also as friendly chat companions who may experience the joy and suffer the pain of the users. In the case of IM conversation bots, IM human users wish to express their emotions through emoticons, abbreviations and acronyms; features which need to be added into AINI's knowledge bases.

9. Conclusion and future work

This paper emphasizes two contributions. The first contribution of our proposed system is the construction of a novel Domain Matrix Knowledge Model for AINI. This model is to establish metamorphosis in a conversation agent; which is something that's good not just for knowledge customization for an AINI, but also to converse naturally and efficiently with a human user. The second contribution follows by developing techniques to retrieve and dynamically construct web knowledge from semi-structured data. We implemented this technique and algorithms in the AKEA system, an intuitive and easy to use tool that will alleviate the work of domain experts in the construction and maintenance of complex knowledge bases. We believe that in spite of the use of semi-structured data as the source for knowledge, we achieved an acceptable degree of confidence identifying and matching knowledge on the web.

Based on this experiment, IM conversation between human users and machines shows an interesting pattern of behaviour in the natural conversation bots. In this paper, we only worked based on the MSN Messenger application run on top of multi-domain knowledge bases. Although we simulated the proxy conversation log that contains clients' requests, there is a possibility that new simulations resulting from other traces may differ from the results referred to in this paper.

Our study suggests that IM human–machine conversations display considerable variation both with and across machine and IM human users. From the conversation logs, AINI's "buddies" seem interested in chatting about personal issues, emotion, love, sex, computers, entertainment, etc. These dialogue traits comprise 85.7% of the AAA's knowledge bases. Although commonsense stimulus–response categories comprise the majority of AINI's knowledge bases (49%), human IM users appear to be focused on current everyday life domain knowledge, instead of factoid questions. As with human knowledge, AINI's knowledge also has limitations. Obviously, about 9.82% of the total questions asked by human IM users are not contained in AINI's knowledge bases. Instead of empty strings or infinite replies, AINI generated random answers. Although these replies were mainly inappropriate and amusing, in a few cases they created a new topic of discussion, and prolonged the conversation. Evidence also suggests that AINI's "buddies" are interested in chatting with bots just to seek information, to be friends, to express their emotions, and some just want chat for leisure. Thus, AINI was successful in imitating human conversation through human-like artificial intelligence. Though the standard of conversation isn't exactly astonishing, the bot's responses are "human" enough to make its IM "buddies" feel a sense of companionship.

Nevertheless, we predict the problems of infinite replies could be overcome with appropriate programming using natural language intelligence sentence parsing, and massive but tailor-made databases to provide sufficient knowledge to the bots. For instance, new categories of stimulus–

response for IM knowledge bases such as emoticons, acronyms, abbreviations or shortcuts, could be added into AINI's domain-specific knowledge. We plan a new data collection phase for the near future in order to examine the application of the results presented here with a new framework and set of hypotheses which are more robust and comprehensive.

Acknowledgement

This research project is funded by Murdoch University Research Excellence Grant Scheme (REGS), 2006/07.

References

[1] A.M. Turing Computing Machinery and Intelligence. MIND the Journal of the Mind Association, LIX (236) 433–460.

[2] E. Agichtein, S. Cucerzan, E. Brill, Analysis of factoid questions for effective relation extraction, in: 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, 2005.

[3] Alicebot, The Annotated A.L.I.C.E. AIML, Alicebot.org, 2006.

[4] C. Alphonso, Texting helps teens' grammar, Bell Globemedia, 2006.

[5] D. Avrahami, S.E. Hudson Communication characteristics of instant messaging: effects and predictions of interpersonal relationships CSCW'06, CM Press, Banff, Alberta, Canada (2006)

[6] B. Bailey Human Interaction Speeds (2000)

[7] N. Baron Instant messaging and the future of language Communications of the ACM, 48 (7) (2005), pp. 29-31

[8] M. Baroni, S. Bernardini, Boot CaT: bootstrapping corpora and terms from the web, in: Fourth Language Resources and Evaluation Conference, 2004.

[9] M. Bianchini, M. Gori, F. Scarselli Inside PageRank, ACM Transactions on Internet Technology ACM Transactions on Internet Technology, 5 (1) (2005)

[10] Z.Z. Bien, H.-E. Lee Effective learning system techniques for human–robot interaction in service environment Knowledge-Based Systems, 20 (5) (2007), pp. 439-456

[11] B.S. Boneva, A. Quinn, R.E. Kraut, S. Kiesler, I. Shklovski Teenage communication in the instant messaging era R. Kraut, M. Brynin, S. Kiesler (Eds.), Computers, phones, and the Internet: Domesticating information technology, Oxford University Press, New York (2006), pp. 201-218

[12] Chatterboxchallenge, ALICE Winner of Chatterbox Challenge 2004, 2006.

[13] S.M. Cherry IM Means Business IEEE Spectrum Magazine (2002)

[14] CleverStat, PaRaMeter, 2006.

[15] K. Collins-Thompson, E. Terra, J. Callan, C. Clarke The effect of document retrieval quality on factoid question-answering performance IGIR2004, Sheffield, UK (2004)

[16] D. Craig Instant messaging: the language of youth literacy The Boothe Prize Essays (2003)

[17] S.R. Crenzel, V.L. Nojima, Children and instant messaging, IEA 2006, in: 16th World Congress on Ergonomics, Maastricht, Netherlands, 2006.

[18] S. Cucerzan, E. Agichtein Factoid Question Answering over Unstructured and Structured Web Content, Microsoft (2005)

[19] R.C. Forgas, J.S. Negre, The use of new technologies amongst minors in the Balearic Islands, in: IAARE Conference, Melbourne, 2004.

[20] A. Franz, T. Brants All our N-gram are Belong to You Google Machine Translation Team (2006)

[21] O.S. Goh A Global Internet-based Crisis Communication: A Case study on SARS using Intelligent Agent Kolej Unversiti Teknikal Kebangsaan Malaysia, Malacca, Malaysia (2005)

[22] O.S. Goh, C. Ardil, W. Wong, C.C. Fung A black-box approach for response quality evaluation conversational agent system International Journal of Computational Intelligence, 3 (3) (2006), pp. 195-203

[23] O.S. Goh, C.C. Fung, Automated knowledge extraction from Internet for a crisis communication portal, in: First International Conference on Natural Computation, Lecture Notes in Computer Science (LNCS), Changsha, China, 2005, pp. 1226–1235.

[24] O.S. Goh, C.C. Fung, C. Ardil, K.W. Wong, A. Depickere, A crisis communication network based on embodied conversational agents system with mobile services, Journal of Information Technology 3 (1) 257–266.

[25] O.S. Goh, C.C. Fung, A. Depickere, K.W. Wong, W. Wilson, Domain knowledge model for embodied conversation agent, in: Third International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2005), Singapore, 2005.

[26] O.S. Goh, C.C. Fung, M.P. Lee Intelligent agents for an Internet-based global crisis communication system Journal of Technology Management and Entrepreneurship, 2 (1) (2005), pp. 65-78

[27] O.S. Goh, C.C. Fung, K.W. Wong, Embodied conversational agents for H5N1 pandemic crisis, Journal of Advanced Computational Intelligence and Informatics 11 (2007) (Special Issue on Advances in Intelligent Data Processing).

[28] Y. Hård af Segerstad, S. Sofkova Hashemi, Exploring the writing of children and adolescents in the information society EARLI SIG writing, in: Ninth International Conference of the EARLI – Special Interest Group on Writing, Geneve, Switzerland, 2004, pp. 20–22.

[29] J.D. Herbsleb, D.L. Atkins, D.G. Boyer, M. Handel, T.A. Finholt **Introducing instant messaging and chat in the workplace** CHI'2002, ACM Press, Minneapolis, Minnesota, USA (2002)

[30] T.M. Holtgraves, S.J. Ross, C.R. Weywadt, T.L. Han Perceiving artificial social agents Computers in Human Behavior, 23 (2007), pp. 163-2174

[31] E. Isaacs, A. Walendowski, S. Whittaker, D.J. Schiano, C. Kamm The character functions and styles of instant messaging in the workplace CSCW'02, ACM Press, NY (2002), pp. 11-20

[32] S. L'Abbé Instant msg-ing Messes With Grammar? As If! lol! Teens Adopting Unique Linguistic Shorthand But Not Ruining Syntax University of Toronto (2006)

[33] W. Li, Zipf’s Law, Feinstein Institute for Medical Research, 2006.

[34] D. Lin, Dependency-based Evaluation of MINIPAR, in: Workshop on the Evaluation of Parsing Systems, Granada, Spain, 1998.

[35] S. Livingstone, UK Children Go Online: Surveying the experiences of young people and their parents, 2006.

[36] H. Loebner Loebner Prize (2006)

[37] MindPixel, GAC-80K Mindpixel, 2006.

[38] E. Mueller Common Sense in Humans (2001)

[39] R. Naraine I Want My Active Buddy (2001)

[40] B. Nardi, S. Whittaker, E. Bradner, Interaction and outeraction: instant messaging in action, in: CSCW’2000, 2000, pp. 79–88.

[41] C. Neustaedter, A 3D Instant Messenger Visualization Using a Space Metaphor, Calgary, Alberta, Canada, 2001.

[42] NIST, Text REtrieval Conference (TREC), 2006.

[43] J. O’Neill, D. Martin, Text chat in action, in: GROUP’03, Sanibel Island, Florida, USA, 2003.

[44] R. Planta, S. Murrell A natural language help system shell through functional programming, 18(18) (2005), pp. 19-35

[45] K.R. Hoffman Messaging Mania in Time for Kids World Report, 8 (25) (2003)

[46] M. Richardson, P. Domingos, Building large knowledge bases by mass collaboration, in: K-CAP’03, ACM Press, Sanibel Island, Florida, USA, 2003.

[47] Rovers, A.F. and Van Essen, H.A., Him: A framework for haptic instant messaging, in: CHI, Viena, April 2004.

[48] E. Shiu, A. Lenhart How Americans use instant messaging Pew Internet & American Project (2004)

[49] C. Uhrhan, O.S. Goh, Features of a mobile personal assistant robot, in: International Conference on Robotics, Vision, Information and Signal Processing, IEEE-ROVISP 03, Penang, Malaysia, 2003, pp. 340–345.

[50] USAToday, Agents pursue terrorists online, USA Today, June 24, 2002.

[51] J.D. Vela’squez, V. Palade A knowledge base for the maintenance of knowledge extracted from web data Knowledge-Based Systems, 20 (2007), pp. 238-248

[52] J. Weizenbaum **Computer Power and Human Reason** W.H. Freeman and Company (1976)

[53] J. Weizenbaum, ELIZA – a computer program for the study of natural language communication between man and machine, Communications of the ACM 9 (1) 36–45.

[54] E.W.D. Whittaker, J. Hamonic, D. Yang, T. Klingberg, S. Furui, Monolingual Web-based Factoid Question Answering in Chinese, Swedish, English and Japanese, in: EACL 2006 Workshop on Multilingual Question Answering – MLQA06, 2006.

[55] T. Winograd Understanding Natural Language Academic Press (1972)

[1] https://www.robotinvaders.com.

[2] http://www.microsoft.com/uk/press/content/presscentre/releases/2006/06/PR03717.mspx.

[3] https://www.robotinvaders.com.

[4] https://www.buddyscript.colloquis.com.

[5] http://www.incesoft.com.

[6] http://www.improvcenter.com/msn.

[7] http://hypothetic.org/docs/msn/.

[8] www.openmind.org.

[9] www.cyc.com.

[10] www.alicebot.org/aiml/aaa/.

[11] http://www.google.com/apis.

[12] The stimulus–response categories of the AINI's knowledge bases calculated on 1 December 2006 and can be access at the experiment portal http://ainibot.murdoch.edu.au.

Fig. 1. AINI and MSN authentication process.

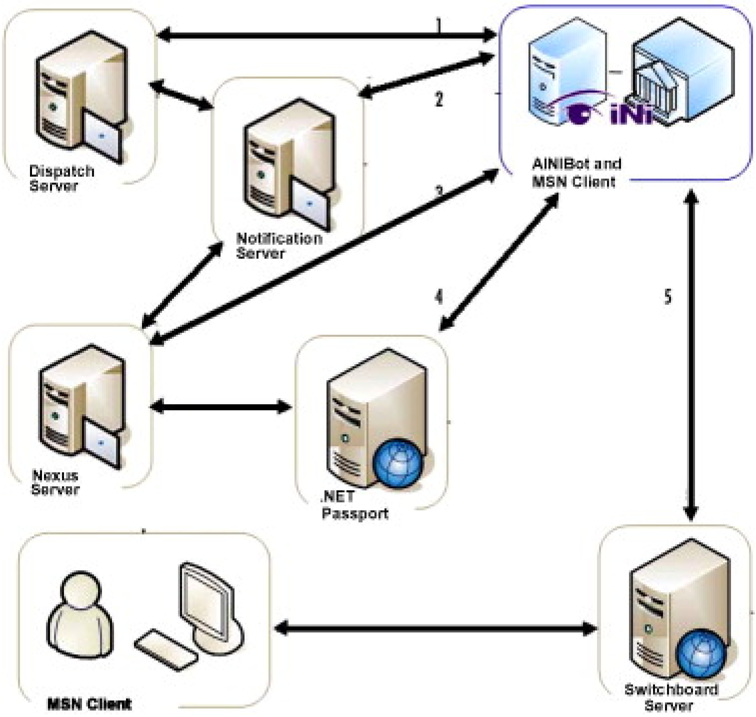


Fig. 2. AINI and MSN messenger interface.

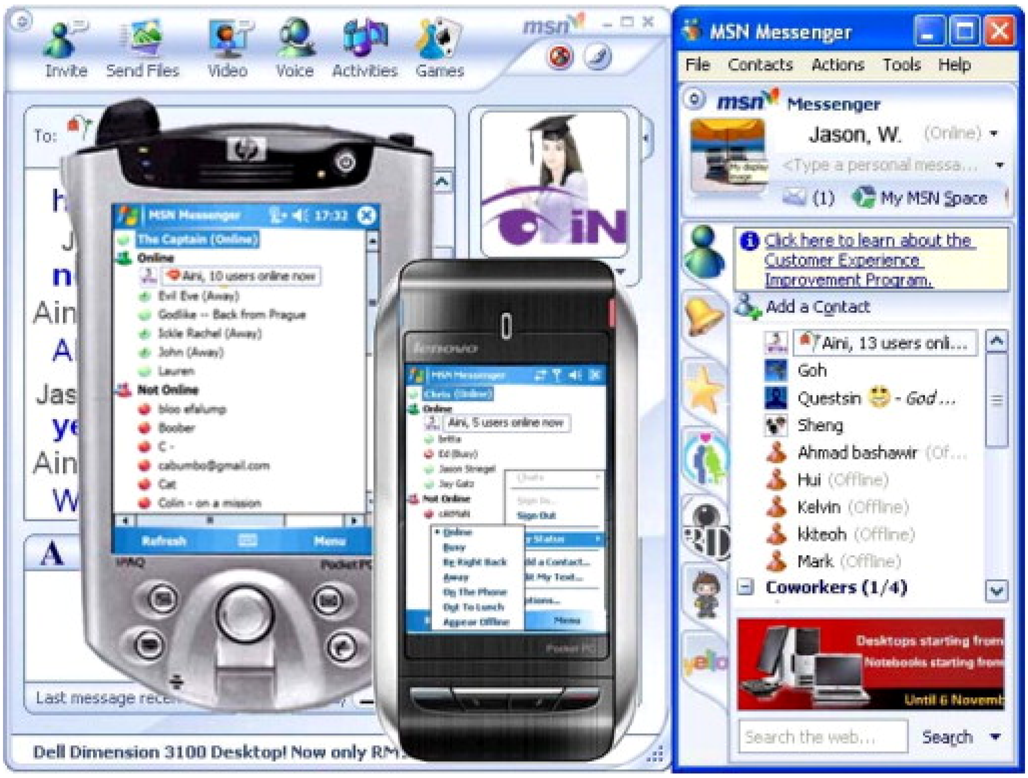


Fig. 3. AINI's Domain Matrix Knowledge Model.

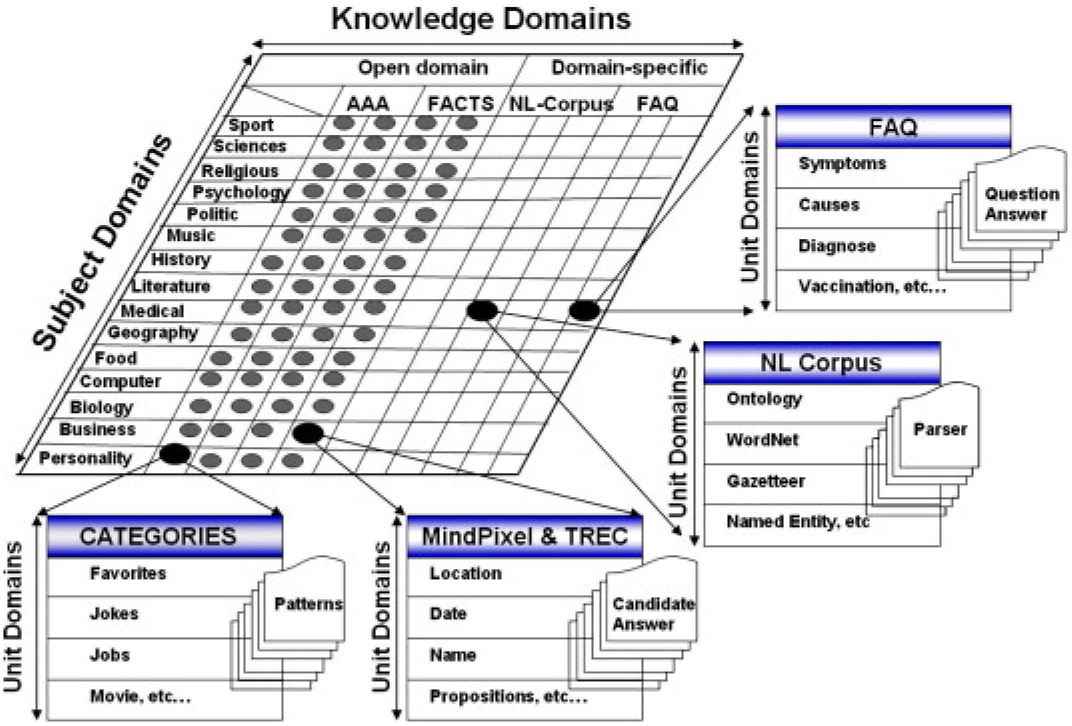
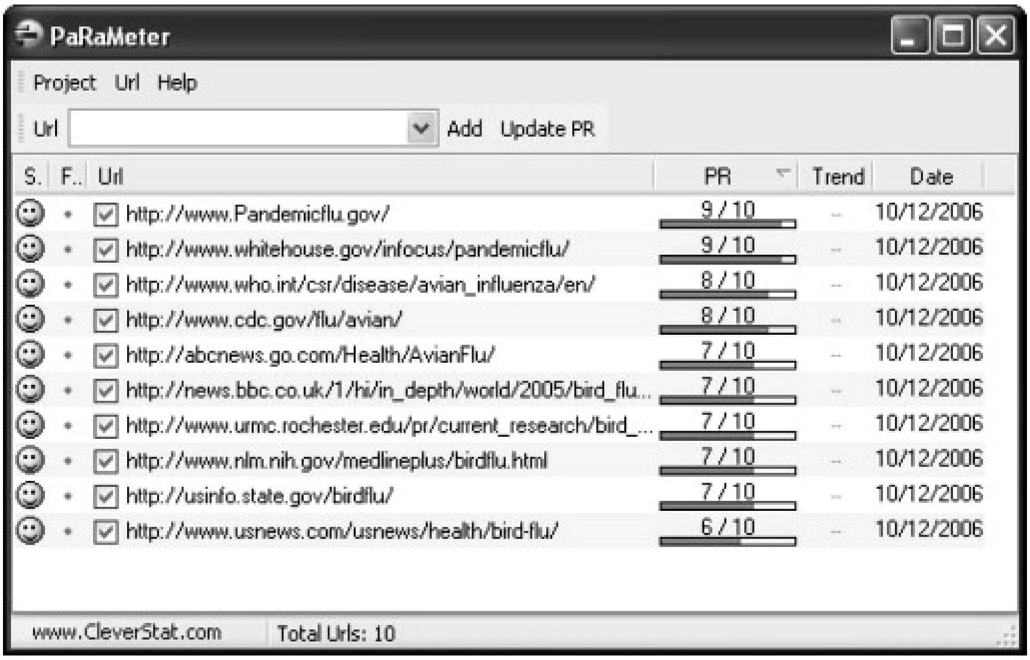


Fig. 4. Top 10 Pagerank^{TM} scale for bird flu domain



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

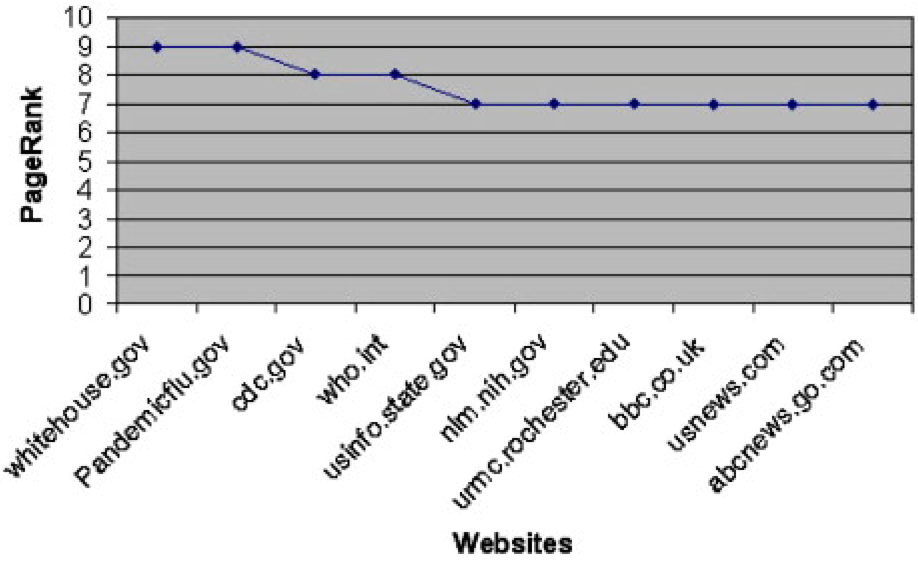
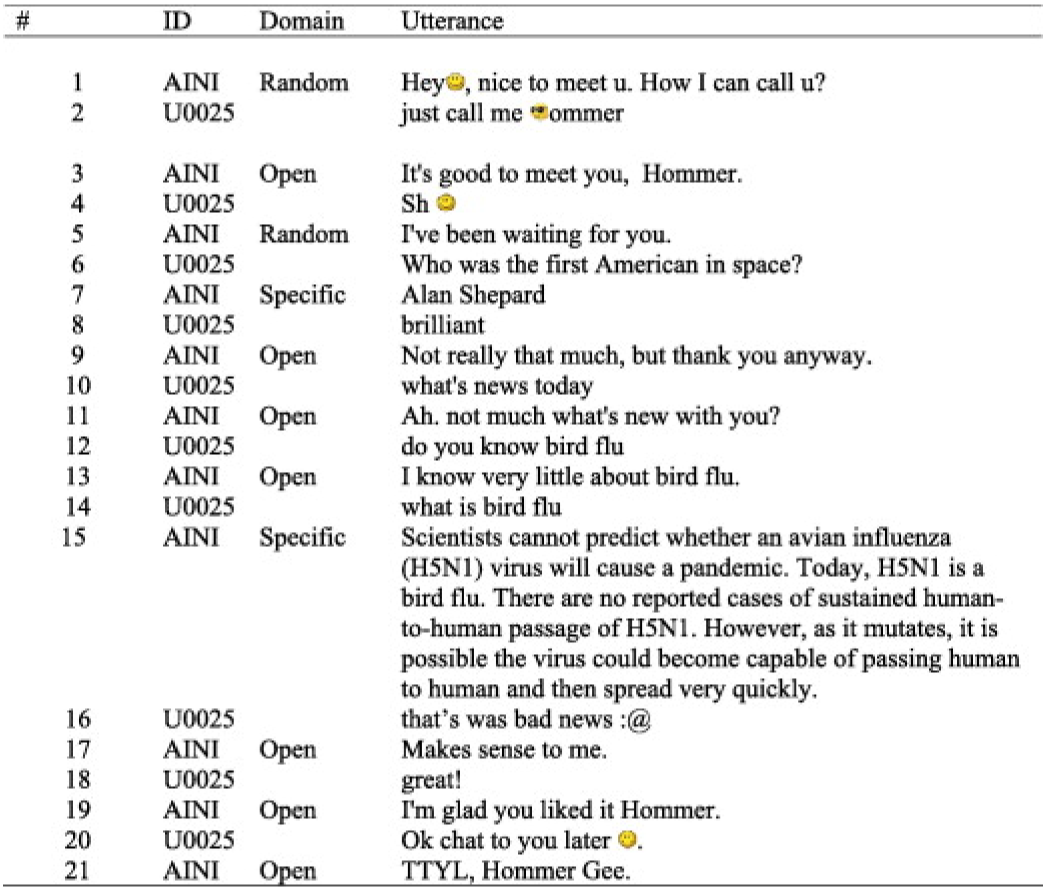


Fig. 5. AINI conversation log with U00301 IM human user.



#	ID	Domain	Utterance
1	AINI	Random	Hey😊, nice to meet u. How I can call u?
2	U0025		just call me 😎ommer
3	AINI	Open	It's good to meet you, Hommer.
4	U0025		Sh 😊
5	AINI	Random	I've been waiting for you.
6	U0025		Who was the first American in space?
7	AINI	Specific	Alan Shepard
8	U0025		brilliant
9	AINI	Open	Not really that much, but thank you anyway.
10	U0025		what's news today
11	AINI	Open	Ah. not much what's new with you?
12	U0025		do you know bird flu
13	AINI	Open	I know very little about bird flu.
14	U0025		what is bird flu
15	AINI	Specific	Scientists cannot predict whether an avian influenza (H5N1) virus will cause a pandemic. Today, H5N1 is a bird flu. There are no reported cases of sustained human-to-human passage of H5N1. However, as it mutates, it is possible the virus could become capable of passing human to human and then spread very quickly.
16	U0025		that’s was bad news :@
17	AINI	Open	Makes sense to me.
18	U0025		great!
19	AINI	Open	I'm glad you liked it Hommer.
20	U0025		Ok chat to you later 😊.
21	AINI	Open	TTYL, Hommer Gee.

Fig. 6. Frequency of dialogue topics.

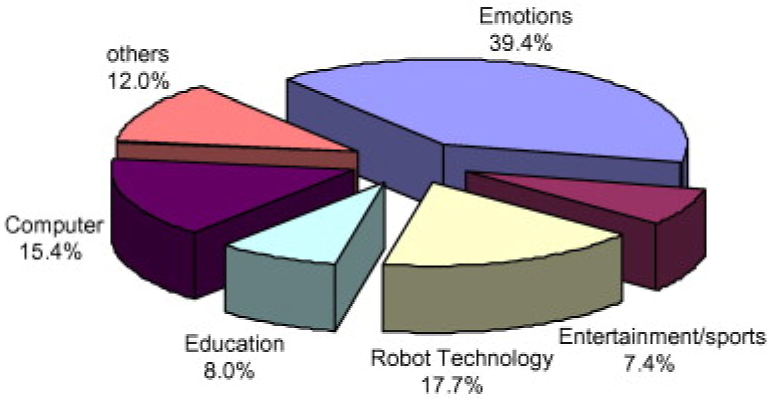


Fig. 7. Frequency of AINI's responses based on domain knowledge base.

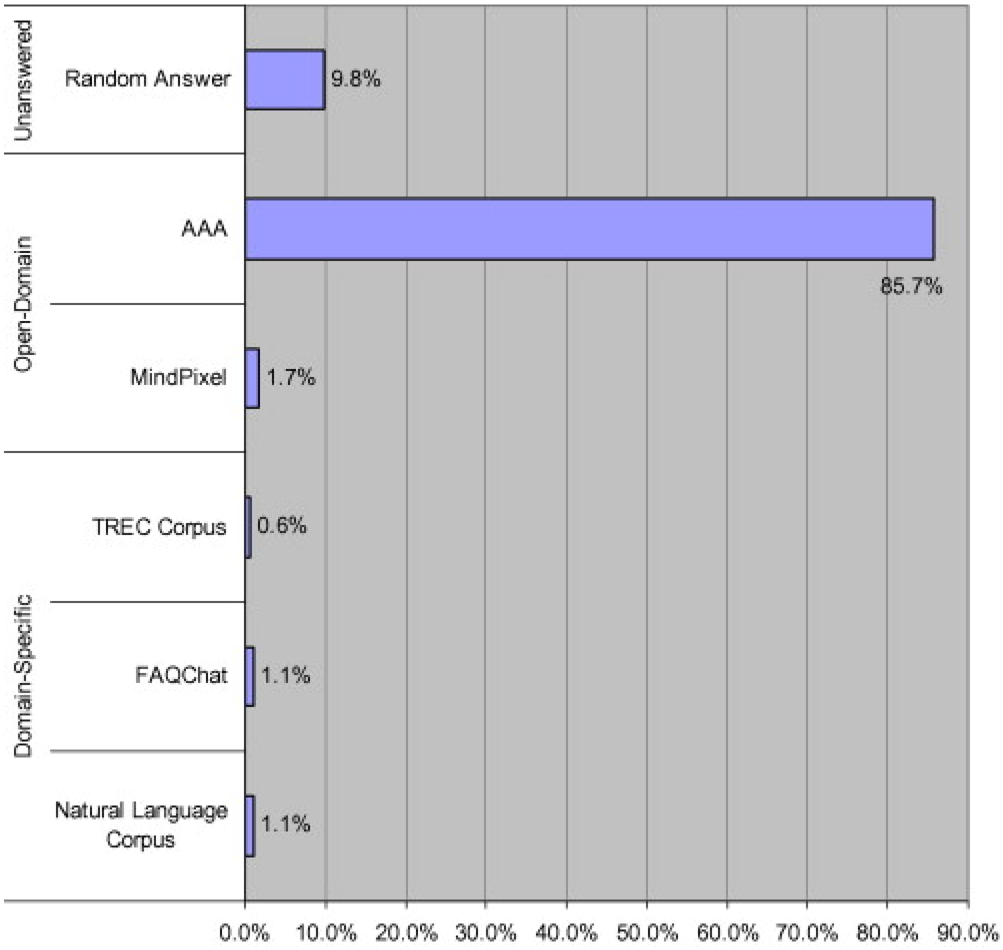


Table 1. Factoid question from TREC 8, 9 and 11

TREC	**Factoid question**	**Text research collection**
8	196	• Financial Times Limited (1991, 1992, 1993, 1994) • The Congressional Record of the 103rd 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 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 and AskJeeves logs donated by Ask Jeeves

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	

Table 3. Frequency of word from conversation logs

	AINI	**Human**	**Total**
Word	18,358	11,089	29,447
Unique word	1368	1173	2541
Character count	79,884	49,876	129,760
Sentence count	2840	1411	4251
Utterance	1721	1559	3280
Average sentence	6.46	7.85	6.90