Leveraging Question Answer Technology to Address Terrorism Inquiry

Robert P. Schumaker, Hsinchun Chen
Artificial Intelligence Lab, Department of Management Information Systems
The University of Arizona, Tucson, Arizona 85721, USA
{rschumak, hchen}@eller.arizona.edu

Word Count: 7512

Abstract

This paper investigates the potential use of dialog-based ALICEbots in disseminating terrorism information to the general public. In particular, we study the acceptance and response satisfaction of ALICEbot responses in both the general conversation and terrorism domains. From our analysis of three different knowledge sets: general conversation, terrorism, and combined, we found that users were more favorable to the systems that exhibited conversational flow. We also found that the system that incorporated both conversation and terrorism knowledge performed better than systems with only conversation or terrorism knowledge alone. Lastly, we were interested in what types of questions were the most prevalently used and discovered that questions beginning with 'wh*' words were the most popular method to start an interrogative sentence. However, 'wh*' sentence starters surprisingly proved to be in a very narrow majority.

Index Terms

Dialog Platform, Knowledge Delivery, Evaluation, Domain-Specific Knowledge, chatterbot, ALICE, XML, AIML

Formatted: Font: 11 pt
Deleted: ,

1 Introduction

Since the events of September 11th, terrorism education and awareness has become a national priority. The Department of Homeland Security (DHS) has written many training guides instructing first responders and citizens how to react in a terrorist event. DHS has also launched several initiatives in this area, such as Community Emergency Response Teams (CERT) and Ready.gov that contain online preparation resources. Independent organizations such as the American Red Cross have also prepared terrorism education materials which instruct citizens in basic emergency procedures. Preparing citizens to act in a terrorism incident is important because citizens are often the first ones to respond to a scene and can be highly effective during an emergency [1].

As a part of mobilizing citizen and post-terrorism response initiatives, two post Sept. 11th reports, "Making the Nation Safer" and "The 9-11 Commission Report," present the importance of a mobile Command, Control, and Communications (C3) system as a vehicle of information delivery. The C3 system would be quickly deployed to the scene of a terrorist attack and would be responsible for both providing emergency communications channel capacity for first responders, and acting as a focal point for information collection, management, and decision-making tasks. It would also have the ability to provide communications to the public by providing citizens with timely information that could aid in rescue. One of the secondary roles of a C3 system is to get information to the public and fill the void before rumors, hoaxes, and misinformation become mainstream [2].

One system with promise to meet many of the C3 system guidelines is ALICEbot. ALICEbots are a class of Question Answer programs developed in 1995 by Richard Wallace [3]. These Question Answer systems can be programmed with terrorism-specific knowledge and

quickly deployed in a terrorism-affected area. ALICEbots have the advantage of holding an immense library of knowledge, they can converse in the English language, and are robust and scalable in addressing new areas of knowledge. This class of system shows promise in the field of terrorism education by allowing citizens to seek specific terrorism knowledge as well as disseminate real-time critical knowledge to those in need, such as family members, in a more interactive way other than a unidirectional television broadcast.

One of the major hurdles to using ALICEbots and Question Answer systems in general, has been returning responses that are appropriate and acceptable to the context of the question. With the many different ways to communicate information and structure sentences, computer systems have had major difficulty in the basic understanding of language. The unique aspect of ALICEbot has been its ability to mimic conversational English without regard to understanding. This primitive approach to contextual accuracy has garnered ALICE the Loebner Prize for most human computer in 2000, 2001, and 2004. This simple conversational technique has made the goals of attaining a working C3 system appear to be within grasp.

This rest of this paper is organized in the following sections. Section 2 is a literature review and talks about how ALICEbots, Question Answer, and C3 systems are all related to each other. Section 3 asks a series of research questions about the retrieval potential of ALICEbots. Section 4 introduces the reader to the Terrorism Activity Resource Application (TARA) system created as a design model for C3 research. Section 5 is the experimental design segment and lays a framework for evaluation. Section 6 is the experimental results and discussion of what the results mean. Finally, Section 7 wraps up the study with conclusions and possible future avenues of research.

2 Literature Review

Question Answer systems are a branch of Natural Language Dialog Systems that are more information retrieval oriented with a close parallel to search engine technology. These types of systems are presented with a user-based input and must search their knowledge bases much like a chatterbot in an attempt to return an answer. While the simpler chatterbot technology of directly matching an input to a response is considered to be a toy domain, the complexities involved in the querying and retrieval aspects of Question Answer systems lend this branch its research worthiness.

When discussing Question Answer (QA) systems and how ALICEbots fit into the model, there are several frameworks available. The first of which is the Natural Language Dialog Systems (NLDS) framework [4]. In the NLDS framework, QA systems can be broken into three distinct camps: Semantic theory, Structure theory, and Intention theory.

Semantic theory deals with finding and inferring relationships from system input. An example of such a system is Doug Lenat's CYC project that attempts semantic understanding through complex word relations [5]. Within Semantic theory, meaning is derived from the words used. This approach is more computationally more difficult than other approaches, as it requires a large corpus of word meanings and word sense disambiguation in order to be effective.

Structure theory is instead more concerned with syntactic sentence markers or cue words, and analyzes their communicative function to the system input [6]. Syntactic systems are generally more shallow than semantic ones and require less processing overhead [7]. In QA systems, the syntactic relations of objects generally deal with interrogatives such as who, what, and where. Focusing in on the use of interrogatives can help the syntactic system narrow the scope of possible response choices and provide a more accurate reply. Many of the current

systems handle simple fact-driven questioning but fail at tasks requiring any level of reasoning [8]. An example of such a syntactic system is ALICE (Artificial Intelligence Linguistic Chat Entity). ALICE's syntactic nature will be described in later sections.

In Intention theory, system input is used to communicate plans and beliefs. This is one of the most difficult areas of research. Through this theory, systems possess reasoning abilities and are able to work through problems of a very constrained nature. An example of such a system is TRAINS 95, where the system's purpose was to route locomotives to meet economic demands and to minimize disruptions due to track outages and other simulated disasters [9]. This system is constrained by the rules given to it, and is unable to abstract itself to meet the challenges of a changing problem that violates the rules.

The other framework concerning QA systems is the class system developed by [8]. In this framework, systems are classified according to how they function.

- Class 1 – QA systems that can process fact-based questioning
- Class 2 – QA systems that can use simple reasoning
- Class 3 – QA systems that can fuse together answers from different sources
- Class 4 – QA systems that remember and use previous dialog to form answers
- Class 5 – QA systems that can perform analogical reasoning

To give the reader some background in the area of QA systems, a special Question Answer track was started in 1999 from the Text REtrieval Conference (TREC-8), in which systems were given a standardized test of information retrieval tasks [10]. Systems are given various class questions, from class 1 questions like, "Who is Bin Laden", to class 5 questions like, "Why do terrorists hate the West?" [8].

2.1 ALICE

ALICE is a derivative of structure theory [4] and class 1 QA systems [8]. This fact-driven syntactic parser uses Case-Based Reasoning (CBR) to determine its responses. The knowledge used by ALICE is contained in XML-based Artificial Intelligence Markup Language (AIML) files that can easily be extended to meet particular knowledge demands. The process of adding new knowledge to the system allows ALICE to convey an 'expert appearance' in narrow domains of knowledge [3]. Because of ALICE's dependence on syntactic parsing [11], this lack of cognitive theory will result in ALICE missing certain types of user interactions. This has become the case in several studies done using ALICE chatterbots where ALICE typically scores an 80% accuracy rate in general conversation, and slightly higher when coupled with specific knowledge domains.

What has become interesting is the use of ALICE as a tutor in several research studies. In a University of Huddersfield experiment, ALICE was carefully given the personalities of four famous social psychologists. The intent of the project was to use ALICE as an interactive tutoring tool where students could pose questions and gain further insight into the thinking style of these four individuals [12]. The experiment unfortunately failed when students used ALICE as a search engine to obtain assignment answers and not as a conversational tutor. Another research study conducted in China used ALICE as a conversational partner to teach Chinese students either English or German [13]. This study had surprising results that focused attention directly on ALICE's 80% accuracy rating. Participating students developed negative attitudes toward the system and generally chatted for no more than 10 turns before leaving. This study also reviewed a log of chat dialog and discovered that negative comments about the system outweighed the positive ones. It is thought that these results are in part due to the limited amount

of conversational knowledge used by the system, which was roughly half of the AIML knowledge available at the time.

In a study conducted at the University of Arizona, several ALICE chatterbots were examined to measure the effects of using subjects as unfiltered tutors to build new chatterbot knowledge patterns [14]. Along with the main thrust of the research, measures for response appropriateness and satisfaction levels for individual responses were performed. This study incorporated three similar chatterbots that differed only their knowledge bases. One system used a limited conversational knowledge base; the Standard AIML knowledge set. The second one used a limited conversational knowledge set from Standard AIML, coupled with several hundred telecommunications definitions. The third chatterbot was a combination of the other two. From this experiment, it was found that subjects did not correct the combination chatterbot as often as the stand-alone systems. However, the conversation-only chatterbot was found to exhibit a significantly higher user satisfaction level. It was found that this result was due to the rejection of the telecommunications knowledge set because it contained an inadequate amount of knowledge base entries.

2.2 Emergency Response Systems and the C3 Model

There are several studies on public emergency response and communications systems worth looking at. The first study, called Citizen Awareness System for Crisis Mitigation (CAMAS), is an event extraction tool that utilizes 'humans as sensors' [15]. This tool is deployed at a disaster site and collects first-hand information from victims as they exit the scene. This information could be about potential problems witnessed by the victims such as downed power lines, ruptured gas lines, etc., or could be information concerning where a group of victims may be located. This information is concentrated and analyzed by CAMAS and

appropriate rescue personnel are then notified. The disadvantage of CAMAS is that it does not contain a mechanism for bidirectional information flow. CAMAS is unable to alert victims of impending danger, instead it requires first responders to relay such information.

The second study, called Integrated Community CAre (INCA), is more tailored to the healthcare field [16]. INCA is a network of health-monitoring agents that can oversee the well-being of a community by coordinating emergency services when a problem is detected. The disadvantages of INCA are similar to CAMAS. There is no bidirectional flow of information from the emergency personnel to the affected victims. The unidirectional design of both of these systems preclude them from attaining the goals set forth for C3 systems.

The C3 Model has several key points that we will address.

- "In a crisis, channels to provide information to the public will be clearly needed." [17]
- A C3 system must be tailored to help people in specific areas with specific needs.
- It must be effective in providing information to a targeted population with substantial infrastructure damage.
- It should be able to communicate the status of affected individuals to people outside of the disaster area.

The first point, channels to provide information to the public, may be one of the more obvious ones. Channels are needed to provide the public with information to mitigate further damage following a terrorist event. This information flow helps to calm the public by fixing their concentration on immediate concerns and empowers individuals to commit themselves in useful and needed ways. It also establishes a medium of trust that officials are in charge of the scene, eliminating some of the information vacuum described earlier [2].

The second point, a C3 system must be tailored to help people in specific areas with specific needs, directly addresses the need for bidirectional information flow. Such a system

would be able to direct targeted groups to safety or provide them with information to satisfy a need. Such a system could also be of benefit to individual rescuers by helping them to locate and free trapped victims.

The third point, provide information to a targeted population with substantial infrastructure damage, illustrates the need for a C3 system to provide its own infrastructure, independent of whatever infrastructure may be remaining. This point could further be refined into issues of portability and durability for the system. In later sections we will discuss aspects of the ALICEbot system which could be placed on mobile communications devices and used in emergency shelters.

The last point, the system should be able to communicate the status of affected individuals to people outside of the disaster area, brings up the concept of "I'm Alive" boards. In the September 11th aftermath, displaced victims and their families fell out of touch and were unable to communicate with one another. Throughout Manhattan, "I'm Alive" boards sprang up where victims could tack their name and current shelter to the board. These victims relied on television news to communicate this information to the victims families. It turned out to be a cumbersome and rarely effective method of communication. However, with so many victims and "I'm Alive" boards, information overload quickly became a problem.

Because of the recent timing of "The 9/11 Commission Report" [18] and "Making the Nation Safer" [17], research has yet to be published regarding a complete instantiation of a C3 system. Several of the existing emergency response systems, such as CAMAS and INCA, only meet some of the criteria set forth. As of publication, no known system currently meets all of the C3 criteria.

2.3 ALICEbots and C3

ALICEbots can provide a unique answer to the problems of C3 systems. ALICEbots have the ability to disseminate information to the public through conversational dialog and are able to carry on multiple independent conversations at the same time. They can quickly be augmented with specific knowledge to help targeted groups and are portable in nature for infrastructure-damaged areas. The bidirectional flow inherent in ALICEbots makes the implementation of "I'm Alive" boards a simple programming exercise.

The knowledge used in the ALICEbot can be acquired in one of two ways; manual effort, or automated means. The first method of manual effort, allows information specific to a disaster to be incorporated in the system. This is usually custom-tailored information such as safe escape routes and victim location. The second method, of automated means, is for general types of terrorism questions. These static answers can be automatically scooped up from reputable sources such as definitional websites or news organizations and incorporated into the knowledge base. The optimum solution would be a mixture of both where the automated information gathering is performed periodically to ensure timely and relevant information.

There are several reputable terrorism definition websites of interest to automated knowledge gathering. The first of which is www.terrorismanswers.com which is authored by the Council on Foreign Relations. This website provides a wealth of information on terrorist groups, individuals, states, and events. The other website is www.11-sept.org which also contains an extensive glossary of terms on groups and weapon descriptions. Both of these websites appear to be non-biased in nature and high-quality resources for ALICEbot to use.

2.4 Literature Review Summary

In reviewing the existing literature, the key facts are as follows:

- A focus on interrogatives is an important aspect of QA systems [7].
- C3 systems need to be able to distribute timely and relevant material.
- ALICEbots can potentially fill the role of a C3 system.

Since most users appear to use ALICEbots as specialized search engines [12], it would make sense to approach ALICEbot input in the same terms. The main benefit to using ALICE over a search engine is the ALICEbot's ability to return information in a conversational context, rather than displaying a list of probable sites and asking that users find the information on their own. Following up on the search engine context, [10] noted that search terms are predominately definitional in nature. This would imply that the best method of acquiring knowledge for the ALICEbot would be to obtain definitional responses that are keyed on interrogative input.

The ability to provide information during a crisis is paramount. This flow of information serves to soothe public fears and calm citizens by providing direct and useful information in an appropriate manner.

Finally, as discussed earlier, ALICEbots already meet many of the challenges required of a C3 system. ALICEbots were built first and foremost for conversation and they can leverage that ability in specialized knowledge domains.

3 Research Questions

Because of the versatility of the ALICEbot system, we have identified several research questions we feel need to be investigated. Based on our prior study on the usefulness and performance of a telecommunications Question Answering system [14], we ask:

- **What is the perceived usefulness of a terrorism Question Answer system?**
- H1: It is expected that users will prefer the system with "Both" conversational and domain knowledge to acquire terrorism-related knowledge.

- **How will a Question Answer system perform with:**
    - **Pure conversational knowledge**
    - **Pure domain knowledge**
    - **Both conversational and domain knowledge**
- H2: The Question Answer system with "Both" conversational and domain knowledge will perform better than either conversation or domain knowledge alone.

From our prior research, we believe that a sufficiently large terrorism knowledge base coupled with the general conversation knowledge will yield higher conversationally acceptable and response satisfaction scores than either alone. This is based on our belief that both elements will work together by helping users to reform their queries or pursue tangential areas of knowledge. Prior research with a more limited knowledge base proved that the "Both" chatterbot could not accomplish this task, however, we believe that a larger terrorism knowledge base will improve cross-cooperation between the two knowledge bases.

For our third research question we would like to investigate the particulars of subject sentence interaction in order to tweak future knowledge base entries on those rules most likely to be executed.

- **What are the most frequent types of questioning found in Question Answering systems?**

- H3: It is expected that interrogative sentences beginning with "Wh*" will compose a large segment of user input.

From the preceding question we will have a better idea of what question types to focus future data gathering activities on and as a consequence provide for better response-oriented accuracy gains.

4 TARA System Design

To answer the research questions posed, we created the Terrorism Activity Resource Application (TARA) system, which is based on a modified version of ALICE Program D that can be freely obtained from www.alicebot.org. Both TARA and ALICE share many of the same components, however, for the sake of clarity, Table 1 shows the differences between the two.

	Chat Interface	Chat Engine	AIML	Logging	Evaluation
Original ALICE	Uses XML to chat with users	Uses off the shelf ALICE ProgramD	Uses the freely available Standard and Wallace set (Dialog AIML)	Logs everything to a monolithic XML Log file	Does not provide for response evaluation
TARA	Uses a customized perl skin to chat and for evaluation purposes	Same chat engine as ALICE Program D	Depends on the chatterbot as to whether it is Dialog AIML or customized Terrorism AIML	Keeps XML logs on a per user basis	Customized perl script that allows users to evaluate and suggest new patterns

Table 1. Differences between ALICE Program D and TARA

From Table 1 the reader will notice that ALICE and TARA have five major components. The Chat Interface was one of the more modified components in TARA. Both ALICE and TARA work by taking the user input, tagging it with XML-style tags, and passing it to the chat engine through a Jetty web server, as illustrated in Figure 1.

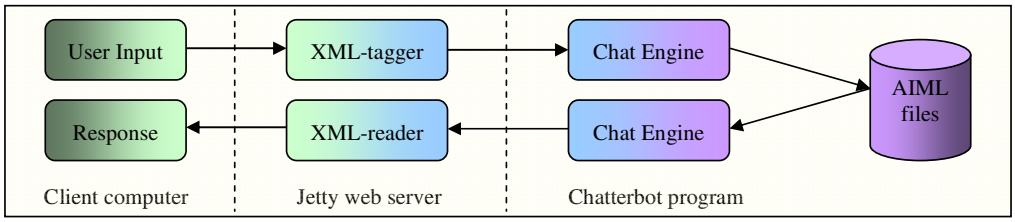

Figure 1. Diagram of User Input to Chatterbot Response

The chat engine is the core element of the system. When the system boots up, the chat engine loads all of the AIML rules or categories into a memory-resident directed graph. As user input is passed to the chat engine from the XML-tagger, the chat engine traverses this graph to find the most suitable match called a pattern. After the match has been made, the corresponding response or template is fed back as XML to Jetty. Jetty then unwraps the XML and posts the chatterbot response to the chat interface.

In the TARA implementation, we modified the chat interface by overlaying a perl skin on top of the client computer. This perl skin allowed for input and response with the Jetty web server as described above, however, it also allowed us to introduce an evaluation component which we will describe later. In both TARA and ALICE Program D, the chat engine and Jetty web server were identical.

The AIML files can be thought of as the brains of the system. The chat engine only organizes and retrieves information stored in AIML, while the AIML itself contains the input and response categories. The following is an example category:

```
<category>
<pattern>WHAT IS POTASSIUM IODIDE</pattern>
<template>FDA-approved nonprescription drug for use as a blocking
   agent to prevent the thyroid gland from absorbing radioactive
   iodine.
</template>
</category>
```

In order for the system to return the template as a response, the user will have to exactly ask "What is potassium iodide". Although the system is not case-sensitive in matching patterns, it is sensitive to wildcard matching. The above example does not use any ALICE-specific wildcards, i.e., '*' or ''. So adding a question mark 'What is potassium iodide?', would not retrieve the same template. This shortcoming becomes more evident when using automated means of gathering terrorism knowledge. The simplest fix is to create multiple patterns that add wildcards both in front of and behind the key term. For example:

```
<category>
<pattern>* PENTAGON *</pattern>
<template>Headquarters of the U.S. Department of Defense. The
   five-sided building, built in 1943, was one of the targets of
   September 11. American Airlines Flight 77 was flown into the
   Pentagon, killing 189 people in all, including 125 people
   inside the building, 64 passengers, and 5 terrorist
   highjackers.
</template>
</category>
```

This category will return the template if the pattern contains the word 'Pentagon' in the middle of the text. This is a more palatable solution, however, this may not be only response desired when talking about the Pentagon. To give the reader an idea of how the ALICE wildcard system works, consider the following four patterns:

- Bin Laden
- Bin Laden *
- * Bin Laden
- * Bin Laden *

In ALICE's scheme of wildcard handling, each of these patterns are distinct which prevents two or more patterns to be matched to the user's input. In the first case, 'Bin Laden' will be activated *only if* the user's input is precisely, 'Bin Laden' with no other characters before

or after the input. In the second case, '* Bin Laden' will be activated *only if* 'Bin Laden' is at the end of a sentence with preceding characters. This leads us to the use of qualifying definitions with the use of interrogatives, such as 'What is the Pentagon' or 'Where is the Pentagon', and generating different templates for those types of questions. The following is a sample interaction to show how qualifying the templates can lead to difficulty.

```
User: Who is our president?
System: George W. Bush

User: What is our president respond to terrorist attack?
System: Any act or series of acts by an enemy causing substantial
   damage or injury to property or persons. In any manner by
   sabotage or by the use of bombs shellfire or atomic
   radiological chemical biological means or others or processes.
```

In this example, the system recognized the term 'terrorist attack', but failed to respond to it in the expected context. It is for instances like these that a larger and more specific corpora of AIML knowledge may be able to address.

The AIML files between TARA and ALICE are slightly different depending upon the chatterbot knowledge desired. For a general conversational chatterbot, it is recommended that the Standard and Wallace AIML knowledge sets be used. These knowledge sets helped ALICE to win the Loebner contest. This pattern of success against deeper reasoning systems is justification enough to use these free knowledge sets. For terrorism knowledge, we put together our own AIML sets which we will describe in the Experimental Design section.

The logging component has also undergone several changes from the ALICE implementation. In ALICE, logging was performed within the chat engine before passing the response back to Jetty. The logs used by ALICE were monolithic XML files that captured every interaction with the program. In our design of TARA, we incorporated logging into the perl skin of the chat interface. This allowed us the freedom to make programming changes without

modifying the chat engine and risking code integrity problems. We logged each participant's conversation into separate XML files which helped limit data overload challenges in the analysis phase.

The evaluation component was new in the TARA design. In TARA, we provided a method of integrated evaluation where the user is given the opportunity to rate their satisfaction level after each chatterbot response is given.

5 Experimental Design

5.1 AIML Knowledge

For our experiment we chose to test three chatterbots, all with different types of AIML knowledge. Table 2 breaks down the AIML knowledge and shows the number of categories for each chatterbot.

	Number of categories		
Chatterbot	Conversational AIML	Terrorism AIML	Total
Dialog	41,873 †		41,873
Domain		10,491	10,491
Both	41,873	10,491	52,354 ‡

Table 2. AIML distribution for the three chatterbots

† - Some of the Conversational AIML categories did include some terrorism terminology.
‡ - Ten categories between Conversational and Terrorism AIML overlapped, decreasing the total number of categories by ten.

The 41,873 categories of conversational AIML consisted of the Standard and Wallace AIML sets. For terrorism knowledge, we created AIML files through a mixture of automated and manual entry processes. Automated entry collected the most categories, 10,476, while 318 categories were obtained from manual methods. This represents an overlap of 303 categories, which occurs when two categories have the exact same pattern. When there is overlap, the chat engine ignores the duplicate categories thus decreasing the total number of categories.

From the 10,476 automatically gathered categories, 2,619 are unique patterns. This factor of four difference is because of the way the ALICE chat engine handles wildcards, as described earlier. These 2,619 patterns came from several terrorism-related websites that were identified as reputable places to gather terrorism definitions by a terrorism expert.

From the 2,619 unique patterns collected:

- 1,322 came from www.terrorismanswers.com
- 174 came from www.11-sept.org
- 1,212 came from www.wmd-nm.org

An overlap of 89 categories occurred between these three sites.

For manual terrorism knowledge, all 318 categories were gathered from www.highvolumemedia.com/thebullhorn/GlossaryA-Z.htm. These categories are represented by the following terrorism groupings:

- Countries – 29 categories
- Terrorist Groups – 122 categories
- WMD – 24 categories
- Definitions – 131 categories
- People – 4 categories
- Bin Laden – 8 categories

Because these terms were manually gathered, appropriate wildcards could be custom added. This avoided the 4x wildcard handling experienced by the automated method. An example category is listed below.

```
<category>
<pattern>What is Zyklon B</pattern>
<template>A form of hydrogen cyanide.  Symptoms include increased
   respiratory rate, restlessness, headache, and giddiness
   followed later by convulsions, vomiting, respiratory failure
   and unconsciousness.  Used in the Nazi gas chambers in WWII.
</template>
</category>
```

5.2 Study Participants

Participants were recruited from several undergraduate and graduate Management of Information Systems classes. Each participant was assigned to one of the three chatterbots through a pseudo-random algorithmic process based on their University login ID's. Participants were encouraged to interact with the system for approximately one-half hour before completing a final survey. Students were asked to discuss terrorism-related topics, but the system did not force them to do so. Completion of the study was further encouraged through the prospect of randomly awarded gift certificates.

An integrated evaluation method was used in our study. Participants would chat a line, and immediately evaluate their satisfaction level of the chatterbot response. Figure 2 shows a sample screenshot of the evaluation process.

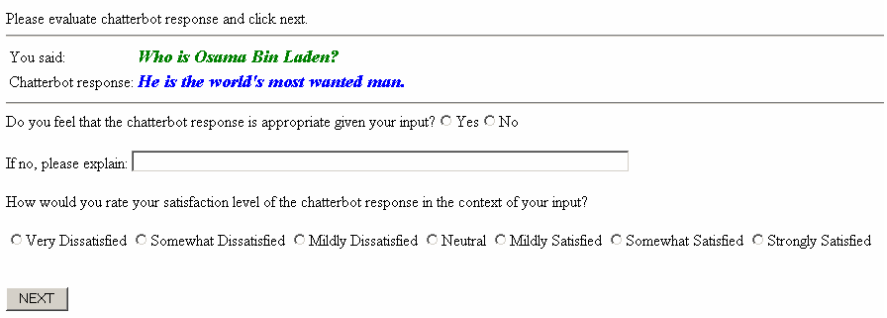

Figure 2. Screenshot of TARA's evaluation process

The final survey was composed of seven questions, four open-ended, and three dichotomous. The four open-ended questions were a qualitative look at user impression of system strengths and weaknesses.

- What do you believe are the strengths of the system?
- What do you believe are the weaknesses of the system?
- What do you believe the system excels at?
- What do you believe the system needs help on?

The three dichotomous questions quantified user reaction concerning their comfort level with the system, whether or not they would use such a system to find terrorism information, and whether or not they would recommend the system to friend. Each of these three questions also had space for open-ended comment where users could optionally justify their answers.

- Do you feel comfortable using this system?
- Would you use such a system to find terrorism information?
- Would you recommend this system to a friend who was interested in terrorism knowledge?

A pilot study was also conducted to trial run the system and to evaluate that the responses received were indeed what we had intended to capture.

5.3 Metrics

To answer the research questions posed, we set up three chatterbots, to use as a dependent variable, that differed only on the knowledge based used: general conversation, terrorism knowledge, and a combined system. The following independent variable metrics were designed: User Impression, Response Rating, and Response Appropriateness. User impression is actually three subtle measures that tie directly to the three dichotomous questions asked in the final survey. These measures of user comfort, usability, and recommendation potential are measured as an aggregate percentage for each chatterbot. The open-ended comments

accompanying these questions are used to qualitatively determine causes behind the user's choices.

Response rating is introduced in the evaluation component and is a measure of the user's satisfaction level of the chatterbot response. This measure uses a Likert scale and is an aggregate computation of all response ratings divided by the number of user inputs.

Response appropriateness also comes from the evaluation component. It is a dichotomous measure that asks participants whether or not the response is appropriate given the user input. This is different from response rating, because an interaction may produce an unexpected response leading to a low response rating yet be an appropriate response. This measure is aimed more towards determining whether a participant will accept the response or not.

6 Experimental Results

6.1 Testing H1

In our first hypothesis, we postulated that users will prefer the chatterbot with Both dialog and domain knowledge. Using the qualitative questions in the final survey, the results of Table 3 are generated.

User's 'Final Survey' Analysis	Chatterbot		
	Dialog	Domain	Both
Number of Users completing the survey	27	25	29
User Comfort	22.2%	44.0%	31.0%
Usability	11.1%	28.0%	31.0%
Recommendation Potential	22.2%	28.0%	24.1%

Table 3. Analyzing User Behavior towards the system

In looking at the percentages for the 'Both' chatterbot, it was preferred the most in the category of usability (31.0%). For the other two categories the Domain chatterbot was preferred, 44.0% and 28.0% respectively. This counters our intuition of the hypothesis and was further investigated.

In looking at the first category of user comfort, we analyzed the open-ended comment of the Domain and 'Both' chatterbots. It soon became clear that some users had selected other criteria to judge the system. Table 4 shows the Domain and 'Both' chatterbot comments broken down by common groupings.

User comment groupings for 'User comfort'				
	Domain chatterbot		'Both' chatterbot	
	Comfort:Yes	Comfort:No	Comfort:Yes	Comfort:No
Ease of use	3	0	0	0
Responses	7	14	8	20
Other	1	0	1	0

Table 4. User comments about comfort using the system

To demonstrate how we arrived at our "Ease of Use" categorization numbers in Table 4, we list the open-ended comments received: "easy to use", "it is easily deployed via the web…", and again "easy to use". The "Other" category was generally reserved for meaningless comments such as "N" which we received and thought was rather ambiguous. Select comments (misspelling et al.) regarding "Responses" are as follows: "It gives very specific details in the responses with exact dates and places", "It recognizes key words well and gives good replies", and "It is like dictionary search engine. So storng about definition".

We had expected that users would rate the system solely based on the chatterbot responses given. However, a minority of participants in the Domain chatterbot instead rated it favorably based upon the system's ease of use. If we were to look at only those values that dealt with Responses, the Domain chatterbot would still be preferred 33.3% to 'Both' at 28.6%. In looking further into the comments, two key findings emerged:

- Users were uncomfortable with the system returning definitions instead of conversation
- Users seemed to like the breadth of knowledge and speed of the system

Several users made comments that the system felt more natural when it is not talking about terrorism and they were displeased with its definition-style approach rather than placing knowledge in the context of the conversation. Others were impressed with the level of detailed knowledge they received from their queries.

The other non-cooperative category of recommendation potential was very similar to what we found in user comfort. Table 5 breaks down user comments between the Domain and 'Both' chatterbot. Again a minority influence of system ease crept into the results.

User comment groupings for 'Recommendation potential'				
	Domain chatterbot		'Both' chatterbot	
	Recmnd:Yes	Recmnd:No	Recmnd:Yes	Recmnd:No
Ease of use	2	1	0	0
Responses	4	17	6	22
Other	1	0	1	0

Table 5. User comments about system recommendation

A minority of users again cited ease of system use instead of the expected category of Responses. However, if we were to only look at the value for Responses, this time the 'Both' chatterbot leads in preference 21.4% to Domain's 19.0%, as shown in Table 6.

Modified 'Final Survey' - Responses Only	Chatterbot		
	Dialog	Domain	Both
Number of Users completing the survey	27	21	28
User Comfort	22.2%	33.3%	28.6%
Usability	11.1%	19.0%	28.6%
Recommendation Potential	22.2%	19.0%	21.4%

Table 6. User behavior towards the system – looking at comments on Responses only

Table 6 shows a snapshot view of the final survey questions where only comments about the system responses were made. Even from this look, the 'Both' chatterbot was clearly not preferred as we had expected. Through analyzing the open-ended comments further, several key findings emerged:

- The system needs a larger corpus of knowledge
- Users appreciated the dialog (conversational) component of the system

The first finding directly contradicts one of the findings from user comfort. However, with approximately an equal number of users making comments both for and against the amount of terrorism-specific knowledge, it would appear that a sufficient level of knowledge was used in this experiment. The second finding was more obvious as users pointed out that the dialog helped them to reform their queries in order to obtain their desired response.

From analyzing the three final survey questions, we found that users did not prefer the 'Both' chatterbot. In particular, it appeared from the comments that users disliked the domain's definitional treatment of responses and would have preferred responses in a conversational context. This definitional treatment did emerge as a hindrance to the 'Both' chatterbot.

6.2 Testing H2

For the second hypothesis, we believe that the 'Both' chatterbot will perform better than stand-alone Dialog and Domain. To test our assumption we use the results of response appropriateness and response rating. Tables 7 and 8 show the results of response appropriateness and response rating respectively.

Response Appropriateness	Chatterbot		
	Dialog	Domain	Both
Response Appropriateness	66.3%	21.6%	62.0%

Table 7. Response appropriateness across three chatterbots

Chatterbot	Number of Response Ratings	Avg Response Rating	Std Dev
Dialog	1,524	4.04	2.00
Domain	849	2.43	1.90
Both	1,138	4.21	2.20

Table 8. Response rating statistics

In analyzing the response rating from Table 8, the 'Both' chatterbot was preferred over Dialog and Domain (p-value < 0.001). But, in Table 7 the 'Both' chatterbot was second-best to Dialog (62.0% to 66.3%). The reader should be cautioned that the above numbers are a bit misleading, because the 'Both' chatterbot is composed of elements from both dialog and domain. Table 9 places everything on equal footing by breaking the 'Both' chatterbot responses into its constituent parts for a more comparable analysis.

	Both's components	
Breaking apart the numbers of 'Both'	dialog	domain
Number of user queries	888	250
Avg Response Appropriateness	68.4%	39.6%
Avg Response Rating	4.51	3.14
Std Dev of Response Rating	2.12	2.17

Table 9. The 'Both' chatterbot's dialog and domain responses

From Table 9, we can now compare the dialog component of 'Both' against the Dialog chatterbot which has a response appropriateness of 68.4% compared to 66.3% (from Table 7, p-value < 0.001). Similarly, we compare the domain component of 'Both' against the Domain chatterbot and obtain a response appropriateness of 39.6% and 21.6% respectively (p-value < 0.001). From this analysis we see that the dialog and domain components of 'Both' are performing better than the Dialog and Domain chatterbots.

6.3 Testing H3

In hypothesis H3, we believe that the interrogatives beginning with 'wh*' will be most frequently observed. To test this hypothesis we look at the 1,138 user inputs given to the 'Both' chatterbot. From this number, 250 or 22.0% were terrorism-related. Of the 250, 171 or 68.4% were interrogatives. Eighty-eight of 171 interrogatives began with the 'wh*' interrogative (51.5%). Table 10 lists the most frequently encountered interrogatives.

Sentence Starter	# Instances	% of Total
What	47	27.5%
Do	27	15.8%
Who	19	11.1%
How	14	8.2%
Where	10	5.8%
Is	9	5.3%

Table 10. The most popular interrogatives used

It was surprising to note the high percentage of 'Do', 'How', and 'Is' instances. These words were not entirely expected in our study. The other fact worth mentioning was that 'Who' questions fetched the highest response appropriateness and rating scores. It would appear that 'Who' questions were well represented by the AIML knowledge.

Looking at the dialog side of 'Both' produces a similar picture. There were 245 'wh*' interrogatives posed in the Both-dialog, and of those 136 began with 'what' (55.5%). In looking at those 136 instances of 'what', 57 contained a terrorism keyword but failed to return a terrorism-related response. Looking further into the 57 terrorism classification failures, the following informal observations were made:

- 31 of the failures were due to an insufficient knowledge base
- 10 were from overly complicated knowledge base entries
- 5 were from improper knowledge base formatting
- 1 was from a better match found in the Dialog AIML

Addressing these failure types regarding knowledge base entries as well as using the interrogative frequency counts from Table 10 will help refine future systems. Efforts should be made to concentrate similar knowledge gathering activities on knowledge base rules that begin with the 'what' interrogative. In summary, we found that 'wh*' interrogatives made up a majority of user questioning and furthermore, the word 'what' was the most frequent interrogative sentence starter.

7 Conclusions and Future Directions

In conclusion, several key findings were made. The first finding was that users appear to prefer a natural flow of conversation over a definitional approach. This means that the terrorism knowledge bases need to be adjusted to reflect a more conversational tone. Perhaps in the future a more careful screening of domain-specific terms and a post-acquisition filter can format the knowledge to fit the demands. The second finding was that the components of the 'Both' chatterbot performed better together than apart. This finding is supported eloquently through the comments of one user that said the dialog kept the user "on track" throughout the conversation. Whenever the user entered a query the system did not understand, the dialog part helped the participant to reform their question to obtain a desired answer. The third finding is consistent with the view of Voorhees that interrogatives are a major source of user inquiries. The 'wh*' interrogatives and 'what' in particular described a substantial number of user queries [10], and should be the focus of future knowledge gathering activities.

We suggest some future attention to be given to spell-checking and a larger knowledge base source. The first suggestion is to implement a spell-check mechanism on user input. Although it may have made a small improvement in our study, the trivial nature of programming would be well worth the effort. The other area is to investigate is the use of a larger corpus of knowledge. In our study, the amount of knowledge appeared to be appropriate, however, the more knowledge a system has, the more potential accuracy of correctly answering highly specific questions the system could have.

From this study we have shown that off-the-shelf ALICEbots can function adequately in the terrorism domain by providing a channel of relevant information to the public. To meet the other challenges of a C3 system, future efforts should also investigate the use of ALICEbots to

disseminate specific information to targeted individuals, as well as its ability to harness other data sources such as the "I'm Alive" boards, to communicate the status of victims to family members outside of a disaster area. We believe that these efforts to establish a working C3 system will be well worth it.

References

1. Furedi, F., *Heroes of the Hour*, in *NewScientist*. 2004.

2. Durodié, B. and S. Wessely, *Resilience or panic? The public and terrorist attack*, in *The Lancet*. 2002. p. 1901-1902.

3. Wallace, R.S., *The Anatomy of A.L.I.C.E.*, in *A.L.I.C.E. Artificial Intelligence Foundation, Inc.* 2004.

4. McKevitt, P., D. Partridge, and Y. Wilks, *Why machines should analyse intention in natural language dialogue.* International Journal of Human-Computer Studies, 1999. 51(5): p. 947-989.

5. Lenat, D., G. Miller, and T. Yokoi, *CYC, WordNet, and EDR: critiques and responses.* Communications of the ACM, 1995. 38(11): p. 45-48.

6. Andernach, T. *A machine learning approach to the classification of dialogue utterances*. in *Proceedings of the Second International Conference on New Methods in Language Processing, NeMLaP*. 1996: Machine Learning.

7. Hammerton, J., et al., *Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing.* Journal of Machine Learning Research, 2002. 2: p. 551-558.

8. Moldovan, D., et al., *Performance issues and error analysis in an open-domain question answering system.* ACM Transactions on Information Systems, 2003. 21(2): p. 133-154.

9. Ferguson, G., et al. *The Design and Implementation of the TRAINS-96 System: A Prototype Mixed-Initiative Planning Assistant*. in *Proceedings of the Third Conference on Artificial Intelligence Planning Systems*. 1996.

10. Voorhees, E.M. *Overview of the TREC 2001 Question Answering Track*. in *Text REtrieval Conference*. 2001.

11. Russell, R.S., *Language Use, Personality and True Conversational Interfaces*. Project Report, AI and CS. 2002, Edinburgh: Univ of Edinburgh.

12. Moore, R. and G. Gibbs, *Emile: Using a chatbot conversation to enhance the learning of social theory*. 2002, Univ. of Huddersfield: Huddersfield, England.

13. Jia, J., *The Study of the Application of a Keywords-based Chatbot System on the Teaching of Foreign Languages*. 2002, University of Augsburg: Augsburg, Germany.

14. Schumaker, R.P., et al., *An Evaluation of the Chat and Knowledge Delivery Components of a Low-Level Dialog System: The AZ-ALICE Experiment.* Decision Support Systems, 2005. Forthcoming.

15. Mehrotra, S., et al. *CAMAS: a citizen awareness system for crisis mitigation*. in *International Conference on Management of Data, Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. 2004. Paris, France.

16. Beer, M.D., R. Hill, and A. SixSmith. *Deploying an agent-based architecture for the management of community care*. in *International Conference on Autonomous Agents, Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. 2003. Melbourne, Australia.

17. National Research Council, *Making the Nation Safer*. 2002, Washington D.C.: The National Academies Press.

18. National Commission on Terrorist Attacks upon the United States, *The 9-11 Commission Report*. 2004.

Robert P. Schumaker is a third-year PhD student in the Department of MIS at The University of Arizona. He received his undergraduate degree in Civil Engineering from the University of Cincinnati and an MBA degree in Management and International Business from the University of Akron. His interests include Stock Price Prediction, Natural Language systems and Textual Analysis techniques.

Dr. Hsinchun Chen is McClelland Professor of Management Information Systems at the University of Arizona and Andersen Consulting Professor of the Year (1999). He received the B.S. degree from the National Chiao-Tung University in Taiwan, the MBA from the SUNY Buffalo, and the Ph.D. degree in Information Systems from New York University. He is author of seven books and more than 120 SCI journal articles covering intelligence analysis, data/text/web mining, digital library, knowledge management, medical informatics, and Web computing.