

Evaluating the Efficacy of a Terrorism Question Answer System: The TARA Project

Rob Schumaker, Ying Liu, Mark Ginsburg, and Hsinchun Chen

Artificial Intelligence Lab, Department of Management Information Systems
The University of Arizona, Tucson, Arizona 85721, USA
{[rschumak](mailto:rschumak@eller.arizona.edu), [yingliu](mailto:yingliu@eller.arizona.edu), [mginsbur](mailto:mginsbur@eller.arizona.edu), [hchen](mailto:hchen@eller.arizona.edu)@eller.arizona.edu}

Word Count: 2874

Introduction

Terrorism education has become a topic of interest lately. With many agencies and private organizations scrambling to provide information, the actual process of finding relevant material can sometimes become lost in the chaos. To the credit of the Department of Homeland Security and several private organizations, there is some valuable information available; however, it is mainly geared towards first responders and not the general public.

In both the “9-11 Commission Report” and “Making the Nation Safer,” the authors propose to bridge this gap through the use of C3; systems embodying Command, Control, and Communications elements. These systems would allow for the deployment of communications channels during an emergency to support decision management as well as communicate instructions to the public during an emergency (Moore and Gibbs 2002).

One potential approach to C3 is through the use of ALICEbots. ALICEbots (Artificial Linguistic Internet Chat Entity robots) are a type of Question Answer (QA) chatterbot developed in 1995 by Richard Wallace. The advantage of these chatterbots is in their ability to be quickly programmed with terrorism-specific knowledge, as well as their robust and scalable nature. ALICEbots are built first and foremost for conversation and are a promising vehicle in disseminating terrorism-related information to the public.

ALICEbots and Question Answer Systems

ALICEbots work by matching user input against pre-existing XML-based input patterns and returning the template response. This simple method of conversational mimicking eliminates the computational overhead that would normally be associated with deeper reasoning systems. The technique can also permit expansion into new knowledge domains, allowing the ALICEbot to convey an ‘expert appearance’ (Wallace 2004). An example knowledge base entry is listed below:

```
<category>
<pattern>WHAT IS AL QAIDA </pattern>
<template>'The Base.' An international terrorist group founded in
approximately 1989 and dedicated to opposing non-Islamic
governments with force and violence.
</template>
</category>
```

In studying the ALICEbot system, it was found that because of its architecture, ALICE cannot properly answer all of the queries given to it. This is the inevitable outcome of using a shallow method of conversational parsing. As simplistic as it may be, this system has proven itself against deeper algorithmic systems by winning the Loebner Prize in 2000, 2001, and 2004 for most human-like conversational partner.

A. Question Answer Systems

Question Answer systems fall into two different categories, domain dependent, and domain independent systems. Systems that are domain independent do not try to understand text. Instead these systems focus on retrieving a snippet of text from the source (Voorhees and Tice 2000). An example system is MURAX which uses external encyclopedic information to answer queries.

A-1. Domain Dependent Systems

Domain dependent systems are those that rely on a specially crafted knowledge-base, and are composed of two subcategories: the traditional or narrow domain, and open domain

(Voorhees and Tice 2000; Pasca and Harabagiu 2001). In the traditional domain, systems attempt conversational fluency in limited domains of expertise. Example systems include Winograd's SHRDLU which could answer natural language queries about a fictitious Block World (Winograd 1977), STUDENT which could solve algebraic word problems (Winograd 1977), and LUNAR which could answer queries on lunar rock data (Woods 1977). Open domain systems are those that still rely on internal knowledge bases, but have a more diverse repertoire of topics. These systems can be further classified into two major camps, information retrieval and information extraction (Voorhees and Tice 2000; Pasca and Harabagiu 2001).

A-2. Information Retrieval versus Information Extraction

In Information Extraction, the goal is to extract the who did what to whom and where type of information from text and fill in pre-defined templates with the extracted data. This field is serviced by the Message Understanding Conferences or MUCs.

In Information Retrieval (IR), there are two basic classes. The document-based and sentence-based retrieval systems (Vrajitoru 2003). The goal of document-based systems is to return a set of relevant documents to the user; which is one of the tenets of the Text Retrieval Conferences (TREC) (Pasca and Harabagiu 2001; Potter 2001). Sentence-based retrieval systems return a small snippet of text to the user, similar to the domain independent systems, except that sentence-based IR uses internal knowledge bases. Crossover systems that utilize both document-based and sentence-based approaches do exist, but are mainly confined to the arena of search engine design (Radiv, Fan et al. 2005). ALICEbots fit into the sentence-based category (Vrajitoru 2003) primarily because of their sentence-directed response capability and use of internal knowledge bases.

B. Question Answer Systems for Emergency Response

Looking at studies done on public emergency communications systems and how they fit into the C3 model, two studies appear most relevant; CAMAS and INCA. CAMAS, Citizen Awareness System for Crisis Mitigation, utilizes a ‘humans as sensors’ approach where event information is extracted from victims and first responders in a disaster area, analyzed, and the appropriate response personnel are then dispatched to priority areas (Light and Maybury 2002). The downside is that CAMAS does not handle bidirectional information flow back to the public. The other system, INCA, Integrated Community Care, is a network of healthcare agents that monitor patients and coordinate emergency services when needed (Beer, Hill et al. 2003). Like CAMAS, INCA also does not handle bidirectional information flow.

There are several aspects to the C3 model that must be maintained. Emergency communications capability should include channels from emergency personnel to the general public, channels from emergency personnel to specific groups or individuals, and channels of communication from within the disaster area to those affected outside of it. Applying QA systems to this problem addresses many of the information dissemination needs. In particular, the use of ALICEbots can help fill the role of information dissemination and can deal with answering many simultaneous disaster-related inquiries in a tireless and personable manner. Specific disaster-related knowledge can be quickly scooped up by scanning news sites or entered manually to meet precise needs. Since most search queries focus on the use of interrogatives and definitional types of answers (Voorhees 2001), ALICEbots can capitalize on this fact by focusing attention on reputable definitional sources (such as www.terrorismanswers.com and www.11-sept.org).

Terrorism Activity Resource Application (TARA)

In our research, we aimed to examine the efficacy of shallow QA systems for disseminating terrorism-related information to the general public. We created three modified ALICEbots, which differed from each other on the dimension of terrorism knowledge bases used. One chatterbot used only general conversational knowledge, the second used only terrorism domain knowledge, and the third was a combination of both conversation and terrorism knowledge. This leads us to the following research questions:

- How will users perceive the usefulness of the different chatterbots?
- How will the different chatterbots perform?
- What are the most frequent types of interrogatives used in the study?

To answer these questions, we created TARA (Terrorism Activity Resource Application). The TARA system design was based on a modified version of the ALICE Program D chatterbot engine which is freely available at www.ALICEbot.org.

There are some notable differences between TARA and ALICE, as illustrated in Figure 1. The only component that remained unaltered was the actual Chat Engine.

	Chat UI	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)	Logs everything to a monolithic XML Log file	None
TARA	Uses a customized perl skin to chat and for evaluation purposes	Same as Original ALICE	Depends on the bot as to whether it is Dialog or customized Terrorism knowledge	Keeps XML logs on a per user basis	Customized perl script that allows users to evaluate and suggest new patterns

Figure 1. Differences between Original ALICE Program D and the TARA chatterbot

As mentioned earlier, we used three chatterbots with differing knowledge bases. The control chatterbot “Dialog,” the general conversationalist, was loaded with the Standard and Wallace knowledge set that allowed ALICE to win the early Loebner contests. This set consists of 41,873 knowledge base entries. The second chatterbot “Domain,” was loaded with 10,491 terrorism-related entries. The third chatterbot “Both,” was a summation of “Dialog” and “Domain,” which can carry on a general conversation and easily handle terrorism inquiries as well. It contains 52,354 entries, 10 less than a true summation because of an overlap between the dialog and domain knowledge bases.

Terrorism entries were collected through a mixture of automatic and manual means. The majority were gathered automatically from several reputable websites including www.terrorismsanswers.com and www.11-sept.org. Manual entry was used sparingly to augment the terrorism knowledge set.

For our research we used ninety participants, thirty for each chatterbot, whom were a mixture of undergraduate and graduate students taking various Management of Information Systems classes. Participants were randomly assigned to one of the chatterbots, and were asked to interact with the system for approximately one-half hour and were permitted to talk about any terrorism-related topic. They were further given an incentive to perform through the random awarding of gift cards to popular local businesses.

The evaluation method of chatterbot responses was an integrated process where users would chat a line and then immediately evaluate the chatterbot’s response. Users were asked to evaluate each line with the following two measures; appropriateness of response (Yes/No), and satisfaction level of the response using a Likert scale of values (1-7). Users were also given the opportunity to provide open-ended comments on a line by line basis. Figure 2 shows the evaluation interface.

Please evaluate chatterbot response and click next.

You said: ***Who is Osama Bin Laden?***
 Chatterbot response: ***He is the world's most wanted man.***

Do you feel that the chatterbot response is appropriate given your input? Yes No

If no, please explain:

How would you rate your satisfaction level of the chatterbot response in the context of your input?

Very Dissatisfied Somewhat Dissatisfied Mildly Dissatisfied Neutral Mildly Satisfied Somewhat Satisfied Strongly Satisfied

Figure 2. The evaluation interface

Following interaction, participants were given an end of the study survey which was aimed at extracting user impressions about the system and its potential impact. The survey asked the following questions:

- Do you feel comfortable using this system to find terrorism information?
- Would you use it to find terrorism information?
- Would you recommend it to a friend who wanted to find terrorism information?

These questions are asked in Yes/No format with space for open-ended comment after each question.

Testing TARA: Experimental Results

A. Users appeared to prefer a natural flow of conversation instead of a definitional approach to knowledge dissemination.

For the first experimental design question, *how will users perceive the usefulness of the different chatterbots*, we studied the results from the end of survey questions that dealt specifically with chatterbot responses. We had expected that the chatterbot with “Both” conversation and terrorism knowledge bases would rate the highest in all three categories; user comfort level, system usability, and recommendation potential. As it turned out, the “Both” chatterbot failed to rate highest in the ‘user comfort’ category (31.0%), beaten unexpectedly by

the terrorism-only “Domain” chatterbot (33.3%) at a p-value < 0.05. Table 1 summarizes the results.

User's 'End Survey' Analysis			
	Dialog	Domain	Both
Number of Users completing the survey	27	21	29
User feels comfortable using the system	22.2%	33.3%	31.0%
User would use the system if available	11.1%	19.0%	31.0%
User would recommend this system to others	22.2%	19.0%	24.1%

Table 1. End Survey analysis of the three chatterbots

From the examination of user comments, we quickly discovered the cause of this discrepancy. The “Domain” chatterbot returned only terrorism-related definitions to its participants. Therefore, participants of the “Domain” chatterbot came to accept these definitions as normal. For users of the “Both” chatterbot, the dialog knowledge set would return responses in conversational form and terrorism domain responses in definitional form which participants found to be in conflict. It was further found that participants of the “Both” chatterbot were uncomfortable with the idea of having the domain-specific knowledge be returned to them in definitional style. Users mentioned that they would have preferred the responses in a conversational context. One user in particular mentioned the usefulness of the system’s conversational aspects. “Some comments are very appropriate and most of them make sense even when they’re off topic.”

B. The ‘Both’ chatterbot performed better than ‘Dialog’ and ‘Domain’.

For the second question, *how will the different chatterbots perform*, measurements were conducted on the appropriateness and satisfaction rating of the chatterbot responses. Because the “Both” chatterbot is composed of dialog and domain parts, we took the “Both” chatterbot and broke its responses into its constituent parts of dialog and domain. We then compared those results against the actual Dialog and Domain chatterbots. This comparison is shown in Table 2.

Breaking apart the numbers (Both)	Both's components		Actual chatterbots	
	dialog	domain	Dialog	Domain
Number of Lines Entered in the Bot	888	250	1,524	849
Average Response Appropriateness	68.4%	39.6%	66.3%	21.6%
Average Response Satisfaction Rating	4.51	3.14	4.04	2.43
Standard Deviation of Response Satisfaction	2.12	2.17	2.00	1.90

Table 2. Comparing the components of “Both” against the Dialog and Domain chatterbots

When comparing the dialog component of “Both” against the actual Dialog chatterbot, the “Both” component rated higher in response appropriateness, 68.4% to 66.3%, as per our expectation. When looking at the domain component of “Both” against the Domain chatterbot, again the “Both” component rated higher, 39.6% compared to 21.6%. Likewise, Response Satisfaction scores from the “Both” chatterbot rate higher than the corresponding “Actual” chatterbots. This analysis shows that the “Both” chatterbot performed better in its constituent areas compared against the stand-alone chatterbots. We believe that this is the result of the dialog portion responding to unrecognized queries and steering communication back to terrorism topics.

C. The ‘wh*’ interrogatives appear to be a good place to focus future knowledge acquisition activities.

For the third question, *what are the most frequent types of questioning used in the study*, we investigated the input/response pairs of the “Both” chatterbot. In particular we were interested in only those user inputs which were in the form of a question, (68.4% of the terrorism domain inputs were interrogatives). Table 3 summarizes the most frequently observed interrogatives.

Interrogative	Percentage Use
What	27.5%
Do	15.8%
Who	11.1%
How	8.2%
Where	5.8%
Is	5.3%

Table 3. Results of the most frequently observed interrogatives

Investigating the interrogatives further, it was found that the interrogative “what” started the most user queries at 27.5% of all queries. We had expected that interrogatives beginning with “wh*” would be the most prevalent, and indeed they were, making up 51.5% of all interrogatives. It is interesting to note how often “Do” and “Is” were used, as these were unexpected surprises. In the vein of work done by Moore and Gibbs (Moore and Gibbs 2002) where students used the chatterbot as a search engine, focusing future efforts of knowledge collection at these selected interrogatives should best improve chatterbot accuracy.

Conclusions and Discussion

In conclusion, it was found that users appeared to prefer a natural flow of conversation instead of a definitional approach to knowledge dissemination. This will mean that terrorism-specific knowledge bases will need to be adjusted, as well as perform a more careful screening of terrorism sources incorporated into the knowledge bases. It was also found that the “Both” chatterbot with dialog and domain knowledge bases performed better than its stand-alone counterparts. This would appear to be the result of dialog and domain working together to provide more appropriate results together rather than apart. Lastly, consistent with Voorhees, interrogatives are a major source of user inquiries. The “wh*” interrogatives, and “what” especially, appear to be a good place to focus future knowledge acquisition activities.

Since most users appear to use ALICEbots as specialized search engines (Moore and Gibbs 2002), it would make sense to approach ALICEbot input in the same terms. Following up on the search engine context, Voorhees (Voorhees 2001) 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.

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.

In the future, it would be a good idea to implement a spell-checking component on user inquiries. This would eliminate 6.6% of the observed bad responses due to horribly misspelled domain terms. It would also be good to investigate adding more knowledge to the system. Although our domain-specific knowledge base appeared to be sufficient for the task, it would be interesting to test even larger corpuses of knowledge and see what impact they may have over dialog knowledge. Another possible aspect worth considering is the addition of a C3 variant, the “I’m Alive” boards. Following the September 11th attacks, multiple boards sprang up around New York City announcing the names and present shelter location of survivors to concerned friends and family members outside of the disaster area. Adding such functionality would be a trivial programming exercise, and would provide a quicker and more concentrated way for bidirectional communications.

References

- Beer, M. D., R. Hill, et al. (2003). Deploying an agent-based architecture for the management of community care. International Conference on Autonomous Agents, Proceedings of the second international joint conference on Autonomous agents and multiagent systems, Melbourne, Australia.
- Light, M. and M. T. Maybury (2002). "Personalized multimedia information access." Communications of the ACM 45(5): 54-59.
- Moore, R. and G. Gibbs (2002). Emile: Using a chatbot conversation to enhance the learning of social theory. Huddersfield, England, Univ. of Huddersfield.
- Pasca, M. A. and S. M. Harabagiu (2001). High Performance Question/Answering. Annual ACM Conference on Research and Development in Information Retrieval, New Orleans, LA, ACM Press.
- Potter, S. (2001). A Survey of Knowledge Acquisition from Natural Language. TMA of Knowledge Acquisition from Natural Language. Edinburgh. 2003:
- Radiv, D., W. Fan, et al. (2005). "Probabilistic Question Answering on the Web." Journal of the American Society for Information Science and Technology 56(6): 571-583.
- Voorhees, E. M. (2001). Overview of the TREC 2001 Question Answering Track. Text REtrieval Conference.
- Voorhees, E. M. and D. M. Tice (2000). Building a Question Answering Test Collection: 200-207.
- Vrajitoru, D. (2003). Evolutionary Sentence Building for Chatterbots. Genetic and Evolutionary Computation Conference (GECCO), Chicago, IL.
- Wallace, R. S. (2004). The Anatomy of A.L.I.C.E. A.L.I.C.E. Artificial Intelligence Foundation, Inc.
- Winograd, T. (1977). Five Lectures on Artificial Intelligence. Fundamental Studies in Computer Science. A. Zampolli. North Holland. 5: 399-520.
- Woods, W. A. (1977). Lunar Rocks in Natural English: Explorations in Natural Language Question Answering. Fundamental Studies in Computer Science. A. Zampolli. North Holland. 5: 521-569.