Interaction Analysis of the ALICE Chatterbot: A Two-Study Investigation of Dialog and Domain Questioning

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: 6775

Abstract

This paper analyzes and compares the data gathered from two previously conducted ALICE chatterbot studies that were focused on response accuracy and user satisfaction measures for six chatterbots. These chatterbots were further loaded with varying degrees of conversational, telecommunications and terrorism knowledge. From our prior experiments using 347 participants, we obtained 33,446 human/chatterbot interactions. It was found that asking the ALICE chatterbots 'are' and 'where' questions resulted in higher response satisfaction levels, as compared to other interrogative-style inputs because of their acceptability to vague, binary or clichéd chatterbot responses. We also found a relationship between the length of a query and the users perceived satisfaction of the chatterbot response; where shorter queries led to more satisfying responses.

1 Introduction

Obtaining relevant yet concise information from online repositories has always been a problem. While search engines have mainly focused on the relevance aspect, they have not paid as much attention to conciseness. One way of addressing this issue has been through the use of natural language dialog systems (NLDS), where users can input natural language queries and expect to receive concise natural language responses.

One of the better conversationalists in NLDS is the ALICE chatterbot. ALICE, which stands for Artificial Linguistic Internet Chat Entity, is a type of dialog-driven chatterbot developed in 1995 by Richard Wallace [1]. ALICE chatterbots are built to function as general conversationalists, but they can be quickly supplemented with specific knowledge to function as a customer service agent, information retrieval agent or language chatting partner [2, 3].

In this paper we analyze data obtained from two similar studies conducted at The University of Arizona using modified ALICE chatterbots. While these two prior studies focused on the knowledge acquisition and delivery aspects of chatterbots in both the telecommunications and terrorism domains, this study differs by analyzing the types of questions posed to the systems and the subject's response satisfaction levels to the chatterbot replies. This type of analysis will be able to provide guidance to future chatterbot knowledge developers as to what types of user queries are most common as well as what areas of knowledge development will need extra emphasis in order to provide the most satisfying responses. We further analyze the effects of query length on response satisfaction measures in order to judge its relative impact. We believe that answers to these questions can have a vital impact on the knowledge acquisition activities for future chatterbot systems.

The rest of this paper is organized as follows. Section 2 is a literature review and discusses the similarities of ALICE chatterbots and how they fit in the NLDS hierarchy. Section 3 asks a series of research questions about the retrieval characteristics of ALICE chatterbots. Section 4 introduces the reader to our two prior ALICE chatterbot studies on telecommunications and terrorism knowledge and presents the system architecture used for both studies. Section 5 is the experimental design segment and lays a framework for our current evaluation. Section 6 is our results section with a discussion of what the results mean. Finally, Section 7 wraps up with conclusions and possible avenues of future research.

2 Literature Review

NLDSs form an interesting and dialog-oriented intersection between human-beings and computers. They allow for simple and natural communication, all while returning concise information to the user. These systems can be wholly automated to perform routine functions such as answer common questions, educate users about a particular topic [4] or be semi-automated as in case of a helpdesk assistant [3] or even try to predict what the user will ask next [5]. These systems can also be loaded with domain-specific knowledge adding to their flexibility of application. Automated knowledge gathering is one possible way of quickly building up a knowledge repository, however, these systems must deal with problems of redundancy, inconsistency and unreachable answers [6].

One particular aspect of NLDS is Question Answer (QA) systems [7]. Figure 1 illustrates a synthesized QA framework extracted from the works of Voorhees, Pasca and Vrajitoru [8-10].

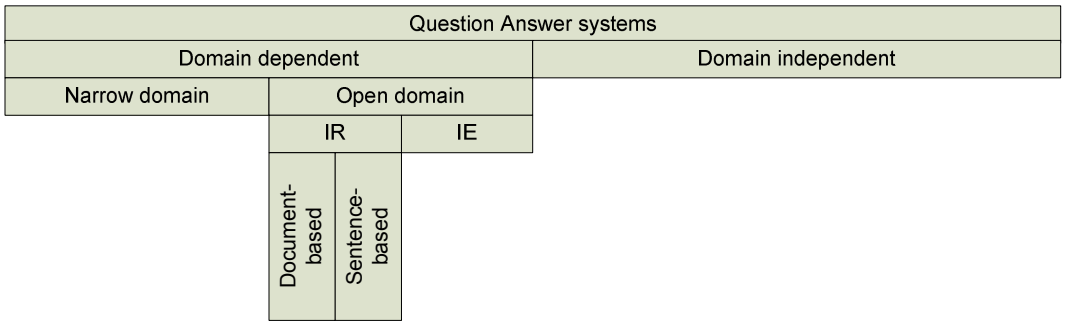

Figure 1. A Question Answer system taxonomy

This framework will describe each sub-area of QA systems; what they do, how they are important, how they differ from one another and provide an overview of how they all link together. We will discuss each component in detail below.

2.1 Question Answer Systems

QA systems use natural language processing methods to select answers based on a search of linguistic features [7]. These features can be syntactic, i.e., relying on the structure of the sentence such as NP-VP-NP patterning; or semantic where ontologies and similar corpora attempt to assign a meaning to the words [11]. These systems can also vary greatly on:

- Knowledge sources used (Domain dependent and independent sources)
- Breadth of domain expertise (Narrow and open domain systems)
- Type of information to obtain (Information retrieval and information extraction)
- And the type of response to give (Document and sentence-based IR systems)

Each of these QA system characteristics can be found as a separate entity in Figure 1.

2.2 Domain Dependent and Independent Systems

QA systems can vary based on the characteristics of their knowledge source(s), such as whether or not the knowledge-base was created specifically for computer usage. They can be classified in one of two different categories; domain dependent or domain independent systems.

Domain independent systems use external knowledge sources, such as an online encyclopedia, that are not specifically built for computational consumption. One example of a domain independent system is MURAX which uses the online Grolier encyclopedia to answer its queries [12].

Conversely, domain dependent systems depend on specially tailored knowledge-bases [8]. These knowledge-bases can be ontology-based as in the case of systems relying on Doug Lenat's Cyc knowledge-bases for semantic meaning and disambiguation [11] or more simply a collection of domain-relevant data that a system uses to answer a specific question.

2.3 Narrow and Open Domain Systems

Domain dependent systems can be further broken into two subcategories: narrow and open domain systems [8, 9]. In narrow domain systems, the goal is to attempt conversational fluency in limited domains of expertise [4]. Example systems include STUDENT which solved algebraic word problems [13], Winograd's SHRDLU which answered natural language queries about a fictitious Block World [13] and LUNAR which responded to geological queries on lunar rock data [14].

Open domain systems possess a more diverse (i.e., generalized) repertoire of topics. These systems are not limited to any one particular area or domain. Instead, these systems can field questions from multiple disciplines and can be further classified into two major categories; information retrieval and information extraction [8, 9].

2.4 Information Retrieval and Information Extraction

In Information Extraction (IE), the goal is to extract relevant contextual information from text and to fill that data into pre-defined templates. This field is well-represented by the Message Understanding Conference (MUC).

In contrast, Information Retrieval (IR) attempts to retrieve a whole or partial document for the user. Examples vary from modern search engines that rely on shallow keyword matching techniques to deeper systems that attempt to retrieve a snippet of text to the user within the context of the query. This field is represented by the Text Retrieval Conference (TReC) [9, 15].

2.5 Document and Sentence-based IR Systems

IR is composed of two smaller classes; the document-based and sentence-based retrieval systems [10]. The objective of document-based systems is to return a set of relevant documents to the user, much like a search engine.

To the contrary, sentence-based retrieval systems return only a small snippet of text to the user. These systems can also vary in the styles of answers given, from binary yes/no or true/false responses [8], to those responses requiring a synthesis of material from various locations (i.e., "Why do terrorists hate the West") [16] and many others between these two extremes. ALICE chatterbots, a type of Question Answer system, fit into the sentence-based retrieval category [10] primarily because of their sentence-oriented response capability.

2.5.1 ALICE

ALICE uses XML-based Artificial Intelligence Markup Language (AIML) files to hold its internal collection of knowledge. This open-source knowledge-base makes ALICE robust and able to quickly extend into new knowledge domains [1]. ALICE seeks to mimic conversation rather than understand it [17]. This method of conversational mimicry has allowed ALICE to win the Loebner Prize for most human chatterbot in 2000, 2001 and 2004. However, because of its simplistic pattern-matching mechanisms, ALICE lacks cognitive ability and will miss certain types of interactions [2]. Wallace argues that ALICEbots use Case Based Reasoning (CBR) to represent responses [3], which is beneficial to system performance because

CBR does not require the computational overhead that other reason-based systems would demand [18].

Prior studies have shown that ALICE is used more like a search engine rather than a conversational tool [19]. This finding complements many of the suggested uses of ALICE as a fact-driven conversationalist that can deliver domain-specific information to the public in a personable and tireless manner. Following ALICE's use as a search engine, a focus on interrogative usage is an important area of QA system research [20].

2.5.2 ALICE Chatterbot Studies

There are several notable research studies where ALICE has been used. The first of which was an English and German conversational partner for Chinese students [21]. This study focused on the usefulness of the ALICE platform as a stand-alone conversationalist and produced some unexpected results. Jia tracked the categories of topical discussion and found that participants most frequently discussed love, the study of the English language and friendship. However, a high proportion of students did not like the chatterbot responses and made 'bad' comments about the system. A majority of subjects interacted for a short period of time before leaving. This system used a smaller than available knowledge-base which may have directly contributed to the unfavorable study observations.

This anthropomorphism of a computer program as a social actor was not unexpected [22]. De Angeli conducted behavioral studies using ALICE and discovered that the friction arose from power differences between users and the system, where users were trying to exert their dominion of control over the system. From De Angeli's work it was found that some users will promote an abusive environment to establish their dominance. In addition, as stated by Pejtersen, users will sometimes use the system in unintended ways [23, 24].

Another teaching-related chatterbot study was that of a Geometry tutor. This ALICE-based instantiation was a prototypical chatterbot designed to assist students with concepts in Euclidean geometry [25]. A more important realization from this study was that Han believed complementing Euclidean domain knowledge with conversational knowledge would augment the ability of the chatterbot to assist users in reformulating misunderstood queries within a natural conversational context. Han's belief was supported in a different study [26], where conversational knowledge coupled with telecommunications definitions performed better than the telecommunications definitions alone.

Another relevant study was the Emile chatterbot at the University of Huddersfield. Emile was designed to emulate four different social theorists and was offered to socio-political students as a teaching tool to better understand the different sociological perspectives [19]. Students were given socio-political class assignments and were instructed to use the Emile chatterbot. Unfortunately it was determined that students were more interested in using Emile as a search engine to quickly answer the assignment questions rather than converse with the system as expected.

In a similar chatterbot study, AutoTutor was tasked with interacting with students on the topics of computer literacy and conceptual physics [27]. From an analysis of dialog interactions, it was again found that students were using the system as a search engine to find answers to definitional queries [28].

Following the findings of the Emile and AutoTutor studies, Voorhees noted that most search engine queries are definitional in nature [29]. Queries such as *Who is Colin Powell* and *What is mold* are common types of interrogative-based methods of obtaining information [30].

In a similar search engine study, Zuckerman looked into the accuracy of returned results as a function of query length [31]. It was found that an inverse relationship exists between the two variables. As query length decreases, system accuracy improves. Zuckerman determined that shorter queries were more apt to provide more accurate results.

3 Research Questions

From the data gathered on our prior ALICE chatterbot studies [4, 26], we pose several research questions which we believe will help provide some insight into the strengths and weaknesses of current ALICE knowledge-bases. We ask our questions in such a way that the answers generated can be directly used by future chatterbot developers to better position chatterbot responses.

The first question we ask relates to what question-types are most commonly posed to chatterbot systems.

- What similarities in interrogative selection frequency exist between different knowledge domains?

With this question, we examine the frequency counts of interrogative selection and determine if any interrogatives are used more frequently than others. The answer to this question will tell us which question-types should be the focus in future knowledge-gathering activities. We adapt a naïve view and assume that interrogatives starting with 'wh*' will occur most frequently.

Our second research question looks at how well users perceive chatterbot responses to particular question-types.

- What interrogative types best answer user queries and why?

In this question, we seek to discover which interrogatives can best answer user queries. As a consequence, we can also examine which interrogatives cannot answer user queries, why

they cannot and address possible ways to correct these instances in the existing ALICE knowledge-bases. We would assume that interrogatives that seek specific responses would return the best answers to the users (i.e., 'what' and 'who' versus 'how' and 'why').

For our third research question, we wish to examine how the length of a user query might impact a user's satisfaction with a chatterbot response.

- How does the length of a query affect the satisfaction level of chatterbot responses?

In the work of Zuckerman and Horvitz on search engine queries, they found that shorter queries led to more accurate document-based results. We will apply this assumption to the sentence-based ALICE chatterbot and assume that shorter queries will also lead to more accurate chatterbot responses.

4 System Design

The chatterbot system we used was composed of five distinct components: Chat Interface, Chat Engine, AIML files, Logging and Evaluation modules. Figure 2 graphically illustrates the different chatterbot components.

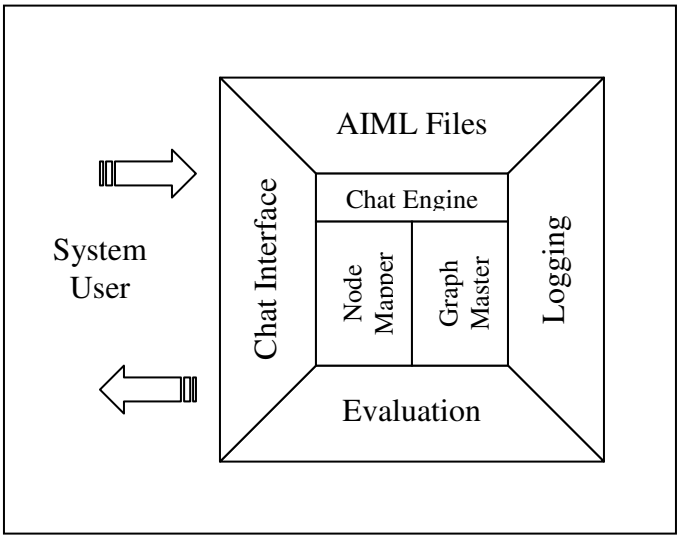

Figure 2. Our system's chatterbot components

The first three components, Chat Interface, Chat Engine and AIML Files, are considered essential components to the ALICE chatterbot. The Chat Interface module allows the system to handle system inputs and responses by performing appropriate conversions of textual data to more friendly XML-based content.

The Chat Engine is the core algorithmic component of the system and is comprised of two sub-entities; the Node Mapper and the Graph Master. When the system is initialized, the Node Mapper constructs a memory-resident directed graph of all the AIML patterns. Later, when a user poses a query to the system, the Graph Master will traverse the directed graph to best match the input.

The AIML files are considered to be the brains behind the system. Specific knowledge input patterns and appropriate chatterbot responses are stored within these files. This flexible arrangement permits the AIML-enabled system to easily migrate into new domains of knowledge with the addition of new domain queries and responses. An AIML category used in the TARA terrorism knowledge-base is listed below. User queries are matched against patterns and chatterbot responses arise from the pre-defined templates.

```
<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>
```

As an example of how this system works, consider the following query “Who is Bin Laden.” The Chat Interface passes the query to the Chat Engine, where the Graph Master resides. The Graph Master then sets out to best match the query from the most general terms to the most specific. Assume that an AIML node map can be represented in Figure 3.

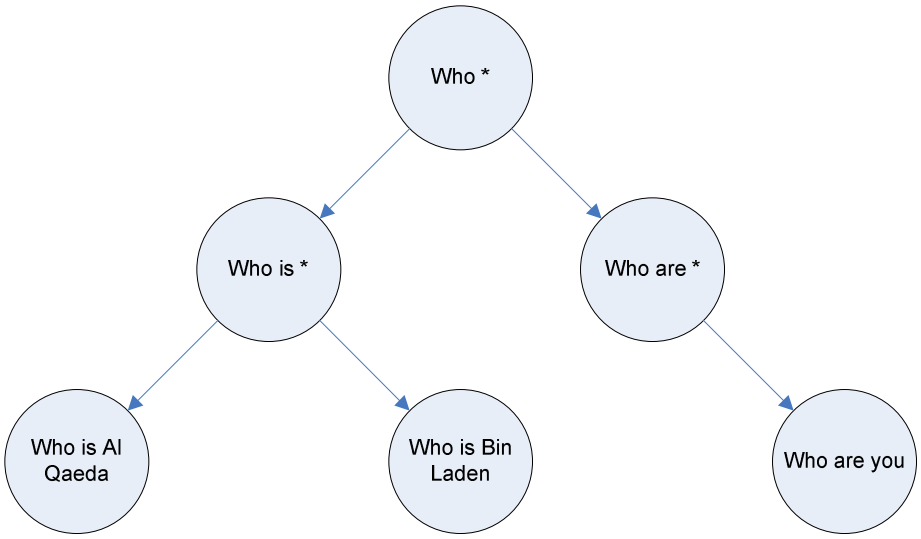

Figure 3. Example AIML Node Mapping

From this query, the Graph Master will first match the node "Who *" to the query where '*' is a wildcard placeholder. The Graph Master will then look at all of the child nodes to determine whether a better, more specific, match can be made. In this case "Who is *" is a better match. Continuing further, we find the exact match in the next level and the appropriate node-specific response is then sent back to the Chat Interface.

In a majority of cases, exact matches are rare. In cases of inexact matches, the Graph Master utilizes wildcard matching to best correspond to queries. Supposing that a query "Who is George Washington" was posed to our hypothetical system. Without having an exact match, the Graph Master will settle on the "Who is *" node and return it's more generic response. It is believed that the more nodes available will directly increase the perceived depth of the system, as illustrated by an actual user interaction.

```
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.

5 Experimental Design

5.1 The AZ-ALICE and TARA Data

For our study we analyzed the data gathered from two prior system experiments; AZ-ALICE and TARA [4, 26]; which were both closely based on the original ALICE system. Both experiments shared similar goals in measuring response satisfaction of chatterbot replies. Table 1 shows the similarities and differences between the six systems.

	Chat Interface	Chat Engine	AIML	Logging	Evaluation
Original ALICE	Uses XML to chat with users through a crude Jetty interface	Uses off the shelf ALICE Program D	Typical setup uses Standard AIML	Logs everything to a monolithic XML Log file	Does not provide a user-based evaluation component
AZ-ALICE	Uses a customized perl skin to chat and for evaluation purposes	Same as Original ALICE	Customized Telecommunications AIML for domain knowledge and Standard AIML for conversation	Keeps XML logs on a per-user basis	Customized perl script that allows users to evaluate and suggest new patterns
TARA	Same as AZ-ALICE	Same as Original ALICE	Customized Terrorism AIML for domain knowledge and Standard and Wallace AIML for conversation	Same as AZ-ALICE	Same as AZ-ALICE

Table 1. Differences between original ALICE, AZ-ALICE, and TARA

In all six systems, two of the three essential parts of the chatterbot remained the same; the Chat Engine and the Chat Interface. The key difference between the chatterbots was the AIML knowledge-bases used. AZ-ALICE focused on the telecommunications domain whereas TARA handled terrorism-related knowledge. Both systems utilized conversational knowledge in their

control chatterbots; AZ-ALICE used Standard AIML and TARA used the Standard and Wallace AIML set. Depending upon the knowledge-bases used by a particular chatterbot, telecommunications, terrorism, general conversation or a mixture of both conversation and domain knowledge may be returned to the user.

5.2 AIML Knowledge

In both AZ-ALICE and TARA, we used three chatterbots apiece where one chatterbot (the control chatterbot for its respective study) was devoted to conversational knowledge, one to specific domain knowledge and the third to a mixture of conversational and domain knowledge. Table 2 illustrates the breakdown of AIML category rules that were used for each chatterbot.

	Dialog categories	Domain categories	Total categories
AZ-ALICE - Dialog	23,735	0	23,735
AZ-ALICE - Domain	3,892	298	4,190
AZ-ALICE - Both	23,735	298	24,032
TARA - Dialog	41,873	0	41,873
TARA - Domain	0	10,491	10,491
TARA - Both	41,873	10,491	52,354

Table 2. Category Breakdown of the Six Chatterbots

From this breakdown, there are several differences that require a further explanation.

- AZ-ALICE – Dialog: This was the control chatterbot for the AZ-ALICE study. Its 23,735 conversational categories were derived from the Standard AIML set which is freely available at www.alicebot.org. These are believed to be the same categories that Jia used in the English part of their conversational partner chatterbot.

- AZ-ALICE – Domain: This one was not a true domain-only chatterbot. It consisted of 298 domain categories as well as a limited set of dialog. Dialog categories were selected because of the limited responses from the domain categories. The dialog categories selected were the core categories that allowed

AZ-ALICE – Domain to provide sufficient responses to a majority of questions. The domain categories were handcrafted and based on telecommunications knowledge.

- TARA – Dialog: This was the conversational control chatterbot for the TARA study. Its 41,873 categories were obtained from the Standard and Wallace set which helped ALICE to win its early Loebner contests.

- TARA – Domain: This chatterbot used true domain-only terrorism knowledge. Its 10,491 terrorism-related categories were gathered from the glossaries of several reputable terrorism websites. Further details of TARA's domain knowledge can be obtained from [26].

The other item of interest is that the total number of categories does not necessarily represent the sum of dialog and domain categories. This is a result of the way the Node Mapper handles duplicate categories. There were several instances where domain knowledge was already represented by the dialog AIML. When this situation occurs, the Node Mapper automatically drops the duplicate node from the directed graph.

5.3 Study Participants

Participants from both studies were self-selected university students. Subjects for the AZ-ALICE project were obtained from several sections of a freshman-level Management of Information Systems (MIS) introductory course in Fall 2003, while TARA participants came from various undergraduate and graduate-level MIS classes in Spring/Summer 2004. Both studies also differed on the number of subjects used, however, given the amount of chatterbot interactions gathered we are able to make statistical comparisons between them. For both of these studies we were more interested in approximating conditions similar to that of 'the wild'

using a demographic subset that would be more likely to use such communicative instruments. Subjects were further assigned to one of the six particular chatterbots by the following metric:

- AZ-ALICE: based on which class section the student belonged
- TARA: based on a random assignment algorithm run against their University ID

Subjects only interacted with one chatterbot and across the two experiments there were no subjects that participated in both studies. In both studies, subjects were asked to interact with the system for approximately one-half hour and were given a participation incentive through the prospect of bonus points or random gift cards, depending upon the study. The break-down of study participants are shown in Table 3.

	Study Participants
AZ-ALICE - Dialog	74
AZ-ALICE - Domain	91
AZ-ALICE - Both	92
TARA - Dialog	30
TARA - Domain	30
TARA - Both	30

Table 3. Study Participants

In both studies, participants were neither told that were was more than one chatterbot, nor which chatterbot they were assigned. Participants were asked to communicate with their chatterbot either on telecommunications (AZ-ALICE) or terrorism (TARA) topics and then evaluate the responses of the chatterbot and rate their satisfaction level of the response using a one-to-seven Likert scale (one – strongly dissatisfied to seven – strongly satisfied). The evaluation phase did differ slightly between studies, where AZ-ALICE asked for the response evaluation at the conclusion of chatting while TARA presented its evaluation after each interaction. Figure 4 illustrates the integrated evaluation mechanism from the TARA studies.

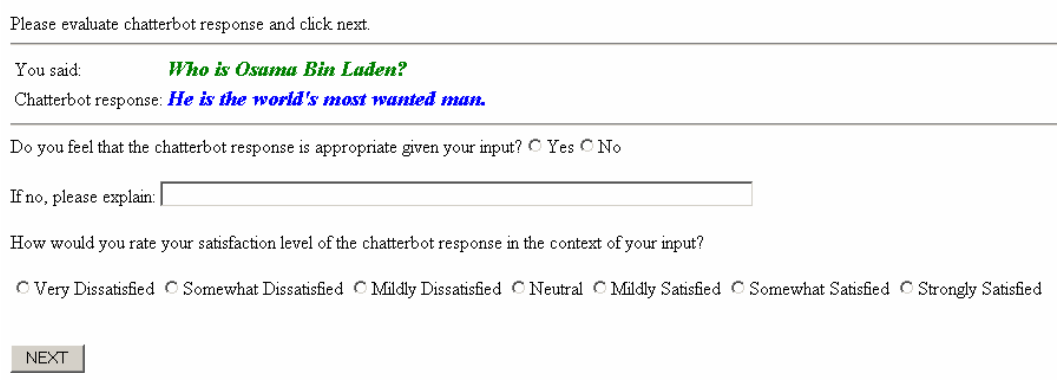

Figure 4. Screenshot of TARA's evaluation process

5.4 System Evaluation Metrics

In this paper we analyzed the data from AZ-ALICE and TARA and arrived at three metrics to answer our research questions; a count of interrogative usage, response satisfaction scores and the length of user queries.

The count of interrogatives was concerned about the frequency that interrogatives are used to begin user queries. Example types include the 'wh*' interrogatives: who, what, when, where and why. Queries ending with question marks are identified as interrogative and the starting words are tabulated in a frequency count. While we agree that identifying interrogatives by their punctuation characteristics is not ideal, this method has a prior basis [32].

Measurements of response satisfaction were common between studies. This metric was subjectively measured on a one-to-seven Likert scale by study participants for each chatterbot input-response pair. Aggregate measures of response satisfaction were then composed for each chatterbot and different interrogatives.

Query length was a simple measure that averaged the number of words in a user input for each chatterbot [33]. The response satisfaction of query length using different interrogatives was also investigated.

6 Experimental Results

6.1 Participants sought definitional facts across knowledge domains

In analyzing the interaction patterns of users across the different chatterbots, it was discovered that participants were generally inquisitive. Table 4 shows a breakdown of chatterbot interactions as well as the percentage of user inputs identified as interrogative.

	# Interrogatives		# Interactions		Interrogative Ratio	
Chatterbot	Average	Std Dev	Average	Std Dev	Average	Std Dev
AZ-ALICE Dialog	49.2	45.2	131.8	94.8	37.2%	0.0236
AZ-ALICE Domain	61.8	41.0	108.8	65.0	58.9%	0.0316
AZ-ALICE Both	44.2	26.6	111.9	59.6	41.8%	0.0392
TARA Dialog	28.8	22.1	50.8	25.7	50.5%	0.0828
TARA Domain	15.3	17.0	28.3	20.5	52.1%	0.1163
TARA Both	23.3	14.8	37.9	20.0	62.2%	0.0721

Chatterbots	F Calculated	F Critical	p-value
Dialog chatterbots	0.2848	0.6162	7.56E-06
Domain chatterbots	0.2717	0.6281	1.09E-06
Both' chatterbots	0.5438	0.6276	0.0156

Table 4. Participant usage of interrogatives across chatterbots

Within this table, there are several things worth further explanation. The top portion of the table shows some basic statistics on the average number of interrogatives and interactions captured for each of the six chatterbots. The right-hand side of the table details the average number of interrogatives captured for each user. From this part of the table, we found that participants of TARA were more inclined to use interrogatives than those of AZ-ALICE (TARA's 50.5% versus AZ-ALICE's 37.2% for Dialog and TARA's 62.2% versus AZ-ALICE's 41.8% for Both), with the exception of the two domain chatterbots (TARA's 52.1% versus AZ-ALICE's 58.9% for Domain). Using a pairwise t-test, we found that the p-values were all less than 0.05. It is interesting to note how much communication was question-oriented.

In testing the interrogative usage variances between the two studies, the bottom portion of Table 4 is an F-test measure and demonstrates that the variances between the chatterbots are within acceptable measures.

Looking further into the frequency counts of the various interrogatives used, eight particular words always appeared frequently in all six chatterbots: *are*, *do*, *how*, *is*, *what*, *where*, *who* and *why*. For the AZ-ALICE studies, these words were consistently in the top 15 and for TARA they were the top 10 most frequently used interrogatives. Investigating these terms further, Tables 5 and 6 illustrate the frequency counts of each of these interrogatives across all six chatterbots.

ALICE Studies - Telecommunications Domain								
Dialog			**Domain**			**Both**		
Interrogative	Count	% Use	Interrogative	Count	% Use	Interrogative	Count	% Use
What	985	36.4%	What	2,159	50.2%	What	1,405	46.3%
Do	628	23.2%	Do	725	16.9%	Do	522	17.2%
How	503	18.6%	How	447	10.4%	How	448	14.8%
Who	168	6.2%	Why	276	6.4%	Who	188	6.2%
Why	164	6.1%	Who	261	6.1%	Why	185	6.1%
Are	130	4.8%	Are	199	4.6%	Are	158	5.2%
Where	79	2.9%	Is	126	2.9%	Where	71	2.3%
Is	49	1.8%	Where	107	2.5%	Is	59	1.9%

Table 5. Interrogative frequency use of Telecommunications knowledge

TARA Studies - Terrorism Domain								
Dialog			**Domain**			**Both**		
Interrogative	Count	% Use	Interrogative	Count	% Use	Interrogative	Count	% Use
What	228	36.3%	What	136	37.4%	What	176	33.8%
Do	112	17.8%	Who	67	18.4%	Do	106	20.3%
How	57	9.1%	How	44	12.1%	Who	67	12.9%
Who	57	9.1%	Where	35	9.6%	How	57	10.9%
Is	50	8.0%	Do	32	8.8%	Where	38	7.3%
Are	49	7.8%	Are	21	5.8%	Are	31	6.0%
Why	40	6.4%	Is	18	4.9%	Is	24	4.6%
Where	35	5.6%	Why	11	3.0%	Why	22	4.2%

Table 6. Interrogative frequency use of Terrorism knowledge

From these two tables, the interrogative *What* was the most frequently used in all six chatterbots. It was interesting that several of the frequently observed interrogatives do not fit

under the traditional 'wh*' interrogative family (i.e., *do*, *are*, *is*). *Do* was one such occurrence that was frequently second to *What* except in pure Terrorism domain knowledge. Another interesting fact was the frequency of interrogative *Why* and the complete absence of *When*. *When*'s conspicuous absence was not expected. It would appear that temporal questioning was not very popular at least within these two domains of knowledge.

Since *What* appeared with the greatest frequency, we further investigated this interrogative by expanding its frequency counts to include the second word as well. Tables 7 and 8 show the *What* expansion.

ALICE Studies - Telecommunications Domain					
Dialog		**Domain**		**Both**	
Interrogative	Count	Interrogative	Count	Interrogative	Count
What is	404	What is	1265	What is	802
What do	148	What do	227	What do	114
What are	82	What are	151	What are	105

Table 7. *What* expansion frequency counts of Telecommunications knowledge

TARA Studies - Terrorism Domain					
Dialog		**Domain**		**Both**	
Interrogative	Count	Interrogative	Count	Interrogative	Count
What is	63	What is	69	What is	69
What do	50	What are	13	What are	23
What are	34	What do	8	What do	21

Table 8. *What* expansion frequency counts of Terrorism knowledge

From this expansion, we found that *What is*, *What do* and *What are* were in the top 4 frequency counts for each chatterbot. Furthermore, *What is* appeared most often suggesting that users were seeking definitional types of responses.

6.2 *Are* interrogatives had the highest Response Satisfaction rating

Looking further into which interrogatives led to better chatterbot responses, we analyzed the Response Satisfaction levels for each of the eight frequently observed interrogatives. Tables 9 and 10 show the Response Satisfaction ratings for the various interrogatives.

ALICE Studies - Telecommunications Domain					
Dialog		**Domain**		**Both**	
Interrogative	Avg Rating	Interrogative	Avg Rating	Interrogative	Avg Rating
Are	4.7077	Are	4.4372	Do	4.6034
Is	4.6531	Where	4.2897	Are	4.5949
Where	4.5443	Who	4.2222	How	4.3705
Do	4.5000	Do	4.1310	What	4.2135
Who	4.4345	How	4.0962	Who	4.2128
What	4.3107	Why	3.9529	Where	3.9155
How	4.3101	Is	3.7063	Why	3.8541
Why	4.2439	What	3.6281	Is	3.6780

Table 9. Interrogative Response Satisfaction scores for Telecommunications knowledge

TARA Studies - Terrorism Domain					
Dialog		**Domain**		**Both**	
Interrogative	Avg Rating	Interrogative	Avg Rating	Interrogative	Avg Rating
Are	4.9388	Where	3.2571	Are	4.8387
Where	4.6286	Who	2.9552	Who	4.5821
How	4.1404	What	2.5074	How	4.2281
Who	3.9825	Are	2.0476	Where	3.9474
Do	3.9821	Is	2.0000	Do	3.9434
Is	3.9000	How	1.9545	Is	3.7083
Why	3.8750	Do	1.8438	What	3.6818
What	3.6798	Why	1.3636	Why	3.5909

Table 10. Interrogative Response Satisfaction scores for Terrorism knowledge

From examining the above tables, it becomes apparent that *Are* interrogatives appear to be more satisfying than the others in AZ-ALICE-Dialog, AZ-ALICE-Domain and TARA-Both (p-value < 0.001 of *Are*'s satisfaction versus other interrogatives in each of these three chatterbots). The *Are* interrogative was statistically equivalent to *Do* in AZ-ALICE-Both and to *Where* in TARA-Dialog. The only case where *Are* did not outright rate the highest or statistically tie, was in the pure terrorism domain of TARA. In further probing, we found that *Are*'s success in dialog-enabled chatterbots appears to come from vague and general responses, such as offering a pre-programmed pseudo-opinion rather than a fact. These responses can be binary, non sequitur or clichéd, yet still be highly satisfying (as shown in the examples below).

```
User: Are there any weapons of mass destruction in the middle east?
System: Yes I think there are.
```

```
User: Are you scared of terrorists?
System: I fear nothing.
```

To take our analysis further, we broke apart the Both chatterbots into their constituent parts of Dialog and Domain knowledge. From there we can cross-check our initial findings to determine what question types are best suited to the various conversation and knowledge domains. Tables 11 and 12 show the Both chatterbots on the left column and their breakdown into Dialog and Domain to the right.

ALICE Studies - Telecommunications Domain					
Both		**Both (Dialog)**		**Both (Domain)**	
Interrogative	Avg Rating	Interrogative	Avg Rating	Interrogative	Avg Rating
Do	4.6034	Are	4.6026	Do	5.8333
Are	4.5949	Do	4.5891	What	4.9545
How	4.3705	How	4.3647	How	4.5833
What	4.2135	Who	4.2246	Is	4.2000
Who	4.2128	What	4.1367	Why	3.2500
Where	3.9155	Where	3.9429	Are	4.0000
Why	3.8541	Why	3.8674	Where	2.0000
Is	3.6780	Is	3.6296	Who	2.0000

Table 11. Response Satisfaction breakdown of Both in Telecommunications

TARA Studies - Terrorism Domain					
Both		**Both (Dialog)**		**Both (Domain)**	
Interrogative	Avg Rating	Interrogative	Avg Rating	Interrogative	Avg Rating
Are	4.8387	Are	5.1481	Who	4.3684
Who	4.5821	Who	4.6667	Where	3.4000
How	4.2281	How	4.6512	What	3.3830
Where	3.9474	Do	4.3291	Is	3.1111
Do	3.9434	Where	4.1429	How	2.9286
Is	3.7083	Is	4.0667	Do	2.8148
What	3.6818	Why	4.0556	Are	2.7500
Why	3.5909	What	3.7907	Why	1.5000

Table 12. Response Satisfaction breakdown of Both in Terrorism

From the view given by the preceding tables, *Are* interrogatives appear to be more satisfying in dialog than domain knowledge (TARA has a p-value < 0.01 but AZ-ALICE lacks a sufficient number of instances to draw any conclusions). This is not an unexpected result when correlating *Are* back to the pure Dialog chatterbot results of Tables 9 and 10, where *Are*'s

Response Satisfaction was 4.7077 for AZ-ALICE-Dialog and 4.9388 for TARA-Dialog, compared to AZ-ALICE-Both (Dialog) of 4.6026 and TARA-Both (Dialog) of 5.1481. In each case, the Dialog-based *Are* interrogative has the highest Response Satisfaction rating (p-value < 0.01).

The other important finding is the low Response Satisfaction of *Why* and *What*. These interrogatives generally scored very low across all six chatterbots. It is believed that this is the result of chatterbot vagueness when a specific answer was desired, as shown in the examples below.

```
User: What is a terrorist attack?
System: I think you already know the answer.

User: Why is he considered a terrorist?
System: Perhaps his creative genius perhaps there is a another reason.
```

6.3 Response satisfaction scores show a weak relation to query length

We further investigated the role that query length may have on response satisfaction scores. Applying the search engine research of Zuckerman, we had hypothesized that our chatterbot response satisfaction scores will show a similar inverse relationship to query length. Table 13 illustrates the query lengths observed for each chatterbot.

	# of Inputs	Query Length Average	Std Dev
AZ-ALICE - Dialog	3,906	5.6521	2.9233
AZ-ALICE - Domain	5,906	5.5320	2.8195
AZ-ALICE - Both	4,284	5.3987	2.8160
TARA - Dialog	931	6.6584	3.0042
TARA - Domain	487	6.7351	3.3225
TARA - Both	731	7.0766	3.7453

Table 13. Query lengths across chatterbots

Tables 14 and 15 show a break-down of the average number of query words for each of the eight most frequently observed chatterbot interrogatives.

ALICE Studies - Telecommunications Domain					
Dialog		Domain		Both	
Interrogative	Word length	Interrogative	Word length	Interrogative	Word length
Why	5.8415	Is	6.1746	Why	5.9135
Is	5.6735	How	5.9821	How	5.8817
Do	5.5748	Do	5.9779	Is	5.7288
What	5.5360	Why	5.8225	Do	5.6782
How	5.2366	Are	5.1910	Are	5.0886
Where	5.1392	What	5.1408	What	4.9374
Are	5.0154	Where	4.9346	Where	4.7606
Who	4.5000	Who	4.0766	Who	4.4309

Table 14. Query lengths for each interrogative in Telecommunications

TARA Studies - Terrorism Domain					
Dialog		Domain		Both	
Interrogative	Word length	Interrogative	Word length	Interrogative	Word length
Do	7.6875	Do	8.9688	Do	8.5094
Why	6.8750	Why	8.8182	How	7.5088
Is	6.6800	Is	7.8333	What	6.3977
How	6.6140	How	7.5909	Is	6.3333
What	6.4868	What	6.8750	Are	5.9032
Are	6.2449	Are	6.5714	Who	5.0597
Where	5.5429	Where	5.2286	Why	5.0000
Who	5.2105	Who	4.9254	Where	4.3421

Table 15. Query lengths for each interrogative in Terrorism

From these tables, *Why* questions appear to have the longest average query length. *Who* interrogatives are generally the shortest and questions in Terrorism are generally longer than their counterparts in Telecommunications.

Aggregating all interrogative-based response satisfaction scores on a per chatterbot level and analyzing their relation to query length, yields Table 16.

	Regressed Slope of Reponse Satisfaction scores to Query Length	p-value
AZ-ALICE - Dialog	-0.0231	0.0050
AZ-ALICE - Domain	0.0173	0.0276
AZ-ALICE - Both	-0.0127	0.1501
TARA - Dialog	-0.0790	0.0003
TARA - Domain	-0.1786	0.0000
TARA - Both	-0.0830	0.0002

Table 16. Relating query lengths to response satisfaction scores

From Table 16, we found that five of the six chatterbots exhibited a negative relation between query length and response satisfaction, which is consistent with our expectations. We did not expect that any of the chatterbots would show a positive relation, but AZ-ALICE-Domain did. When looking at a p-value < 0.001, we found that all three AZ-ALICE chatterbots had no significant slope, while the three TARA chatterbots did.

One possible explanation for these findings may come from the disparate levels of knowledge used between the two studies. It may be that AZ-ALICE did not possess a sufficient critical mass of knowledge which resulted in poorer ratings regardless of query length. This conclusion of insufficient critical mass is best supported by the observations of AZ-ALICE-Domain which possessed the least amount of knowledge. This unusual chatterbot exhibited a positive relationship between query length and response satisfaction at p-value < 0.05, where longer queries were rated more satisfying. The effect of longer query lengths in a knowledge-starved system would lead to a better chance of the system picking up some keyword and returning a response.

From our interpretation, the TARA chatterbots support the observations of Zuckerman that accuracy is inversely proportional to query length, while the AZ-ALICE chatterbots would appear inconclusive.

As an example, consider the following interactions within the TARA-Both chatterbot.

```
User: what is 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.
Query Length: 4 words
User Satisfaction Level: 7 - Strongly Satisfied

User: What is the terrorist group which toppled the two towers?
System: which are involved in terrorism.
Query Length: 10 words
User Satisfaction Level: 1 - Strongly Dissatisfied
```

From this example interaction, the TARA-Both was better able to key in on the important terms and return appropriate responses with the shorter query. The longer queries had the effect of confusing the chatterbot, with the exception of AZ-ALICE Domain which needed the longer queries to offset its lack of critical knowledge.

7 Conclusions and Future Directions

We believe that the findings generated by this paper have a direct applicability to existing ALICE chatterbots and can help shed light on the design of natural language dialog systems.

In particular, we found that interrogatives are a major source of user inquiries [29]. *Wh** interrogatives and *What* especially, appear to be a good place to focus future knowledge-gathering activities. In both the AZ-ALICE and TARA studies, *What is* * was found to be the most prevalent question type. Collecting more *What is* *, *What are* *, and *What do** types of knowledge categories should help improve dialog systems by providing more specific and more likely to be triggered responses.

We also found that interrogatives beginning with *Are* and *Where* were the most satisfying. We believe this is because general and vague chatterbot responses fit these types of questions best because of their binary, non sequitur or clichéd nature. Conversely, *Why* and similar

interrogatives that expected a specific answer were found the least satisfying when a general response was given. To capitalize on this finding we would further suggest a focus on gathering more knowledge categories with an emphasis on specific responses.

Query length would appear to have an impact on response satisfaction, where longer queries correlated with decreasing user response satisfaction scores. It would be advisable in the future to focus more on gathering knowledge for longer query lengths than shorter to provide some equilibrium. The TARA studies agreed with the shorter query and increased satisfaction result, while the AZ-ALICE studies were inconclusive. We believe that this inconsistency comes from an insufficient critical mass of knowledge used in the AZ-ALICE chatterbots. Future work should evaluate just how many AIML patterns are needed for minimal conversational saliency.

There are a few caveats that we should impart to readers. First off, the use of student subjects in our studies may lead to generalization problems. Although student subjects are a more likely demographic to use such communicative instruments, their comfort-level with imitation-style technology may not reflect the entire stratum of potential users. Secondly, we acknowledge that response satisfaction scores can be a highly subjective measure. However, in our pseudo-random assignment of participants, we feel that we gained a sufficient number of chatterbot responses to offset any random error fluctuations that may arise in the response satisfaction results. Thirdly, the results of this paper may not generalize to all knowledge domains.

Future research should focus on implementing a spell-check mechanism to catch user misspelled domain terms. This problem was identified in the TARA studies, accounting for around 6% of missed responses from the chatterbot. Another suggestion is to add other corpora

of knowledge to the system. Although we drew our conclusions from the telecommunications and terrorism domain, perhaps other areas of interest may result in further insights. Finally, a study to investigate where the balancing point of critical knowledge mass may prove to be useful to future researchers.

References

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

[2] R. S. Russell, *Language Use, Personality and True Conversational Interfaces*. Edinburgh: Univ of Edinburgh, 2002.

[3] R. S. Wallace, "The Elements of AIML Style," in *A.L.I.C.E. Artificial Intelligence Foundation, Inc.*, 2003.

[4] R. P. Schumaker, M. Ginsburg, H. Chen, and Y. Liu, "An Evaluation of the Chat and Knowledge Delivery Components of a Low-Level Dialog System: The AZ-ALICE Experiment," *Decision Support Systems*, vol. 42, pp. 2236-2246, 2007.

[5] M. Awad and L. Khan, "Web Navigation Prediction Using Multiple Evidence Combination and Domain Knowledge," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 37, pp. 1054-1062, 2007.

[6] V. Shen and T. Juang, "Verification of Knowledge-Based Systems Using Predicate/Transition Nets," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 38, pp. 78-87, 2008.

[7] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, and C. Jacquemin, "Terminological Variants for Document Selection and Question/Answer Matching," presented at Association for Computational Linguistics, Toulouse, France, pp. 1-8, 2001.

[8] E. M. Voorhees and D. M. Tice, "Building a Question Answering Test Collection," 2000, pp. 200-207.

[9] M. A. Pasca and S. M. Harabagiu, "High Performance Question/Answering," presented at Annual ACM Conference on Research and Development in Information Retrieval, New Orleans, LA, pp. 366-374, 2001.

[10] D. Vrajitoru, "Evolutionary Sentence Building for Chatterbots," presented at Genetic and Evolutionary Computation Conference (GECCO), Chicago, IL, pp. 315-321, 2003.

[11] D. Lenat, G. Miller, and T. Yokoi, "CYC, WordNet, and EDR: critiques and responses," *Communications of the ACM*, vol. 38, pp. 45-48, 1995.

[12] J. Kupiec, "MURAX: A Robust Linguistic Approach for Question Answering Using an On-Line Encyclopedia," presented at ACM-SIGIR, Pittsburgh, PA, pp. 181-190, 1993.

[13] T. Winograd, "Five Lectures on Artificial Intelligence," in *Fundamental Studies in Computer Science*, vol. 5, A. Zampolli, Ed. North Holland, 1977, pp. 399-520.

[14] W. A. Woods, "Lunar Rocks in Natural English: Explorations in Natural Language Question Answering," in *Fundamental Studies in Computer Science*, vol. 5, A. Zampolli, Ed. North Holland, 1977, pp. 521-569.

[15] S. Potter, "A Survey of Knowledge Acquisition from Natural Language," in *TMA of Knowledge Acquisition from Natural Language*, vol. 2003. Edinburgh, 2001, http://www.aiai.ed.ac.uk/project/akt/work/stephenp/TMA%20of%20KAfromNL.pdf.

[16] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu, "Performance issues and error analysis in an open-domain question answering system," *ACM Transactions on Information Systems*, vol. 21, pp. 133-154, 2003.

[17] J. L. Hutchens and M. D. Alder, "Introducing MegaHAL," presented at Proceedings of the Human-Computer Communication Workshop, pp. 271-274, 1998.

[18] J. S. Breese and D. Heckerman, "Decision-theoretic case-based reasoning," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 26, pp. 838-842, 1996.

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

[20] J. Hammerton, M. Osbourne, S. Armstrong, and W. Daelemans, "Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing," *Journal of Machine Learning Research*, vol. 2, pp. 551-558, 2002.

[21] J. Jia, "The Study of the Application of a Keywords-based Chatbot System on the Teaching of Foreign Languages," University of Augsburg, Augsburg, Germany 2002.

[22] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Televisions and New Media Like Real People and Places*: Cambridge University Press, 1996.

[23] A. M. Pejtersen, "Semantic Information Retrieval," *Communications of the ACM*, vol. 41, pp. 90-92, 1998.

[24] A. De Angeli, G. I. Johnson, and L. Coventry, "The unfriendly user: exploring social reactions to chatterbots," presented at Proceedings of The International Conference on Affective Human Factors Design, London, pp. 467-474, 2001.

[25] S. Han and Y. Kim, "Intelligent Dialogue System for Plane Euclidean Geometry Learning," presented at International Conference on Computers in Education, Seoul, Korea, 2001.

[26] R. P. Schumaker and H. Chen, "Leveraging Question Answer Technology to Address Terrorism Inquiry," *Decision Support Systems*, vol. 43, pp. 1419-1430, 2007.

[27] M. Louwerse, A. Graesser, and A. Olney, "Good Computational Manners: Mixed-Initiative Dialog in Conversational Agents," in *Papers from the 2002 Fall Symposium, Technical Report FS-02-02*, E. f. H.-C. W. C. Miller, Ed. Menlo Park, CA: AAAI Press, 2002, pp. 71-76.

[28] A. C. Graesser, N. K. Person, and D. Harter, "Teaching Tactics and Dialog in AutoTutor," *International Journal of Artificial Intelligence in Education*, vol. 12, pp. 257-279, 2001.

[29] E. M. Voorhees, "Overview of the TREC 2001 Question Answering Track," presented at Text REtrieval Conference, pp. 42-51, 2001.

[30] E. M. Voorhees, "Overview of the TREC 2003 Question Answering Track," presented at Text REtrieval Conference, pp. 2003.

[31] I. Zuckerman and E. Horvitz, "Using Machine Learning Techniques to Interpret WH-questions," presented at Association for Computational Linguistics, Toulouse, France, pp. 547-554, 2001.

[32] O. Uzuner, R. Davis, and B. Katz, "Using empirical methods for evaluating expression and content similarity," presented at Proceedings of the 37th Annual Hawaii International Conference on System Sciences, pp. 1-8, 2004.

[33] A. F. Bissell, "Weighted Cumulative Sums for Test Analysis Using Word Counts," *Journal of the Royal Statistical Society. Series A*, vol. 158, pp. 525-545, 1995.