Evaluating Mass Knowledge Acquisition using the ALICE Chatterbot: The AZ-ALICE Dialog System

Robert P. 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, yingliu, mginsbur, hchen@eller.arizona.edu}

Word Count: 5824

Abstract

In this paper, we evaluate mass knowledge acquisition using modified ALICE chatterbots. In particular we investigate the potential of allowing subjects to modify chatterbot responses to see if distributed learning from a web environment can succeed. This experiment looks at dividing knowledge into general conversation and domain specific categories for which we have selected telecommunications. It was found that subject participation in knowledge acquisition can contribute a significant improvement to both the conversational and telecommunications knowledge bases. We further found that participants were more satisfied with telecommunications responses rather than general conversation.

1 Introduction

The process of knowledge acquisition is to transfer existing knowledge and its structure into a computer-interpretable form (Potter 2001). This knowledge can come from humans or other sources such as textual documents or encyclopedias. When coupled with the Internet, knowledge acquisition inherits new problems of scale such as information quality and reliability issues.

This paper investigates the knowledge acquisition activities of a chatterbot program that mimics human conversation. Web-based chatterbot systems can provide an easy, natural extension to knowledge acquisition. This style of dialog system, due to its robustness, scalability, and ease of connecting to the web for information retrieval, appears to be a viable approach for knowledge acquisition. One of the better performers in the field is the ALICEbot. ALICE, or Artificial Linguistic Internet Chat Entity, was developed by Richard Wallace in 1995. This system has had marked success, winning the Loebner Prize for most human-like computer in 2000, 2001, and 2004.

In this paper, we will investigate the existing literature in Section 2, going from knowledge acquisition and its many approaches, down to the ALICE chatterbots and how they fit into the framework. In Section 3, we introduce a set of research questions and offer possible hypotheses. In Section 4, we explain the system design implemented in our study. Section 5 looks at the experimental design in detail. Section 6 describes the results from the experiment and offers a discussion of their meaning. Finally, in Section 7, the conclusions and future directions are provided.

2 Literature Review

Knowledge acquisition has been a sought after goal since the early days of Artificial Intelligence. Newell posited that psychology and structure are important elements to perform a sequence of complex tasks, and noted the similarities between cognitive tasks and existing programming languages that are engineered to use logic and conditional operators (Newell 1973) to mimic human ability and to simulate human behavior (Feigenbaum and Simon 1962).

Under the broad umbrella of knowledge acquisition there are several approaches; those that are manual, and those that are automated or semi-automated. In the manual approaches, (Potter 2001) describes several of the more common ones; the expert interview where knowledge is captured from a human expert through an interview-style process, and the protocol analysis where an expert's thought process is analyzed step-by-step in solving a particular problem. The strengths of using these processes rest in the quality of information gained. However, the process of acquiring knowledge using these methods is laborious and time-consuming. The automatic or semi-automatic methods to knowledge acquisition attempt to automate all or some part of the acquisition process in a human-like fashion using models or examples to learn from (Lavrac and Mozetic 1992; Potter 2001). The draw of using such an approach is to decrease the amount of time needed to capture knowledge. These approaches work best when dealing with highly structured knowledge. But, this approach does not generally increase information quality and it is difficult to qualify what a system 'should have' learned.

2.1 Approaches to Knowledge Acquisition

The automated or semi-automated approaches have several sub-approaches; pattern and template matching, machine learning, and text data mining (Lavrac and Mozetic 1992; Potter 2001). In pattern and template-matching, the natural language input is matched against some preconceived template where the response is pre-coded and dictated by the matching template. This type of system is simple, scalable, easy to develop, and has a good performance history. The drawback to this approach is the amount of time required to acquire initial knowledge and place in a structured template. For machine learning, where the system attempts to identify concept pairs or co-occurrences, the system can handle exceedingly large data sets. However, this system requires a priori knowledge of the domain and the input must be in a highly structured and machine-friendly style. In text data mining, new knowledge can be discovered by the system by unearthing previously unknown relations using data mining techniques. However, the new knowledge in a conversational setting will generally require user guidance to format it in a semantically correct context.

2.2 Dialog Systems

For the pattern and template matching approach, there are several sub-categories as well. The first of which is the trigger approach where the system identifies trigger phrases within the natural language input. This method lends itself to a high degree of precision, however, it is not very robust and does not handle conflicts between triggers very well. The second method is that of the semantic approach. In this system, a context-free environment is maintained while focusing only on the word meanings themselves. The downside to this method is that it is not very scalable when pursuing new knowledge domains. In the third approach, syntactic analysis is used to acquire knowledge, typically from highly structured technical texts. This approach can induct new knowledge, but it will ignore unrecognized words or patterns. The fourth approach is that of the Eliza-type chatterbot, which played the role of a Rogerian psychotherapist to pseudo-extract information from its patients (Weizenbaum 1966). With Eliza, it became possible to carry on generalized conversations in a reasonable manner. The weakness fell on the exceedingly limited initial set of patterns that Eliza had at its disposal.

Following Eliza, a fundamental split of theory occurred when researchers began concentrating of two different methodologies; the theoretically motivated models and the performance led systems (Simmons 1970; Russell 2002). The theoretical models pursued directions in symbolic reasoning and deep understanding systems. An example of such research is Winograd's Blocks World where the artificial environment played a key role in shaping the program's direction. Performance led systems were quite the opposite and became increasingly more interested in obtaining an answer with speed than to distract themselves with system understanding. The favored approach is to omit syntactic analysis in favor of simple pattern matching tactics (Russell 2002; Vrajitoru 2003). Some of the more interesting research in this area came by way of Parry and the ALICEbot family.

Parry (Colby, Weber et al. 1971) was one of the chatterbot pioneers in the performance led systems and expanded upon the Eliza framework. Parry, a variant of Eliza, was a robust entity with many more patterns than its predecessor. However, Parry was restricted to a limited domain of paranoid hospital patient expressions. This under-generalized set of expressions were a limitation, however, in the context of insane hospital patient responses, those interacting with the system accepted Parry's responses even if the responses were non-transferable to a more generalized domain.

Another performance led system that investigated the aspects of social interaction was Cobot (Isbell, Kearns et al. 2000). Cobot, a part of the MUD community called LambdaMOO, interacts with others either by recognizing certain commands tied to its statistical information gathering (i.e., 'who loves me' and 'who acts like me') or by conversational pattern-matching techniques. This process of learning and interaction has allowed others to socially accept Cobot.

The ALICEbot family (Wallace 2004) is one of the more recent additions to the field, utilizing pattern-matching techniques without a syntactic overhead. These series of systems are considered by some researchers as robust, scalable, and convincing conversationalists. The downside is that expanding into new knowledge domains can require some laborious programming. Within the ALICEbot family as defined by (Russell 2002) is Converse, Cartman, and ALICE.

Converse, which won the Loebner Prize in 1997 for most human-like computer (Anonymous 2003), uses scripts and outside corpuses such as WordNet to find responses. The problem with this system is the need to write scripts to expand into new knowledge domains, as well as the reliance on a single exterior source of knowledge.

Cartman, the next notable member of the ALICEbot family, was developed as a dialog agent to tutor students. The benefits to this system are in its flexible, scalable nature, and the lack of client-side software to operate. The weakness, however, is that it was not developed as a general conversationalist and instead operates in particular domains of interest.

2.3 ALICE

ALICE uses XML knowledge bases to match user input against a predefined response set. The shortcoming of this system is that it cannot adequately answer all of the queries given to it. (Russell 2002) contends that ALICEbots have no cognitive theory behind them, instead they blindly rely on canned responses to matched inputs. (Wallace 2003) on the other hand argues that ALICEbots use Case Based Reasoning (CBR) to represent their responses. This becomes beneficial to a performance-led system because CBR does not require the computational overhead that other reason-based systems would demand (Breese and Heckerman 1996). Although ALICEbots are sometimes compared to rule-based procedures, (Gilboa and Schmeidler 2000) clarifies that rules can contradict one another while cases are repetitive and never contradictory.

ALICEbots are also able to expand their present knowledge bases through XML-based AIML (Artificial Intelligence Markup Language) (Wallace 2003). This would imply that ALICEbots could be given an 'expert appearance' within a particular domain of knowledge. This expansion has already been witnessed in the areas of foreign language fluency, and specific domain-related knowledge fields which can either be supervised by an interceding chatterbot master or unsupervised where knowledge is gathered en masse from trusted sources.

In the supervised approaches, there is a responsible party that filters and formats the knowledge into a machine-readable form. This can be a labor intensive task and is the default method of ALICEbot's learning. In unsupervised approaches, massive amounts of knowledge can be gathered in a relatively short period of time, however, (Wallace 2003) makes issue that users can sometimes be untrustworthy sources of knowledge.

Within the realm of unsupervised learning (Fu 1996) describes a system 'ART' which is an unsupervised pattern recognizer. Fu does concede that because of its unsupervised nature, ART can misclassify entire classes of instances. Another similar study in spoken dialog systems made use of a sub-verification routine (Smith 1998). Whenever the system is unsure about the spoken request, i.e., the calculated value of understanding falls below a certain threshold, the system will ask clarifying follow-up questions regarding the unclear statement in order to gain a more unambiguous understanding. It was found that such use of a sub-dialog verifier dramatically increased the overall system accuracy.

In studies conducted on ALICEbots in particular, some interesting results have been obtained. One study focused on using an ALICEbot as a Social Theory tutor for students (Moore and Gibbs 2002). It was discovered that students were more interested in using the system as a search engine to answer assignment questions rather than as the conversational tutor per its design. This sentence-based information retrieval aspect had traditionally been confined to the arena of search engine design (Radiv, Fan et al. 2005). In another learning-style experiment, a modified version of an ALICEbot was used as a learning tool to teach Chinese students either English or German (Jia 2002). This study focused more on user attitudes rather than on chatterbot efficiency. It was discovered that 62% of users chatted for 10 lines or less, and that 8.5% of the time ALICEbot had no specific pattern to match the given input and had to rely on root-level generic responses. These conversational entities all have in common the difficulty of maintaining dialog for a sustainable period of time (Zacharski 2004).

Throughout the examination of the field, several gaps in the research were noticed. The first such gap is that no ALICEbot harnesses a user-centric automated approach to knowledge acquisition. ALICEbots by default have relied solely on manual or semi-automatic methods for obtaining knowledge. The other notable find is that no one has examined the effectiveness of chatterbots in conversational or specific knowledge domains.

3 Research Questions

Dialog systems can function in one of two ways; they can provide brief, concise, or well-detailed answers to a particular query or they can engage the user in providing small talk types of conversational responses.

This leads to our exploration of mass knowledge acquisition where our aim is two-fold. First, we explore using human subjects to train various dialog systems and study the impact of the acquired knowledge. Second, we study the effects of domain answers to those of the conversational dialog. To accomplish this, we address the following questions:

- **How effective is interactive training of a dialog system?**
- **How well can a chatterbot perform in a specific domain compared to general conversation?**

To properly address the first question, we have a dual set of hypotheses.

- H1A: Mass knowledge acquisition will not make a noticeable improvement of responses for the general conversational chatterbot.
- H1B: Mass knowledge acquisition will improve the responses of the domain-specific chatterbot.

It is believed that because of the complexities involved and the limitless breadth of topics that can be covered in general conversation, we do not believe that we will find a noticeable difference. However, it is believed that for a constrained set of domain-specific knowledge, acquiring a set of additional knowledge will provide a noticeable improvement in chatterbot responses.

- H2: We expect that ALICEbots will be more effective in domain responses than in general dialog responses.

In H2, we further suspect that the responses from a constrained set of domain-specific knowledge will be more effective than those from general conversation.

4 System Design

To answer the questions posed, we constructed the AZ-ALICE dialog system. AZ-ALICE is built upon the freely available java-based ALICE Program D from www.ALICEbot.org. Our system can be broken into five component parts; the Chat User Interface, Chat Engine, AIML (Artificial Intelligence Markup Language) knowledge files, a Logging component, and Evaluation module.

The Chat User Interface is an XML-based web page that allows users to authenticate themselves and chat with the system. The system stores the authenticated name as a state variable which allows for personalized communication (e.g., What do you mean by that Sally369?).

The Chat Engine functions as the brain of the system, pattern-matching user input against the intelligence stored in the AIML files and returning the corresponding response.

The AIML files contain the knowledge of the system which is structured in XML form as patterns and templates. When a user inputs a query to the system, the Chat Engine seeks to best match the input against all of the loaded patterns and returns the associated template. Each of the pattern/template pairs are also known as categories.

```
<category>
<pattern>WHAT IS *</pattern>
<template>Do you think I should know that?</template>
</category>
```

The Logging component records the entire conversation history for each user. Finally, the Evaluation module permits the user to evaluate and offer a corrected system response. These evaluations are later analyzed for performance related data. Table 1 summarizes the differences between ALICE and the AZ-ALICE system.

	Chat Interface	Chat Engine	AIML	Logging	Evaluation
ALICE	Uses XML to chat with users	Uses off the shelf ALICE ProgramD	Uses the freely available Standard AIML set	Logs everything to a monolithic XML Log file	None
AZ-ALICE	Uses a customized perl skin for chat and evaluation purposes	Same as Original ALICE	Depends on the bot as to whether it is Standard AIML or customized Telecommunications knowledge	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 and AZ-ALICE

The Chat User Interface used in the AZ-ALICE system combined together the system dialog with the evaluation module. Users were allowed to immediately rate and suggest new chatterbot responses following each line of inputted chat. Figure 1 shows a screenshot of the AZ-ALICE in evaluation mode.

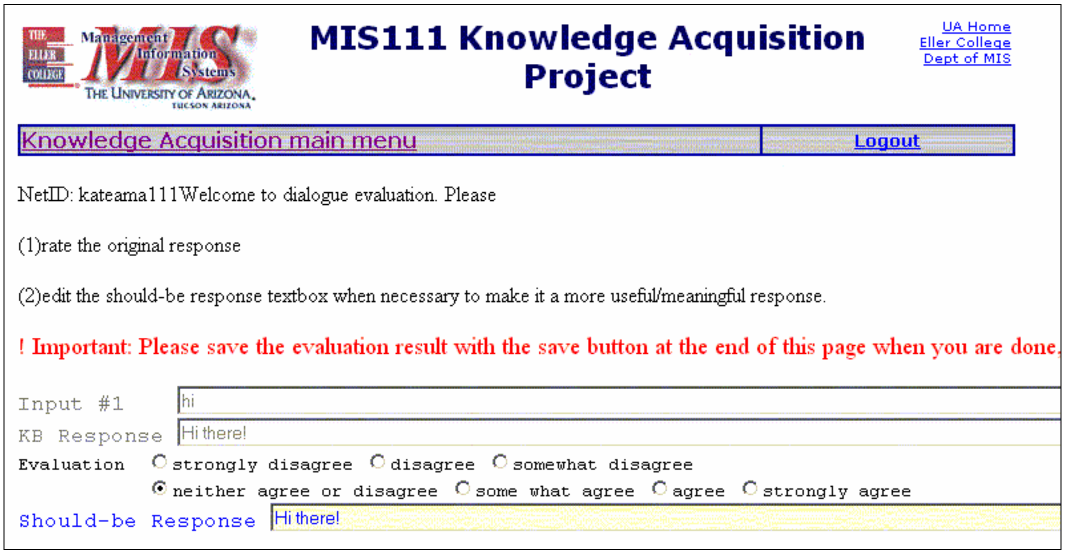

Figure 1. Screenshot of AZ-ALICE evaluation

To illustrate how the AZ-ALICE system works; consider the following input/response interaction.

```
User: Tell me a funny joke.
AZ-ALICE: Did you hear the one about the Mountain Goats in the Andes?
   It was Ba a a a a a d.
User: Tell me another one.
AZ-ALICE: I am the latest result in artificial intelligence which can
   reproduce the capabilities of the human brain with greater speed and
   accuracy.
```

- AZ-ALICE passes the user input "Tell me a funny joke" to the chat engine.
- The chat engine uses a memory-resident directed graph to best match the user input. In the case of the first user input, the best match is "TELL ME A JOKE".
- The chat engine returns the corresponding response to the input, "Did you hear…"

The second user input/response interaction demonstrates the shallow behavior of the ALICE chat engine. In the second user input the context is still about funny jokes, however, ALICE treats this second input as a completely new topical interaction. This lack of context returns the best match to the input without consideration to the topic. The chat engine behavior does at times become problematic when attempting to maintain a conversation for any length of time.

```
User: ok well i think i might go now
AZ-ALICE: Oh. How about that. Since when? Perhaps. How do you plan to
   do this?
User: i'll logoff
AZ-ALICE: Do you think your plan will succeed?
User: yes its not that difficult
AZ-ALICE: That was not a yes or no question. I think you mean 'it's' or
   'it is' not 'its'. What is it?
```

However, the ALICE chatterbot is able to return adequate replies to brief user inquiries. The following interaction demonstrates this behavior.

```
User: I went on a big date last night.
AZ-ALICE: Was it very expensive? Did you have a good time?
User: It was very expensive. I had a great time.
AZ-ALICE: How did you feel about that? What happened to it?
User: I had the salmon.
AZ-ALICE: Do you still have the salmon ?
User: I ate it yesterday.
AZ-ALICE: How did it taste?
User: It tastes very good.
AZ-ALICE: Like chicken?
User: Not like chicken but chicken is good too.
AZ-ALICE: Oh I get it.
User: Do you like chicken?
AZ-ALICE: Yes it is good especially baked.
User: How can you eat chicken if you are a robot?
AZ-ALICE: That is a good epistemological question.
```

In terms of telecommunications knowledge, ALICE exhibited some of the same shortcomings of keeping conversations going for any length of time.

```
User: Do you know what a firewall is?
AZ-ALICE: A firewall is a specialized gateway designed to block or
   restrict certain types of data transfers on the Internet while
   permitting others.
User: Can you list different types of firewalls?
AZ-ALICE: You should just download my software.
User: What software?
AZ-ALICE: Interesting question.
User: How could I download your software?
AZ-ALICE: A lot of practise.
```

5 Experimental Design

5.1 The Experiment

In our experiment, we created two chatterbots, BaseBot and TeleComm. BaseBot, the general conversational chatterbot, was our control chatterbot. It was an off-the-shelf ALICE ProgramD chatterbot loaded with the 'Standard AIML' rule set consisting of 23,735 knowledge categories that can be freely obtained from www.alicebot.org. Each of the knowledge categories consists of a pattern to match against the user input and a template response corresponding to the pattern. The other chatterbot, TeleComm, was essentially identical to BaseBot except that TeleComm was further augmented with 298 telecommunication specific definitions thus raising TeleComm's total rule set to 24,032[1].

Our experiment was further divided into two separate user studies utilizing both chatterbots. In User Study 1, subjects were encouraged to imprint new knowledge patterns onto their particular chatterbot to offset incorrect responses to their queries. In User Study 2, the new knowledge gathered from each chatterbot was then integrated into the preexisting knowledge stores of BaseBot' and TeleComm-2 respectively. Subjects in the second user study were given the same instructions as those of the first.

5.2 Participants

Participant subjects came from four sections of a freshman introductory course in Management of Information Systems. Each class section was given a particular chatterbot. Participation was on a voluntary basis, however, participants who successfully completed the requirements were given bonus point incentives. Table 2 shows the breakdown of participation

[1] One category pattern overlapped between 'Standard AIML' and Telecommunications definitions, decreasing the total by one.

among the experiments. In total, 376 participants were involved in our study. We believe that this is probably the largest systematic study involving a chatterbot.

Number of Study Participants		
Chatterbot	User Study 1	User Study 2
BaseBot	74	98
TeleComm	91	113

Table 2: Participant Breakdown

Participants were asked to interact with the chatterbot for approximately one-half hour. At the end of the interaction period, users were presented with all of their user input and chatterbot responses. Participants were then given the opportunity to correct chatterbot responses and rate their satisfaction level with the response using a one-to-seven Likert scale (one – strongly dissatisfied to seven – strongly satisfied). Subjects were further asked to constrain their chatting to telecommunication topics; however, participants were not forced to do so. This gentle absence of restricting topics further allows us to evaluate Han's claim that the inclusion of general conversational knowledge can help users to self-steer their conversation back to a domain-specific target (Han and Kim 2001).

While we can agree that in this experiment alone where users were provided with an incentive of receiving a meager amount of classroom bonus points, transferring this technology 'to the wild' would derive many of the same benefits as mentioned with open-source software. The best comparison to an existing knowledge-based system would be Wikipedia, warts and all. Users derive no more benefit from adding and correcting the knowledge in Wikipedia, other than self-satisfaction and use of the tool for areas in which they may not as knowledgeable (Wagner 2004). Applying this parallel to a real-world instance of our application, users would be creating and improving knowledge in much the same way as Wikipedia, and for the same reasons.

5.3 Performance Metrics

The first metric that we aim to study is that of Correction Rate. Since users are permitted to correct chatterbot responses, Correction Rate is defined as a percentage of the number of new responses entered into the system divided by the total number of user inputs. This metric is used to evaluate each chatterbot to determine whether the Correction Rates are increasing, decreasing, or remaining steady after mass knowledge acquisition.

The second metric is Response Satisfaction. Response Satisfaction is measured for each user input/response pair on a Likert scale of one to seven. This measure is then aggregated for each chatterbot and is simply the summation of all Response Satisfaction scores for the chatterbot divided by the total number of user inputs.

5.4 Experiment setup

Testing of H1A and H1B, mass knowledge acquisition effectiveness of the BaseBot and TeleComm chatterbot respectively, is carried out between user studies. Users from the first user study whom have provided new responses for BaseBot and TeleComm will have those new responses incorporated into the second user study chatterbots, BaseBot-2 and TeleComm-2. The change in Correction Rate and Response Satisfaction between both BaseBot-1 and BaseBot-2 and TeleComm-1 and TeleComm-2 is then measured.

In testing hypothesis H2 (domain responses will be more effective than dialog), the Response Satisfaction measures of the BaseBot and TeleComm chatterbots are broken down into constituent dialog and domain knowledge responses and calculated against one another. Figure 2 graphically demonstrates the experimental relations.

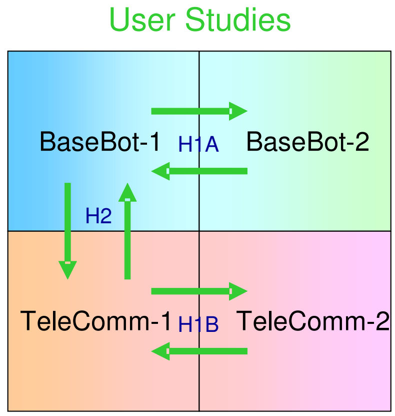

Figure 2 – Experimental Design Testing

6 Experimental Results and Discussion

6.1 Number of Corrections Made

In our investigation of mass knowledge acquisition, users were instructed to make corrections to the knowledge sets of Study 1 that would then be incorporated into the knowledge sets used by Study 2. The following user/chatterbot interaction followed this pattern:

```
User in Study 1: What are you talking about?
AZ-ALICE: The topic is mood are you in.
User's new suggested response: My bad.

User in Study 2: What are you talking about?
AZ-ALICE: My bad.
```

From this acquisition interaction, users entered 1,707 new responses (17.5% of BaseBot-1 responses) for BaseBot. While the TeleComm-1 chatterbot received slightly fewer new responses, 1,632 (16.0% of TeleComm-1 responses). These results are summarized in Table 3.

Stats on Overall Use	User Study 1		User Study 2	
	BaseBot-1	TeleComm-1	BaseBot-2	TeleComm-2
Total Number of Users	74	91	98	113
Lines of User Input	9,751	10,179	14,484	13,641
New Responses Entered	1,707	1,632	2,722	2,430
Correction Rate	17.5%	16.0%	18.8%	17.8%
Corrections per User	23	18	28	22
Number of New Responses used by the chatterbot			156	101
Percentage used			1.1%	0.7%
How many were recorrected			41	31
Percentage recorrected			26.3%	30.7%

Table 3. Chatterbot User Corrections

From the bottom line of Table 3, approximately one-percent of the corrections made to chatterbot responses from Study 1 were used by the chatterbot in Study 2. Of these corrections, 26.3% of the BaseBot-2 New Responses and 30.7% of the TeleComm-2 New Responses were further corrected. This leads us to believe that with more subjects on the web, mass knowledge acquisition could be viable.

However, one problem we encountered was the addition of new suggested responses to already acceptable chatterbot expressions. To continue the previous interaction:

```
User in Study 1: What are you talking about?
AZ-ALICE: The topic is mood are you in.
User's new suggested response: My bad.

User in Study 2: What are you talking about?
AZ-ALICE: My bad.
User's suggested response: Nevermind.
```

In this example, the context of User 1's suggested response of "My bad" was not general enough to encompass the given input. This led to User 2 issuing a second correction of "Nevermind" to address the input. However, not all user interactions that fell into this category were because of inadequate response suggestions. Some of the input/response pairs came as a result of users wanting to provide more entertaining answers such as the following.

```
User in Study 1: What is your favorite movie?
AZ-ALICE: My favorite movie is undefined.  Have you seen it?
User’s suggested response: I don’t have a favorite movie.

User in Study 2: What is your favorite movie?
AZ-ALICE: I don’t have a favorite movie.
User’s suggested response: Star Trek.
```

In this exchange, the user in Study 2 offered a new chatterbot response although it was not really needed. This user behavior presents a problem. We had intended that Correction Rates be indicative of grammatical or communication-related problems with chatterbot responses where users would offer new responses for only those responses that needed it. But with participants instead elaborating on otherwise good responses, this may not be a limitation as much as expected student interaction with a system for a certain demographic subset and as such, this measure needs to be considered and interpreted cautiously.

6.2 Responses for General Conversational Chatterbot

In looking at the Response Satisfaction ratings for the general conversational chatterbot BaseBot, it would appear that mass knowledge acquisition improved. This improvement, 4.31 in the first user study to 4.33 in the second, was found to be statistically significant with a p-value < 0.001. This significance value is the result of the large number of interactions (n) with the system. Table 4 summarizes BaseBot Response Satisfaction and Correction Rates.

BaseBot stats on Satisfaction					
	n	Avg	Std Dev.		Corr. Rates
Study 1	9,751	4.31	1.50		17.5%
Study 2	14,484	4.33	1.76		18.8%

Table 4. BaseBot Satisfaction and Correction Rates

The increase in Response Satisfaction between user studies is likely to be attributed to the refinement of conversational chatterbot responses for Study 2 users. Returning to our hypothesis

H1A, we find that mass knowledge acquisition does in fact make a noticeable improvement of user satisfaction for the general conversational chatterbot.

6.3 Responses for Domain-specific Chatterbot

The TeleComm chatterbot is composed of both conversational as well as telecommunication knowledge. In looking at the Response Satisfaction ratings for the domain-specific knowledge, it would appear that mass knowledge acquisition improved its responses as well. This improvement from a mean of 4.64 in Study 1 to 4.95 in Study 2 was found to be statistically significant with a p-value < 0.001. Correction Rates decreased between studies (p-value < 0.001), which was in line with expectations. Table 5 summarizes TeleComm response satisfaction and correction rates.

TeleComm stats on Satisfaction				
	n	Avg	Std Dev.	Corr. Rates
Study 1	218	4.64	1.70	15.1%
Study 2	346	4.95	1.79	13.6%

Table 5. TeleComm telecommunications knowledge - Satisfaction and Correction Rates

This increase in response satisfaction and decrease in correction rates between user studies can be attributed to the refinement of telecommunications knowledge similar to that observed in the dialog system. It is interesting to note that the magnitude of mean movement between the two studies is much greater than that of the dialog system. This could be due to the limited amount of initial telecommunications knowledge used in Study 1 as compared to the much larger array of conversational knowledge, which made more of an impact on the telecommunications mass knowledge acquisition.

From hypothesis H1B, we had expected that mass knowledge acquisition would improve the domain-specific chatterbot responses. From the analysis above, we find that mass knowledge acquisition improves domain-specific chatterbot responses.

6.4 Comparing Dialog to Domain

To answer our hypothesis H2, that domain-specific chatterbot responses will rate better than general conversation, we examine the chat output from both BaseBot and TeleComm in User Study 1 by breaking the chat output of each chatterbot into either conversational or domain-specific chat. Table 6 illustrates such a breakdown of response satisfaction and correction rates between the chatterbots.

	User Study 1	
BaseBot	Conversational Dialog	TeleComm Domain
Lines of User Input	9,670	81
Response Satisfaction Avg	4.30	4.43
Response Satisfaction Std Dev.	1.50	1.66
Correction Rate	17.4%	28.4%
TeleComm		
Lines of User Input	9,961	218
Response Satisfaction Avg	4.16	4.64
Response Satisfaction Std Dev.	1.50	1.70
Correction Rate	16.1%	15.1%

Table 6. Comparing dialog to domain for the chatterbots

Looking at the BaseBot results, the mean response satisfaction value is higher when the BaseBot is talking about Telecommunications than general conversation (4.43 versus 4.30, p-value < 0.001). Even though the BaseBot was not loaded with a specific Telecommunications knowledge set, BaseBot did have some familiarity with some of the telecommunications terms. From the data, users rated this subset of dialog knowledge better than the rest of the set. The following input/response interaction illustrates the argument.

```
User: What is an IP address?
AZ-ALICE: The number of a specific machine like your 127.0.0.1.
```

The same conclusion can be drawn about the TeleComm chatterbot. Its mean response satisfaction value is higher for telecommunications than general conversation (4.64 versus 4.16, p-value < 0.001). This result meets our expectation that the domain knowledge is better than the conversational knowledge. Therefore, we found that domain responses are clearly more effective than dialog.

7 Conclusions and Future Directions

From our study we can conclude that the use of a chatterbot as a knowledge acquisition tool appears to be a stable instrument in gathering both conversation and domain-related knowledge. We believe that with the decrease in correction rates observed between studies, that after several rounds of such corrections, that the knowledge base will be of sufficient quality to answer domain-related questions. Extending this research in such a way demonstrates the viability of having users train a conversational system. Although a minority of subjects did intentionally leave misleading or false responses, we feel that the iterative process will eventually eliminate such responses. Furthermore, we found that mass knowledge acquisition restricted to a particular knowledge domain had higher Response Satisfaction levels than the corresponding conversational-style responses regardless of the chatterbot involved.

From our examination of natural language dialog systems and ALICEbot in particular, we believe that ALICEbots show a promising future in domain-restricted areas. Although we studied only one particular area of domain expertise in telecommunications, it would be interesting to further pursue other areas of domain interest, and test the flexibility of the ALICEbot system in those areas as well.

However, before concluding this analysis the reader should be aware of some caveats specific to this study. The first of which dealt with Correction Rates. It was hoped that Correction Rates would only be issued for malformed or poorly performed chatterbot responses. Instead, users manipulated the ability to make corrections for social or entertainment value. The second caveat deals with the number of telecommunications terms used in the study. It is recognized that there were a limited number of domain-specific knowledge categories loaded into the TeleComm chatterbot. Although the number may have been quite sparse when compared against the number of conversational dialog categories, the results that arose from this Spartan set speak for themselves. In further studies it would be interesting to explore the use of larger domain sets to determine if similar results can be obtained. With even more subjects on the Web, mass knowledge acquisition could be feasible.

8 Acknowledgements

This work was supported in part by the NSF, ITR: "COPLINK Center for Intelligence and Security Informatics Research" Sept. 1, 2003 – Aug. 31, 2005.

9 References

Anonymous (2003). Home Page of The Loebner Prize--"The First Turing Test". 2004.

Breese, J. S. and D. Heckerman (1996). "Decision-theoretic case-based reasoning." Systems, Man and Cybernetics, Part A, IEEE Transactions on 26(6): 838-842.

Colby, K. M., S. Weber, et al. (1971). "Artificial Paranoia." Artificial Intelligence 2: 1-25.

Feigenbaum, E. A. and H. A. Simon (1962). "Simulation of Human Verbal Learning Behavior." Communications of the ACM 5(4): 223.

Fu, L. (1996). "Incremental knowledge acquisition in supervised learning networks." Systems, Man and Cybernetics, Part A, IEEE Transactions on 26(6): 801-809.

Gilboa, I. and D. Schmeidler (2000). "Case-based knowledge and induction." Systems, Man and Cybernetics, Part A, IEEE Transactions on 30(2): 85-95.

Han, S. and Y. Kim (2001). Intelligent Dialogue System for Plane Euclidean Geometry Learning. International Conference on Computers in Education, Seoul, Korea.

Isbell, C. L., M. Kearns, et al. (2000). Cobot in LambdaMOO: A Social Statistics Agent. American Association for Artificial Intelligence, Austin, TX.

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

Lavrac, N. and I. Mozetic, Eds. (1992). Second generation knowledge acquisition methods and their application to medicine. Deep Models for Medical Knowledge Engineering. Elsevier, New York.

Moore, R. and G. Gibbs (2002). Emile: Using a chatbot conversation to enhance the learning of social theory. Huddersfield, England, Univ. of Huddersfield, (unpublished report obtained from author in 2003)

Newell, A. (1973). You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of this Symposium, W.G. Chase, Academic Press.

Potter, S. (2001). A Survey of Knowledge Acquisition from Natural Language. TMA of Knowledge Acquisition from Natural Language. Edinburgh. 2003: http://www.aiai.ed.ac.uk/project/akt/work/stephenp/TMA%20of%20KAfromNL.pdf.

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.

Russell, R. S. (2002). Language Use, Personality and True Conversational Interfaces. Edinburgh, Univ of Edinburgh.

Simmons, R. F. (1970). "Natural Language Question Answering Systems: 1969." Communications of the ACM 13(1): 15-30.

Smith, R. W. (1998). "An evaluation of strategies for selectively verifying utterance meanings in spoken natural language dialog." International Journal of Human-Computer Studies 48(5): 627-647.

Vrajitoru, D. (2003). Evolutionary Sentence Building for Chatterbots. Genetic and Evolutionary Computation Conference (GECCO), Chicago, IL.

Wagner, C. (2004). "WIKI: A Technology for Conversational Knowledge Management and Group Collaboration." Communications of the AIS 13: 265-289.

Wallace, R. S. (2003). The Elements of AIML Style. A.L.I.C.E. Artificial Intelligence Foundation, Inc.

Wallace, R. S. (2004). The Anatomy of A.L.I.C.E. A.L.I.C.E. Artificial Intelligence Foundation, Inc.

Weizenbaum, J. (1966). "Eliza - a computer program for the study of natural language communication between man and machine." Communications of the ACM 9(1): 36-45.

Zacharski, R. (2004). A Discourse System for Conversational Characters. Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.