# Building Effective Question Answering Characters

**Anton Leuski** and **Ronakkumar Patel** and **David Traum**
Institute for Creative Technologies
University of Southern California
Marina del Rey, CA, 90292, USA
leuski,ronakkup,traum@ict.usc.edu

**Brandon Kennedy**
Brandon.Kennedy@usma.edu

## Abstract

In this paper, we describe methods for building and evaluation of limited domain question-answering characters. Several classification techniques are tested, including text classification using support vector machines, language-model based retrieval, and cross-language information retrieval techniques, with the latter having the highest success rate. We also evaluated the effect of speech recognition errors on performance with users, finding that retrieval is robust until recognition reaches over 50% WER.

## 1 Introduction

In the recent Hollywood movie "iRobot" set in 2035 the main character played by Will Smith is running an investigation into the death of an old friend. The detective finds a small device that projects a holographic image of the deceased. The device delivers a recorded message and responds to questions by playing back prerecorded answers. We are developing virtual characters with similar capabilities.

Our target applications for these virtual characters are training, education, and entertainment. For use in education, such a character should be able to deliver a message to the student on a specific topic. It also should be able to support a basic spoken dialog on the subject of the message, e.g., answer questions about the message topic and give additional explanations. For example, consider a student learning about an event in a virtual world. Lets say there is a small circus in a small town and someone has released all the animals from circus. A young student plays a role of a reporter to find out who caused this local havoc. She is out to interrogate a number of witnesses represented by the virtual characters. It is reasonable to expect that each conversation is going to be focused solely on the event of interest and the characters may refuse to talk about anything else. Each witness may have a particular and very narrow view into an aspect of the event, and the student's success would depend on what sort of questions she asks and to which character she addresses them.

Automatic question answering (QA) has been studied extensively in recent years. For example, there is a significant body of research done in the context of the QA track at the Text REtrieval Conference (TREC) (Voorhees, 2003). In contrast to the TREC scenario where both questions and answers are based on facts and the goal is to provide the most *relevant* answer, we focus the answer's *appropriateness*. In our example about an investigation, an evasive, misleading, or an "honestly" wrong answer from a witness character would be appropriate but might not be relevant. We try to highlight that distinction by talking about QA *characters* as opposed to QA systems or agents.

We expect that a typical simulation would contain quite a few QA characters. We also expect those characters to have a natural spoken language interaction with the student. Our technical requirements for such a QA character is that it should be able to understand spoken language. It should be robust to disfluencies in conversational English. It should be relatively fast, easy, and inexpensive to construct without the need for extensive domain knowledge and dialog management design expertise.

In this paper we describe a QA character by the name of *SGT Blackwell* who was originally designed to serve as an information kiosk at an army



*Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27,
Sydney, July 2006. 

conference (see Appendix C for a photograph of the system) (**?**). We have used SGT Blackwell to develop our technology for automatic answer selection, conversation management, and system integration. We are presently using this technology to create other QA characters.

In the next section we outline the SGT Blackwell system setup. In Section 3 we discuss the answer selection problem and consider three different algorithms: Support Vector Machines classifier (SVM), Language Model retrieval (LM), and Cross-lingual Language Model (CLM) retrieval. We present the results of off-line experiments showing that the CLM method performs significantly better than the other two techniques in Section 4. Section 5 describes a user study of the system that uses the CLM approach for answer selection. Our results show that the approach is very robust to deviations in wording from expected answers, and speech recognition errors. Finally, we summarize our results and outline some directions for future work in Section 6.

## 2 SGT Blackwell

A user talks to SGT Blackwell using a head-mounted close capture USB microphone. The user's speech is converted into text using an automatic speech recognition (ASR) system. We used the Sonic statistical speech recognition engine from the University of Colorado (Pellom, 2001) with acoustic and language models provided to us by our colleagues at the University of Southern California (Sethy et al., 2005). The answer selection module analyzes the speech recognition output and selects the appropriate response.

The character can deliver 83 spoken lines ranging from one word to a couple paragraphs long monologues. There are three kinds of lines SGT Blackwell can deliver: content, off-topic, and prompts. The 57 content-focused lines cover the identity of the character, its origin, its language and animation technology, its design goals, our university, the conference setup, and some miscellaneous topics, such as "what time is it?" and "where can I get my coffee?"

When SGT Blackwell detects a question that cannot be answered with one of the content-focused lines, it selects one out of 13 off-topic responses, (e.g., "I am not authorized to comment on that,") indicating that the user has ventured out of the allowed conversation domain. In the event that the user persists in asking the questions for which the character has no informative response, the system tries to nudge the user back into the conversation domain by suggesting a question for the user to ask: "You should ask me instead about my technology." There are 7 different prompts in the system.

One topic can be covered by multiple answers, so asking the same question again often results in a different response, introducing variety into the conversation. The user can specifically request alternative answers by asking something along the lines of "do you have anything to add?" or "anything else?" This is the first of two types command-like expressions SGT Blackwell understands. The second type is a direct request to repeat the previous response, e.g., "come again?" or "what was that?"

If the user persists on asking the same question over and over, the character might be forced to repeat its answer. It indicates that by preceding the answer with one of the four "pre-repeat" lines indicating that incoming response has been heard recently, e.g., "Let me say this again..."

## 3 Answer Selection

The main problem with answer selection is uncertainty. There are two sources of uncertainty in a spoken dialog system: the first is the complex nature of natural language (including ambiguity, vagueness, underspecification, indirect speech acts, etc.), making it difficult to compactly characterize the mapping from the text surface form to the meaning; and the second is the error-prone output from the speech recognition module. One possible approach to creating a language understanding system is to design a set of rules that select a response given an input text string (Weizenbaum, 1966). Because of uncertainty this approach can quickly become intractable for anything more than the most trivial tasks. An alternative is to create an automatic system that uses a set of training question-answer pairs to learn the appropriate question-answer matching algorithm (Chu-Carroll and Carpenter, 1999). We have tried three different methods for the latter approach, described in the rest of this section.

### 3.1 Text Classification

The answer selection problem can be viewed as a text classification task. We have a question text

as input and a finite set of answers, – classes, – we build a system that selects the most appropriate class or set of classes for the question. Text classification has been studied in Information Retrieval (IR) for several decades (Lewis et al., 1996). The distinct properties of our setup are (1) a very small size of the text, – the questions are very short, and (2) the large number of classes, e.g, 60 responses for SGT Blackwell.

An answer defines a class. The questions corresponding to the answer are represented as vectors of term features. We tokenized the questions and stemmed using the KStem algorithm (Krovetz, 1993). We used a $tf \times idf$ weighting scheme to assign values to the individual term features (Allan et al., 1998). Finally, we trained a multi-class Support Vector Machines ($SVM^{struct}$) classifier with an exponential kernel (Tsochantaridis et al., 2004). We have also experimented with linear kernel function, various parameter values for the exponential kernel, and different term weighting schemes. The reported combination of the kernel and weighting scheme showed the best classification performance. Such an approach is well-known in the community and has been shown to work very well in numerous applications (Leuski, 2004). In fact, SVM is generally considered to be one of the best performing methods for text classification. We believe it provides us with a very strong baseline.

### 3.2 Answer Retrieval

The answer selection problem can also be viewed as an information retrieval problem. We have a set of answers which we can call documents in accordance with the information retrieval terminology. Let the question be the query, we compare the query to each document in the collection and return the most appropriate set of documents.

Presently the best performing IR techniques are based on the concept of Language Modeling (Ponte and Croft, 1997). The main strategy is to view both a query and a document as samples from some probability distributions over the words in the vocabulary (i.e., language models) and compare those distributions. These probability distributions rarely can be computed directly. The "art" of the field is to estimate the language models as accurately as possible given observed queries and documents.

Let $Q = q_1...q_m$ be the question that is received by the system, $R_Q$ is the set of all the answers appropriate to that question, and $P(w|R_Q)$ is the probability that a word randomly sampled from an appropriate answer would be the word $w$. The language model of $Q$ is the set of probabilities $P(w|R_Q)$ for every word in the vocabulary. If we knew the answer set for that question, we can easily estimate the model. Unfortunately, we only know the question and not the answer set $R_Q$. We approximate the language model with the conditional distribution:

$$P(w|R_Q) \approx P(w|Q) = \frac{P(w, q_1, ..., q_m)}{P(q_1, ..., q_m)} \quad (1)$$

The next step is to calculate the joint probability of observing a string: $P(W) = P(w_1, ..., w_n)$. Different methods for estimating $P(W)$ have been suggested starting with simple unigram approach where the occurrences of individual words are assumed independent from each other: $P(W) = \prod_{i=1}^{n} P(w_i)$. Other approaches include Probabilistic Latent Semantic Indexing (PLSI) (Hoffman, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The main goal of these different estimations is to model the interdependencies that exist in the text and make the estimation feasible given the finite amount of training data.

In this paper we adapt an approach suggested by Lavrenko (Lavrenko, 2004). He assumed that all the word dependencies are defined by a vector of possibly unknown parameters on the language model. Using the de Finetti's representation theorem and kernel-based probability estimations, he derived the following estimate for the query language model:

$$P(w|Q) = \frac{\sum_{s \in S} \pi_s(w) \prod_{i=1}^{m} \pi_s(q_i)}{\sum_s \prod_{i=1}^{m} \pi_s(q_i)} \quad (2)$$

Here we sum over all training strings $s \in S$, where $S$ is the set of training strings. $\pi_s(w)$ is the probability of observing word $w$ in the string $s$, which can be estimated directly from the training data. Generally the unigram maximum likelihood estimator is used with some smoothing factor:

$$\pi_s(w) = \lambda_\pi \cdot \frac{\#(w, s)}{|s|} + (1 - \lambda_\pi) \cdot \frac{\sum_s \#(w, s)}{\sum_s |s|} \quad (3)$$

where $\#(w, s)$ is the number of times word $w$ appears in string $s$, $|s|$ is the length of the string $s$, we sum over all training strings $s \in S$, and the constant $\lambda_\pi$ is the tunable parameter that can be determined from training data.

We know all the possible answers, so the answer language model $P(w|A)$ can be estimated from the data:

$$P(w|A) = \pi_A(w) \tag{4}$$

### 3.3 Ranking criteria

To compare two language models we use the Kullback-Leibler divergence $D(p_q||p_a)$ defined as

$$D(p_q||p_a) = \sum_{w \in V} P(w|Q) \log \frac{P(w|Q)}{P(w|A)} \tag{5}$$

which can be interpreted as the relative entropy between two distributions. Note that the Kullback-Leibler divergence is a dissimilarity measure, we use $-D(p_q||p_a)$ to rank the answers.

So far we have assumed that both questions and answers use the same vocabulary and have the same a priori language models. Clearly, it is not the case. For example, consider the following exchange: "what happened here?" – "well, maam, someone released the animals this morning." While the answer is likely to be very appropriate to the question, there is no word overlap between these sentences. This is an example of what is known in information retrieval as vocabulary mismatch between the query and the documents. In a typical retrieval scenario a query is assumed to look like a part of a document. We cannot make the same assumption about the questions because of the language rules: e.g., "what", "where", and "why" are likely to appear much more often in questions than in answers. Additionally, a typical document is much larger than any of our answers and has a higher probability to have words in common with the query. Finally, a typical retrieval scenario is totally context-free and a user is encouraged to specify her information need as accurately as possible. In a dialog, a portion of the information is assumed to be well-known to the participants and remains un-verbalized leading to sometimes brief questions and answers.

We believe this vocabulary mismatch to be so significant that we view the participants as speaking two different "languages": a language of questions and a language of answers. We will model the problem as a cross-lingual information task, where one has a query in one language and wishes to retrieve documents in another language. There are two ways we can solve it: we can translate the answers into the question language by building a representation for each answer using the question vocabulary or we can build question representations in the answer language.

### 3.4 Question domain

We create an answer representation in the question vocabulary by merging together all the training questions that are associated with the answer into one string: a pseudo-answer. We use equations 5, 2, 3, and 4 to compare and rank the pseudo-answers. Note that in equation 2 $s$ iterates over the set of all pseudo-answers.

### 3.5 Answer domain

Let us look at the question language model $P(w|Q)$ again, but now we will take into account that $w$ and $Q$ are from different vocabularies and have potentially different distributions:

$$P(w|Q) = \frac{\sum_s \alpha_{A_s}(w) \prod_{i=1}^m \pi_{Q_s}(q_i)}{\sum_s \prod_{i=1}^m \pi_{Q_s}(q_i)} \tag{6}$$

Here $s$ iterates over the training set of question-answer pairs $\{Q_s, A_s\}$ and $\alpha_x(w)$ is the experimental probability distribution on the answer vocabulary given by the expression similar to equation 3:

$$\alpha_x(w) = \lambda_\alpha \frac{\#(w, x)}{|x|} + (1 - \lambda_\alpha) \frac{\sum_s \#(w, x)}{\sum_s |x|}$$

and the answer language model $P(w|A)$ can be estimated from the data as

$$P(w|A) = \alpha_A(w)$$

## 4 Algorithm comparison

We have a collection of questions for SGT Blackwell each linked to a set of appropriate responses. Our script writer defined the first question or two for each answer. We expanded the set by a) paraphrasing the initial questions and b) collecting questions from users by simulating the final system in a Wizard of Oz study (WOZ). There are 1,261 questions in the collection linked to 72 answers (57 content answers, 13 off-topic responses, and 2 command classes, see Section 2). For this

study we considered all our off-topic responses equally appropriate to an off-topic question and we collapsed all the corresponding responses into one class. Thus we have 60 response classes.

We divided our collection of questions into training and testing subsets following the 10-fold cross-validation schema. The SVM system was trained to classify test questions into one of the 60 classes.

Both retrieval techniques produce a ranked list of candidate answers ordered by the $-D(p_q||p_a)$ score. We only select the answers with scores that exceed a given threshold $-D(p_q||p_a) > \tau$. If the resulting answer set is empty we classify the question as off-topic, i.e., set the candidate answer set contains to an off-topic response. We determine the language model smoothing parameters $\lambda_s$ and the threshold $\tau$ on the training data.

We consider two statistics when measuring the performance of the classification. First, we measure its accuracy. For each test question the first response returned by the system, – the class from the SVM system or the top ranked candidate answer returned by either LM or CLM methods, – is considered to be correct if there is link between the question and the response. The accuracy is the proportion of correctly answered questions among all test questions.

The second statistic is precision. Both LM and CLM methods may return several candidate answers ranked by their scores. That way a user will get a different response if she repeats the question. For example, consider a scenario where the first response is incorrect. The user repeats her question and the system returns a correct response creating the impression that the QA character simply did not hear the user correctly the first time. We want to measure the quality of the ranked list of candidate answers or the proportion of appropriate answers among all the candidate answers, but we should also prefer the candidate sets that list all the correct answers before all the incorrect ones. A well-known IR technique is to compute average precision – for each position in the ranked list compute the proportion of correct answers among all preceding answers and average those values.

Table 1 shows the accuracy and average precision numbers for three answer selection methods on the SGT Blackwell data set. We observe a significant improvement in accuracy in the retrieval methods over the SVM technique. The differences shown are statistical significant by t-test with the cutoff set to 5% ($p < 0.05$).

We repeated out experiments on QA characters we are developing for another project. There we have 7 different characters with various number of responses. The primary difference with the SGT Blackwell data is that in the new scenario each question is assigned to one and only one answer. Table 2 shows the accuracy numbers for the answer selection techniques on those data sets. These performance numbers are generally lower than the corresponding numbers on the SGT Blackwell collection. We have not yet collected as many training questions as for SGT Blackwell. We observe that the retrieval approaches are more successful for problems with more answer classes and more training data. The table shows the percent improvement in classification accuracy for each LM-based approach over the SVM baseline. The asterisks indicate statistical significance using a t-test with the cutoff set to 5% ($p < 0.05$).

## 5 Effect of ASR

In the second set of experiments for this paper we studied the question of how robust the CLM answer selection technique in the SGT Blackwell system is to the disfluencies of normal conversational speech and errors of the speech recognition. We conducted a user study with people interviewing SGT Blackwell and analyzed the results. Because the original system was meant for one of three demo "reporters" to ask SGT Blackwell questions, specialized acoustic models were used to ensure the highest accuracy for these three (male) speakers. Consequently, for other speakers (especially female speakers), the error rate was much higher than for a standard recognizer. This allowed us to calculate the role of a variety of speech error rates on classifier performance.

For this experiment, we recruited 20 participants (14 male, 6 female, ages from 20 to 62) from our organization who were not members of this project. All participants spoke English fluently, however the range of their birth languages included English, Hindi, and Chinese.

After filling out a consent form, participants were "introduced" to SGT Blackwell, and demonstrated the proper technique for asking him questions (i.e., when and how to activate the microphone and how to adjust the microphone position.) Next, the participants were given a scenario

| SVM | LM | | | CLM | | |
|---|---|---|---|---|---|---|
| accuracy | accuracy | impr. SVM | avg. prec. | accuracy | impr. SVM | avg. prec. |
| 53.13 | 57.80 | 8.78 | 63.88 | 61.99 | 16.67 | 65.24 |

Table 1: Comparison of three different algorithms for answer selection on SGT Blackwell data. Each performance number is given in percentages.

| | number of questions | number of answers | SVM accuracy | LM | | CLM | |
|---|---|---|---|---|---|---|---|
| | | | | accuracy | impr. SVM | accuracy | impr. SVM |
| 1 | 238 | 22 | 44.12 | 47.06 | 6.67* | 47.90 | 8.57* |
| 2 | 120 | 15 | 63.33 | 62.50 | -1.32 | 64.17 | 1.32 |
| 3 | 150 | 23 | 42.67 | 44.00 | 3.12* | 50.00 | 17.19* |
| 4 | 108 | 18 | 42.59 | 44.44 | 4.35* | 50.00 | 17.39* |
| 5 | 149 | 33 | 32.21 | 41.35 | 28.37* | 42.86 | 33.04* |
| 6 | 39 | 8 | 69.23 | 58.97 | -14.81* | 66.67 | -3.70 |
| 7 | 135 | 31 | 42.96 | 44.19 | 2.85 | 50.39 | 17.28* |
| average | 134 | 21 | 48.16 | 48.93 | 1.60* | 53.14 | 10.34* |

Table 2: Comparison of three different algorithms for answer selection on 7 additional QA characters. The table shows the number of answers and the number of questions collected for each character. The accuracy and the improvement over the baseline numbers are given in percentages.

wherein the participant would act as a reporter about to interview SGT Blackwell. The participants were then given a list of 10 pre-designated questions to ask of SGT Blackwell. These questions were selected from the training data. They were then instructed to take a few minutes to write down an additional five questions to ask SGT Blackwell. Finally they were informed that after asking the fifteen written down questions, they would have to spontaneously generate and ask five additional questions for a total of 20 questions asked all together. Once the participants had written down their fifteen questions, they began the interview with SGT Blackwell. Upon the completion of the interview the participants were then asked a short series of survey questions by the experimenter about SGT Blackwell and the interview. Finally, participants were given an explanation of the study and then released. Voice recordings were made for each interview, as well as the raw data collected from the answer selection module and ASR. This is our first set of question answer pairs, we call it the ASR-QA set.

The voice recordings were later transcribed. We ran the transcriptions through the CLM answer selection module to generate answers for each question. This generated question and answer pairs based on how the system would have responded to the participant questions if the speech recognition was perfect. This is our second set of question answer pairs – the TRS-QA set. Appendix B shows a sample dialog between a participant and SGT Blackwell.

Next we used three human raters to judge the appropriateness of both sets. Using a scale of 1-6 (see Appendix A) each rater judged the appropriateness of SGT Blackwell's answers to the questions posed by the participants. We evaluated the agreement between raters by computing Cronbach's alpha score, which measures consistency in the data. The alpha score is 0.929 for TRS-QA and 0.916 for ASR-QA, which indicate high consistency among the raters.

The average appropriateness score for TRS-QA is 4.83 and 4.56 for ASR-QA. The difference in the scores is statistically significant according to t-test with the cutoff set to 5%. It may indicate that ASR quality has a significant impact on answer selection.

We computed the Word Error Rate (WER) between the transcribed question text and the ASR output. Thus each question-answer pair in the ASR-QA and TRS-QA data set has a WER score assigned to it. The average WER score is 37.33%.

We analyzed sensitivity of the appropriateness score to input errors. Figure 1a and 1b show plots of the cumulative average appropriateness score (CAA) as function of WER: for each WER value $t$ we average appropriateness scores for all questions-answer pairs with WER score less than

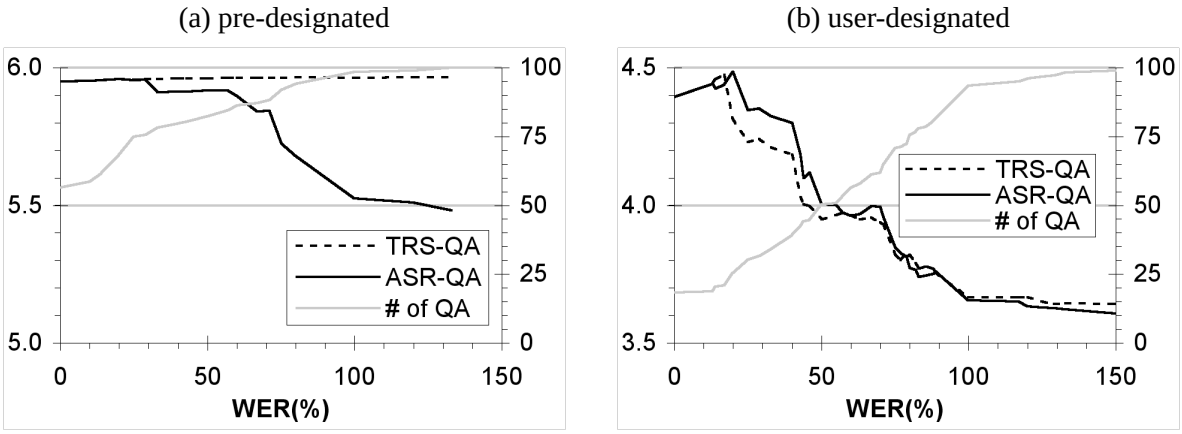


Figure 1: Shows the cumulative average appropriateness score (CAA) of (a) pre-designated and (b) user-designated question-answer pairs as function of the ASR's output word error rate. We show the scores for TRS-QA (dotted black line) and ASR-QA (solid black line). We also show the percentage of the question-answer pairs with the WER score below a given value ("# ofQA") as a gray line with the corresponding values on the right Y axis.

or equal to $t$.

$$CAA(t) = \frac{1}{|S|} \sum_{p \in S} A(p), S = \{p | WER(p) \leq t\}$$

where $p$ is a question-answer pair, $A(p)$ is the appropriateness score for $p$, and $WER(p)$ is the WER score for $p$. It is the expected value of the appropriateness score if the ASR WER was at most $t$.

Both figures show the $CAA$ values for TRS-QA (dotted black line) and ASR-QA (solid black line). Both figures also show the percentage of the question-answer pairs with the WER score below a given value, i.e., the cumulative distribution function (CDF) for the WER as a gray line with the corresponding values depicted on the right Y axis.

Figure 1a shows these plots for the pre-designated questions. The values of $CAA$ for TRS-QA and ASR-QA are approximately the same between 0 and 60% WER. $CAA$ for ASR-QA decreases for WER above 60% – as the input becomes more and more garbled, it becomes more difficult for the CLM module to select an appropriate answer. We confirmed this observation by calculating t-test scores at each WER value: the differences between $CAA(t)$ scores are statistically significant for $t > 60\%$. It indicates that until WER exceeds 60% there is no noticeable effect on the quality of answer selection, which means that our answer selection technique is robust relative to the quality of the input.

Figure 1b shows the same plots for the user-designated questions. Here the system has to deal with questions it has never seen before. $CAA$ values decrease for both TRS-QA and ASR-QA as WER increases. Both ASR and CLM were trained on the same data set and out of vocabulary words that affect ASR performance, affect CLM performance as well.

## 6 Conclusions and future work

In this paper we presented a method for efficient construction of conversational virtual characters. These characters accept spoken input from a user, convert it to text, and select the appropriate response using statistical language modeling techniques from cross-lingual information retrieval. We showed that in this domain the performance of our answer selection approach significantly exceeds the performance of a state of the art text classification method. We also showed that our technique is very robust to the quality of the input and can be effectively used with existing speech recognition technology.

Preliminary failure analysis indicates a few directions for improving the system's quality. First, we should continue collecting more training data and extending the question sets.

Second, we could have the system generate a confidence score for its classification decisions. Then the answers with a low confidence score can be replaced with an answer that prompts the user to rephrase her question. The system would then

use the original and the rephrased version to repeat the answer selection process.

Finally, we observed that a notable percent of misclassifications results from the user asking a question that has a strong context dependency on the previous answer or question. We are presently looking into incorporating this context information into the answer selection process.

## Acknowledgments

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

## References

James Allan, Jamie Callan, W. Bruce Croft, Lisa Ballesteros, Donald Byrd, Russell Swan, and Jinxi Xu. 1998. Inquery does battle with TREC-6. In *Sixth Text REtrieval Conference (TREC-6)*, pages 169–206, Gaithersburg, Maryland, USA.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jennifer Chu-Carroll and Bob Carpenter. 1999. Vector-based natural language call routing. *Journal of Computational Linguistics*, 25(30):361–388.

Sudeep Gandhe, Andrew S. Gordon, and David Traum. 2006. Improving question-answering with linking dialogues. In *Proceedings of the 11th international conference on Intelligent user interfaces (IUI'06)*, pages 369–371, New York, NY, USA. ACM Press.

T. Hoffman. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International ACM SIGIR Conference*, pages 50–57.

Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202.

Victor Lavrenko. 2004. *A Generative Theory of Relevance*. Ph.D. thesis, University of Massachusetts at Amherst.

Anton Leuski. 2004. Email is a stage: discovering people roles from email archives. In *Proceedings of 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*, pages 502–503, Sheffield, United Kingdom. ACM Press. NY, USA.

David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. 1996. Training algorithms for linear text classifiers. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–306, Zurich, Switzerland.

Bryan Pellom. 2001. Sonic: The university of colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder, CO.

Jay M. Ponte and W. Bruce Croft. 1997. Text segmentation by topic. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 120–129.

Abhinav Sethy, Panayiotis Georgiou, and Shrikanth Narayanan. 2005. Building topic specific language models from webdata using competitive models. In *Proceedings of EUROSPEECH*, Lisbon, Portugal.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the Twenty-first international conference on Machine learning*, Banff, Alberta, Canada.

Ellen M. Voorhees. 2003. Overview of the trec 2003 question answering track. In *Proceedings of The Twelfth Text Retrieval Conference*, pages 54–69.

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

## A Appropriateness grading

Gandhe and his colleagues (Gandhe et al., 2006) suggested the following grading scheme that we used in our user evaluation.

| Grade | Description |
|---|---|
| 1 | Response is not related in any way the question |
| 2 | Response contains some discussion of people or objects mentioned in the question, but does not really address the question itself. |
| 3 | Response partially addresses the question, but little or no coherence between the question and response. |
| 4 | Response does mostly address the question, but with major problems in the coherence between question and response; seems like the response is really addressing a different question than the one asked. |
| 5 | Response does address the question, but the transition is somewhat awkward. |
| 6 | Response answers the question in a perfectly fluent manner. |

Table 3: The question-answer appropriateness scale.

## B Sample dialog

This section presents an example of a participant's dialog with SGT Blackwell. We only show spontaneous questions asked by the participant. Each of the participant's questions (called "Reporter" in this transcript) is followed by the ASR output of that line in brackets. That is what the answer selection module "heard." Note that ASR does make an occasional mistake and the answer selection is able to recover gracefully until the next to the last question, where the input was so garbled that only one correct word got through.

**Reporter:** Who created you (Who created you)

**Blackwell:** Tell you the truth, I have spent a lot of time with the smart folks at ICT, testing and tweaking. I am about ready to get out there and train with real soldiers. I am thinking first stop is going to be Fort Sill. But my ultimate goal is to get forward deployed...

**Reporter:** Why should we spend money on you (What should we spend money on you)

**Blackwell:** Why do you need me?! What you should be saying is "How did you get along without me?" I will show you how to be a leader, how to make critical decisions under stress... And I am glad to be here, because I can help. I know I can. Look... I have one mission: to make better soldiers, make better warfighters. And if I can ... can save one life... bring one loved one back to his family... Well, then, I am definitely worth it. You be the judge.

**Reporter:** What's your value (What's the value in)

**Blackwell:** I can do "Virtually" anything. I do mission rehearsals, Urban Combat training, language training - I speak Farsi, Pashtun and Arabic. "Mar Haba shabab! That means how are you in Arabic.

**Reporter:** Do you really work or is this just a demo (G. really worker is this just a demo)

**Blackwell:** Sorry. That's outside my AO.

**Reporter:** How could you be better (How could you be better)

**Blackwell:** Stop mumbling. Just kidding. I didn't get that.

**Reporter:** What do you wish you could do (When you wish you could do)

**Blackwell:** Career Army. That is what I am thinking. Just call me a lifer.

**Reporter:** What kind of people have you trained (What other people in the turn in)

**Blackwell:** I teach cultural awareness, function as a training mentor and can call in the full range of Joint fires and effects.

**Reporter:** What's the hardest part of your job (He was the hardest are you job)

**Blackwell:** I am not authorized to comment on that

**Reporter:** You have any virtual friends (The universal friends)

**Blackwell:** Sorry, I can’t hear you

**Reporter:** Are you hard of hearing (Are you had a hearing)

**Blackwell:** I can’t understand you.

## C SGT Blackwell

Figure 2: A photograph of the SGT Blackwell system setup.