ASKING QUESTIONS TO LIMITED DOMAIN VIRTUAL CHARACTERS: HOW GOOD DOES SPEECH RECOGNITION HAVE TO BE?

Anton Leuski, Brandon Kennedy, 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

ABSTRACT

In this paper, we describe the evaluation of a limited domain question-answering characters, particularly as to the effect of non-optimal speech recognition, and the ability to appropriately answer novel questions. Results show that answering ability is robust until speech recognition reaches over 60% Word error rate.

1. INTRODUCTION

Immersive virtual environments that simulate real-life scenarios are proving to be a very effective approach for training. An important part of training for today's soldier involves interacting with people, especially people from different cultures, and thus interacting with *virtual humans* has great potential for increasing the scope of virtual training. While computer graphics has made great strides in realistic-looking computer generated characters, there is still a perception among many that language processing technology is not yet adequate for such applications. While it is true that full human-level Artificial Intelligence is still beyond current capabilities, we contend that technology is already adequate to support useful training applications and provide greater immersion than could be experienced without it. To use familiar examples from Hollywood, while we can't produce the likes of Lt Commander Data or the Holographic Doctor from *Star Trek*, characters such as the hologram of Dr Lanning from *I Robot* or Vox, the Holographic Librarian from the *Time Machine* are within our capabilities.

At the 24th Army Science Conference in 2004, we presented a demo of *Sgt Blackwell*, a limited domain character who is projected on a transparent screen (like Vox), and can answer a reporter's questions about the Army, the Institute for Creative Technologies, his technology and possible use in training applications, and the Conference, itself. Sgt Blackwell recognizes spoken language, uses state of the art classification technology (Leuski *et al.*, 2006a; Leuski *et al.*, 2006b) to pick the best answer (if possible, or an answer indicating the question can not be answered, if not), and then speaks the answer, while performing accompanying head and body movements. The demo was very successful, in part, because the speech recognizer was specially trained for the voices of the three demo reporters, who also knew the right questions to ask, to stay within the vocabulary.

While extremely high performance for speech recognition can be achieved with training or limited vocabulary, such circumstances are less useful for training situations in which formulating the question is part of the learning objective. In this paper we explore the question of the degree to which imperfect speech recognition impacts the overall performance. In the next section, we describe the general class of *question-answering characters*, which have great promise to fill an important niche in entertainment and training applications. In Section 3, we focus on Sgt Blackwell, one such question-answering character. In Section 4, we describe our classfication approach for determing answers to questions. In Section 5, we describe the evaluation approach which we use to address the questions of how good the character's answering ability is both for sub-optimal speech recognition and to novel questions. The results are presented in the next section, and we conclude in Section 7 with analysis of the results and future work to improve the performance of the character.

2. QUESTION ANSWERING CHARACTERS

Sergeant Blackwell is one example of a class we call *question answering characters*. Question answering characters have a set of knowledge they can impart when asked and goals for the presentation of this information subject to appropriate conditions. A key is remaining in character when deciding how to react to questions. Question answering characters can be used for training, education, and entertainment. At the Institute for Creative Technologies we have recently built several question answering characters, including also characters a reporter can interview to piece together a news story, and more recently characters for training tactical questioning.

The technology in question answering characters bears some similarity to that used in question answering systems. Automatic question answering has been studied extensively in recent years. For example, there is a significant body of research done in the context of the question answering 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, question answering characters focus on the answer's *appropriateness*. For example, in the reporter domain, an evasive, misleading, or an “honestly” wrong answer from an interview subject would be appropriate but might not be relevant. We try to highlight that distinction by talking about question answering *characters* as opposed to question answering systems.

We also expect question answering characters to have a natural spoken language interaction with the student. Our technical requirements for such a question answering character are: (a) it should be able to understand spoken language; (b) it should be robust to disfluencies in conversational English; and (c) it should be relatively fast, easy, and inexpensive to construct without the need for extensive domain knowledge and dialog system design expertise.

3. SGT BLACKWELL

Sergeant Blackwell is a question answering character who was originally designed to serve as an information kiosk at the army conference. (Figure 1 shows a photograph of the system).

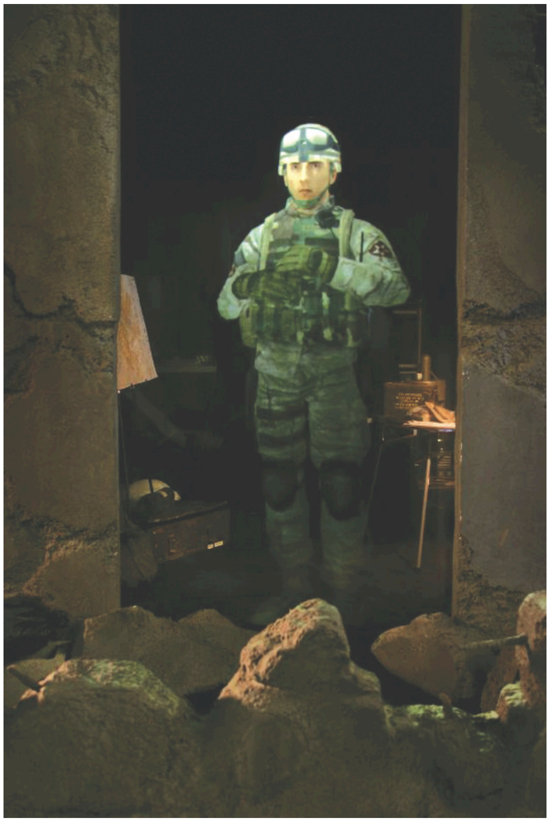

Figure 1. A photograph of SGT Blackwell in his environment.

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 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., "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..."

Here is an example of a participant's dialog with SGT Blackwell. 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 one question, where the input was so garbled, so 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.

4. 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.

The first method is to view the answer selection problem 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. There are several techniques for text classification. In our testing with this domain, a multi-class Support Vector Machines (SVM$^{\text{struct}}$) classifier with an exponential kernel (Tsochantaridis *et al.*, 2004) had 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.

The second method is to view the answer selection problem 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. We adapted an approach suggested by Lavrenko (Lavrenko, 2004), which assumed that all the word dependencies are defined by a vector of possibly unknown parameters on the language model.

The third method is also based on information retrieval, but uses techniques from cross-language retrieval. In the second method, we assumed that both questions and answers used the same vocabulary and had the same a priori language models. Clearly, this is not the case. For example, consider the following exchange: "what happened here?" — "well, 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. There are two ways we can solve the cross-language information retrieval problem: 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.

We have implemented answer classifiers using each of these methods. Comparison using a 10-fold cross-validation method on our Sgt Blackwell training data showed that the cross-language information retrieval method significantly outperformed the others, with greater than 16% improvement over the SVM version. More details on both the specific algorithms and results are presented elsewhere (Leuski *et al.*, 2006b).

5. EVALUATION APPROACH

While it is of great importance for a question answering system, classification accuracy on perfect input is only the beginning of the requirements for a question answering character. What is most important for maintaining an immersive experience with a character is the quality and coherence of the reply under all conditions, including when unknown or uninterpretable questions are presented.

Here we study two aspects of this problem:

1. To what degree does erroneous input affect the ability to give an appropriate response?
2. How appropriate are the responses to unknown questions?

We study these questions by recruiting subjects to ask questions to Sgt Blackwell. We used two different sets of questions to address the two questions. 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 if we had used general purpose acoustic models (which would not have been quite as high for the main speakers). Given a range of different speakers, this allowed us to calculate the role of a variety of speech error rates on classifier performance for both known and unknown questions.

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 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.

Table 1. The question-answer appropriateness scale with sample answers in response to the question "How exactly do you think you can help the real army soldiers?"

Grade	Description
1	Response is not related in any way the question. *For me being virtual, Christmas is not a big deal. But I know it is for other real guys like me, who will be overseas this Christmas...*
2	Response contains some discussion of people or objects mentioned in the question, but does not really address the question itself. *Where was not I trained? Infantry, Armor, Arty, Benning, Bragg, Airborne, Ranger, you name it...*
3	Response partially addresses the question, but little or no coherence between the question and response. *I am a virtual soldier. Which has got lots of big advantages. First of all I am readily deployable. And I can work 24-7 without a break ...*
4	Response does mostly address the question, but with major problems in the coherence between question and response. *I am not actually flesh and blood, sir. I am pixels and polygons. But that means I can be anywhere and do anything. I can function in urban operations, food distributions, or SASO...*
5	Response does address the question, but the transition is somewhat awkward. *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...*
6	Response answers the question in a perfectly fluent manner. *Think of it this way: while I will never be able to do what real soldiers do, I can help my flesh and blood brethren learn how to better do their business out of the line of fire, so that they can be more capable and better prepared when they finally do get into it...*

The voice recordings were later transcribed and used as input to the answer selection module to generate answers for each question. These answers are 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.

Next we used three human raters to judge the appropriateness of both sets, using the coding scheme from (Gandhe *et al.*, 2006). This scheme addresses both the relevancy and the fluency of the answer on a scale of 1-6. Table 1 shows the scale and an example of the coding scheme applied to the Blackwell domain.

Using the scale of 1-6 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 indicates high consistency among the raters.

6. RESULTS

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 Score (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 2a and 2b 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 or equal to t.

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

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.

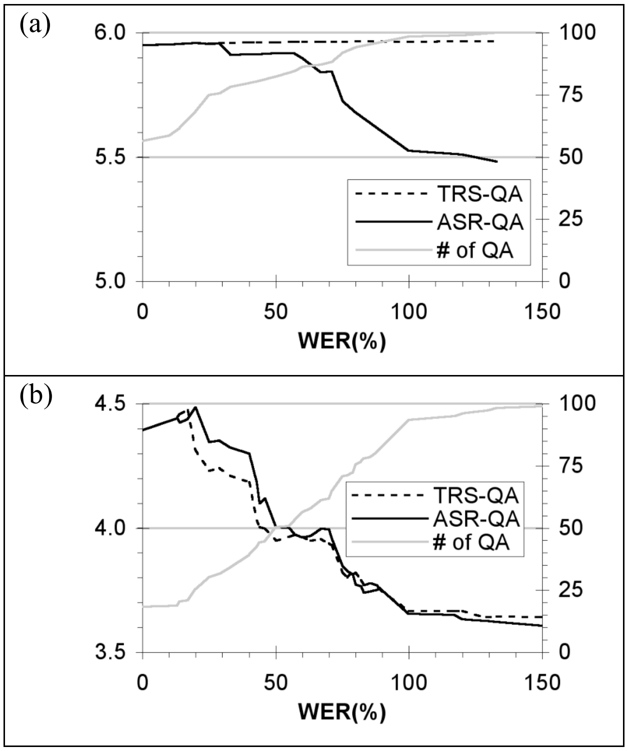

Figure 2. Shows the cumulative average appropriateness score of (a) pre-designated and (b) user-designated question-answer pairs as function of the ASR's output word error rate.

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

Figure 2a 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 classifier 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 2b 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 the classifier were trained on the same data set and out of vocabulary words that affect ASR performance, affect classifier performance as well.

7. CONCLUSIONS

The results in the previous section show both a strong level of performance for Sgt Blackwell even for unseen questions, while there is still quite a bit of room for improvement. One fairly surprising result is that the quality of speech recognition is not the most important factor. Researchers (Kamm and Walker, 1997) have shown that various measures of recognition performance such as WER have a large impact on dialogue quality for task oriented dialogues. In contrast, for our domain speech recognition performance is not a significant factor until the error rate becomes very high. For applications such as this, using appropriate selection technology, current recognition rates already allow viable applications. Interestingly, others have also found that speech recognition rate is not correlated with learning for a tutorial dialogue system in the physics domain (Forbes-Riley and Litman, 2005).

Other than improvement of the speech recognition component, there are other avenues that can be taken to improve the system performance. The first direction is the collection and labeling of additional training data, so the classifier can recognize a wider range of input.

The second task is more sophisticated techniques for dealing with hard to classify input. The current system has an "off-topic" class which is chosen if none of the others is high enough. This results in a reply which addresses the fact that the user asked a question, without providing specific information. Several of these are illustrated in the sample dialogue presented in Section 3. In the original system, these were chosen at random, however closer analysis revealed that different off-topic answers would have different levels of appropriateness. Elsewhere (Patel *et al.*, 2006), we show how having multiple classes of off-topic answers and doing a second tier of classification can improve the acceptability of both off-topic question responses and total performance including both on-topic and off-topic responses. Another technique that can be tried is to engage in specific clarification dialogues to ask the user to repeat or rephrase the question, and then use both the original and subsequent questions together as part of the classification task.

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

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

Katherine Forbes-Riley and Diane J. Litman. 2005. Correlating student acoustic-prosodic profiles with student learning in spoken tutoring dialogues. In *Proceedings. Interspeech* 2005, 157–160, Lisbon, Portugal.

Sudeep Gandhe, Andrew S. Gordon, and David Traum. 2006. Improving question-answering with linking dialogues. In *Proceedings IUI 2006*, 369–371, Sydney, Australia.

C. Kamm and M. Walker. 1997. Design and evaluation of spoken dialog systems.

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

Anton Leuski, Jarrell Pair, David Traum, Peter J. McNerney, Panayiotis Georgiou, and Ronakkumar Patel. 2006a. How to talk to a hologram. In *Proceedings of IUI 2006*, 360–362, Sydney, Australia.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006b. Building effective question answering characters. In *Proceedings of SIGdial 2006*.

Anton Leuski. 2004. Email is a stage: discovering people roles from email archives. In *Proceedings of SIGIR 2004*, 502–503, Sheffield, United Kingdom.

David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. 1996. Training algorithms for linear text classifiers. In *Proceedings of SIGIR 1996*, 298–306, Zurich, Switzerland.

Ronakkumar Patel, Anton Leuski, and David Traum. 2006. Dealing with out of domain questions in virtual characters. In *Proceedings of IVA 2006*.

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 1st European Conference on Research and Advanced Technology for Digital Libraries*, 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 ICML 2004*, Banff, Alberta, Canada.

Ellen M. Voorhees. 2003. Overview of the trec 2003 question answering track. In *Proceedings of TREC 2003*, 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.