Dealing with Out of Domain Questions in Virtual Characters

Ronakkumar Patel, Anton Leuski, and David Traum

Institute for Creative Technologies
University of Southern California
Marina del Rey, CA 90292, USA
{patelr, leuski, traum}@ict.usc.edu

Abstract. We consider the problem of designing virtual characters that support speech-based interactions in a limited domain. Previously we have shown that classification can be an effective and robust tool for selecting appropriate in-domain responses. In this paper, we consider the problem of dealing with out-of-domain user questions. We introduce a taxonomy of out-of-domain response types. We consider three classification architectures for selecting the most appropriate out-of-domain responses. We evaluate these architectures and show that they significantly improve the quality of the response selection making the user's interaction with the virtual character more natural and engaging.

1 Introduction

Previous work has shown that limited domain virtual humans that use spoken interaction can be quite successful in terms of delivering quality answers to in-domain questions [2, 3]. Question-answering characters can serve a number of purposes, including entertainment, training, and education. For a question-answering character, a key point is to give human-like responses to questions when no answer is available. The character should act like a person who either does not know or does not want to reveal the answer: recognizing explicitly that something is "off-topic" and giving a response indicating this recognition is better than providing an inappropriate in-domain answer. While a character could be constructed to always reply with something generic like "I don't know", this can lead to repetitive behavior that breaks a sense of immersion. Having a set of such answers allows the character to seem more engaging, by producing some variety in his responses. Thus we have constructed a set of off-topic responses for our characters to choose from.

We have found, however, that not all off-topic responses are equally satisfactory as replies to each of a range of off-topic questions. In this paper we explore whether the general category "off-topic" can be broken down into appropriate sub-categories to achieve higher performance. We use the SGT Blackwell character [2, 3], as a testbed for this exploration, and create a taxonomy of types of off-topic areas, a set of replies for the SGT Blackwell character for each area. We further evaluate performance of several classification-based architectures that

use the off-topic taxonomy, as to how satisfactory the answers are. The results show that the best architecture significantly out-performs the baseline character, – which does not use the taxonomy, – on both on-topic and off-topic questions.

In the next section we give an overview of the SGT Blackwell character and the baseline question-understanding/response. In Section 3 we discuss a taxonomy of off-topic response classes, which we hope can reduce the number of inappropriate off-topic responses. In Section 4 we describe three different classification-based architectures, which are intended to improve the baseline classifier, using the off-topic taxonomy. In Section 5, we present the results of evaluating the three architectures with respect to the quality of answers given. Finally, in Section 6 we summarize our results and outline some directions for future work.

Fig. 1. SGT Blackwell

2 The baseline SGT Blackwell System

SGT Blackwell, shown in Figure 1, is a life-sized character projected on a transparent screen. He is meant to answer questions from a user acting as a reporter

interviewing him about his role in the Army and the technology at the Institute for Creative Technology that created him. A user talks to SGT Blackwell using a head mounted microphone. For speech recognition, we use a hybrid limited domain/general language model [4], built using the SONIC system [5]. A classifier [3] then analyzes the text output and selects the highest scoring answer. The answers are pre-recorded audio clips linked with animation, which are played through the game engine to show SGT Blackwell providing the response. SGT Blackwell's responses include spoken lines ranging from one word to a couple paragraphs. There are 55 content answers with domain information. When SGT Blackwell detects a question that cannot be answered with one of the content (on-topic) answers, he picks a random answer from a pool of 17 off-topic answers.

The classifier is based on statistical language modeling techniques used in cross-lingual information retrieval. It represents a text string with a language model – a probability distribution over the words in the string. The classifier views both questions and answers as samples from two different "languages" the language of questions and the language of answers. Given an input question from the user, the classifier calculates the language model of the most likely answer for the question, – it uses the training data as a dictionary to "translate" the question into an answer, – then it compares that model to the language model of individual answers, and selects the best matching response. We showed that this technique outperforms traditional text classification approaches such as support-vector machines for tasks that have a large number of response classes [3].

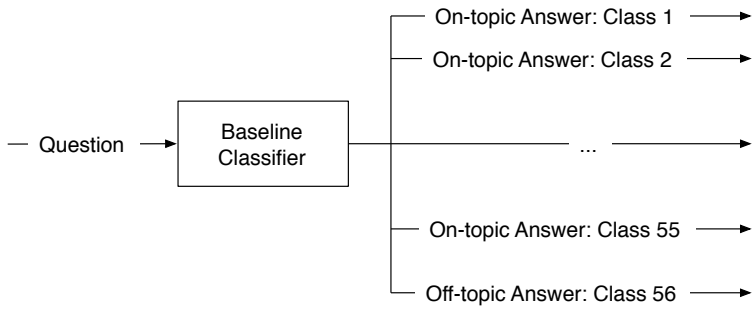

Fig. 2. Baseline classifier architecture

In order to train the classifier, we have created a training corpus of questions linked to either one of the 55 content classes or the "off-topic" class. Questions and answers were created using a multi-stage process, including scripted initial questions, manual paraphrases, and collected questions from a Wizard of Oz study, in which naive users were allowed to ask whatever they wanted, after a brief description of the intended domain. We also used human coders to

link questions to appropriate answers using the question-answer quality ranking scheme suggested by Gandhe and his colleagues [6], described in Section 5. The baseline training set included 1572 questions linked to the 56 answer classes.

Figure 2 shows the design of the question-response part of the initial system, which serves as the baseline for our study. It has one classifier, which can deliver on-topic or off-topic answers based the input question. This design assumes that all off-topic answers are equally appropriate to any off-topic question, so we put them into one class. Thus we have 56 response classes for the baseline, 55 for on-topic and 1 for off-topic answers.

This architecture works well at providing the correct answer when the question is in-domain, and recognizing when a question is out of domain [3], however, the answers for out of domain questions are still often far from satisfactory. We noticed that sometimes one out of domain answer works very well for the question while the answer (randomly) selected by the system did not work so well. Some examples of these kinds of problems are shown in Table 1. Column 1 shows some of out of domain questions asked by users. Column 2 shows the answers SGT Blackwell actually provided, while the third column shows other answers that were available in the off-topic set that our human raters preferred.

Table 1. Improvement in selection of off-topic answers

Question	Randomly Selected	Human Selected
What color are my eyes?	I am not liberty to discuss that	You might want to put that one to a real human.
How is the weather?	I can tell you but I'd have to kill you	I would like to know that too.
Have you got any medals?	I would like to know that too.	No
Where are you going next?	Yes	You'll have to talk to the PAO on that one.

3 Taxonomy of out-of-domain response classes

In order to achieve better conversation with a virtual character, we tried to group the off-topic responses into coherent classes for which the same (set of) answers would apply. First, there is a set of questions which can not be understood as coherent, whether because of speech recognition errors or not enough in-vocabulary words. For these, the character can simply state that he did not hear or understand. Next, there is a simple class of yes-no questions, yielding classes for positive and negative responses. Of the remaining questions, we can make a distinction between those that are really out of the domain vs. those that

Table 2. Taxonomy of out-of-domain response classes with their meaning

Classes	Meaning and Example	
Don't Understand	A question that does not make any sense and/or is very hard to interpret.	
	All the region.	*Stop mumbling. Just kidding. I didn't get that.*
Out of Domain	A question that asks something that is not about the topic(s) the character is prepared to talk about	
	Where is the bathroom?	*I don't have that information*
Unknown	A question that concerns the domain, but for which the character does not have an appropriate answer.	
	What does AO mean?	*I would like to know that too*
Restriction	A question that the character can legitimately refuse to answer.	
	Who do you think of new army uniform?	*I'm not at liberty to discuss.*
Pass	A question that could be better answered by some other domain entity rather than character.	
	When will you become Major?	*You'll have to talk to the PAO on that one.*
Leave to human	A question about specific human characteristics rather than about the domain.	
	How much do you weigh?	*You might want to put that one to a real human.*
Negative	A question that can be answered with a negative answer.	
	Do you have wife?	*No*
Positive	A question that can be answered with a positive answer.	
	Do you like the army?	*Yes*

are appropriately in the domain, but the character's knowledge base does not have an appropriate answer. For those off-topic questions that are in-domain, we can further distinguish between whether it would be legitimate for the character to not know the answer, refuse to tell the answer, or defer the answer to someone else. Of the latter category we also distinguish between specifically asking another domain entity, or a generic "real human". Putting these distinctions together, we end up with the 8 classes shown in Table 2. This table shows the eight classes, along with a brief definition and an example of a question answer pair.

Table 3 shows the complete set of answers for each of these classes for the initial SGT Blackwell character. Since he represents a soldier, his answers are framed in the way a soldier might put them. Things he doesn't want to talk about can be characterized as classified information. He also refers to AO "area of operations", and PAO "Public Affairs Officer" as ways of characterizing the limits of what he can talk about.

Table 3. Response classes and their pool of answers

Classes	Response
Don't Understand	Sorry, I can't hear you.
	I can't understand you.
	Stop mumbling. Just kidding. I didn't get that.
Out of Domain	I don't have that information.
	Sorry. That's outside my AO.
Unknown	I can tell you.....but I would have to kill you (smirks).
	I would like to know that too.
Restriction	That's classified.
	I am not authorized to comment on that.
	No comment.
	I'm not at liberty to discuss.
Pass	You'll have to talk to the PAO on that one.
Leave to Human	You might want to put that one to a real human.
Negative	No.
	Negative, sir.
Positive	Yes.
	Roger.

We had three annotators use the above descriptions and examples to categorize off-topic responses into one of the off-topic classes. After removing duplicate and redundant questions, we end up with collection of 1000 on-topic questions and 300 off-topic questions.

4 Using the off-topic classes to improve classification and answers

Depending on the way we want to mix or separate the on-topic and off-topic classes, there are several different ways that we could use the data to perform classification and answer selection. These methods use different combinations of different classifiers to achieve the desired effect. We built four classifiers, as described in Table 4 and four classifier architectures, as shown in Figures 2, 3, 4, and 5.

Table 4. Classifier descriptions

Name	Description
Baseline	The classifier used for the original SGT Blackwell character, with 55 on-topic classes and one off-topic class
Binary	A binary classifier, which determines only whether a question was on-topic or off-topic
Off-topic	A classifier that assumes its input is off-topic and classifies to one of the 8 off-topic classes
On-topic	A classifier that assumes its input is on-topic and classifies to one of the 55 on-topic classes
Combined	A classifier that treats on-topic and off-topic classes the same, and classifies to one of 63 classes (55 on-topic and 8 off-topic)

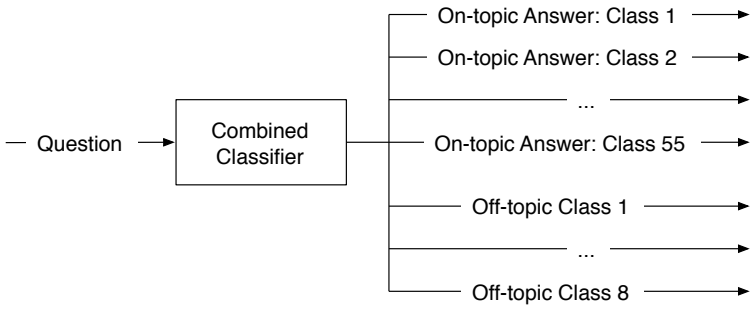

Fig. 3. Architecture 1 with one classifier

The simplest modified architecture is shown in Figure 3 – this is parallel to the baseline architecture shown in Figure 2, although re-trained with new data

including more classes. For this classifier we used all 1300 training question-answer pairs.

Since off-topic and on-topic classes are very different, both in the size of the classes and the specificity of the answer, it also seemed prudent to experiment with other classification architectures and methods for distinguishing on-topic from off-topic questions. Our second architecture, shown in Figure 4, includes two classifiers, the baseline, including one off-topic class and 55 on-topic classes (trained on all 1300 training examples), as well as the off-topic classifier (trained on the 300 off-topic examples). This architecture is most directly comparable to the baseline architecture, since only the method of treating off-topic responses has changed.

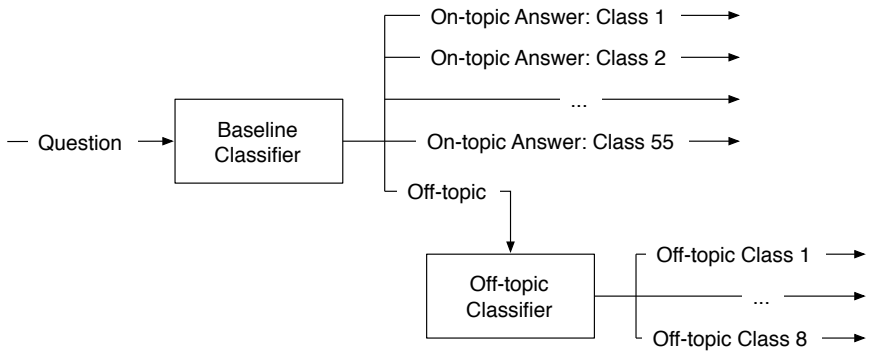

Fig. 4. Architecture 2 with two classifiers

Finally, we separate the decision problem of "on-topic" vs "off-topic" from the classification within those general categories using the architecture shown in Figure 5. Here we have a two pass-classification procedure, first using the binary classifier (trained on all 1300 examples), and then dispatching to either the on-topic classifier (55 classes, trained on 1000 on-topic examples) or the off-topic classifier (8 classes, trained on 300 off-topic examples).

5 Evaluation

We evaluated performance of these architectures in order to address the following questions:

1. Is it really helpful to divide off-topic answers into different disjoint classes?
2. If so, which architecture is best among the proposed three?

To answer these questions, we performed a study parallel to the original SGT Blackwell evaluation of the baseline system [2, 3]. Our test set had 150 questions,

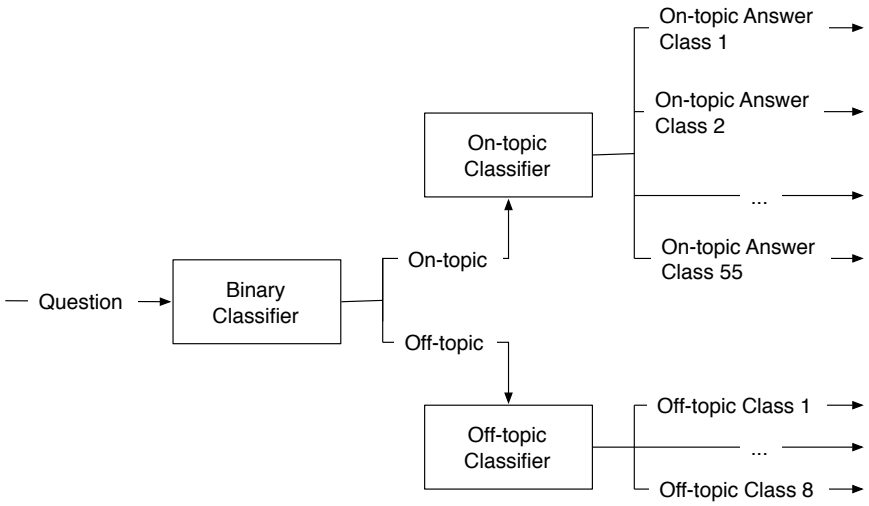

Fig. 5. Architecture 3 with three classifiers

none of which was included in training set. Out of 150 questions 100 were on-topic and 50 were off-topic. This ratio of 1/3 off-topic questions and 2/3 on-topic questions for the test set, was derived from our previous data [3]. The same 150 questions were classified by our baseline architecture shown in Figure 2 as well as the new classifier architectures in Figures 3, 4, and 5. This resulted in 150 question-answer pairs for each architecture.

Three human raters were asked to judge the appropriateness of all Q-A sets, using the 1-6 scale [6], shown in Table 5. We evaluated the agreement between raters by computing Cronbach's alpha score, which measures consistency in the data [1]. The alpha score is 0.885 for baseline, 0.849 for first, 0.781 for second and 0.835 for third architecture respectively, which indicate high consistency among the raters.

The average appropriateness scores for all four architectures is displayed in Table 6. The first row shows the overall average for all 150 examples (100 on-topic and 50 off-topic questions). The differences in the scores are statistically significant according to pair-wise t-test with the cutoff set to 5% except for the difference between the baseline and Architecture 1 with one classifier. We can see a marked improvement for Architecture 3, and a slighter improvement for Architecture 2. Looking more closely, the next rows break out the scores for on and off-topic questions. Architecture 3 with 3 classifiers outperforms the baseline on both on-topic and off-topic responses. While all three architectures outperform the baseline on off-topic responses, we can see that Architecture 1 slightly under-performs for on-topic answers (presumably because they are more likely to be confused with the individual off-topic classes than the overall "off-

Table 5. Appropriateness coding scheme

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 do 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 6. Average appropriateness score for all architectures

	Architecture			
	Baseline	1	2	3
Avg. Score	3.92	3.89	4.16	4.63
Avg. Score (On-topic)	4.58	4.44	4.65	5.14
Ave. Score (Off-topic)	2.59	2.77	3.17	3.62

topic" category), and is not to be preferred. We can thus conclude that the off-topic categories are indeed useful, but we must be careful in how we use them in a classifier. Architecture 3 with three classifiers dominates the other three, and is thus preferred, at least for this data set.

Table 7. On and off-topic mis-classification

	Architecture			
	Baseline	1	2	3
On label as Off	0.05	0.03	0.05	0.03
Off label as On	0.07	0.09	0.07	0.13

Table 7 shows how often the classification architectures confuse on-topic from off-topic questions. For the baseline architecture and Architecture 2, an off-topic label means the question was assigned to the off-topic class, while an on-topic label means that one of the other classes was chosen by the baseline classifier.

For Architecture 1 an on-topic label means that one of the on-topic classes was chosen, while an off-topic label means that one of the eight off-topic classes was chosen. Finally, for Architecture 3, the label comes from the results of the binary classifier. The first row in Table 7 displays the ration of total classifications in which the system produces an off-topic answer to an on-topic question. The second row displays the ratio of total classifications in which an on-topic answer was given for an off-topic question. Interestingly, even though Architecture 3 has higher response scores, it has a higher error rate at the binary classification task than the others, which indicates a hybrid architecture might perform even better.

6 Conclusion and Future work

In this paper we examined the issue of how to improve the performance of a limited-domain question-answering character in the case where the character is given an out of domain question. After some analysis of the types of problems that arise, we constructed a taxonomy of 8 types of off-topic questions and answers. After annotating our data with these classes, we experimented with different classification architectures, and evaluated the performance of these architectures, showing significant improvement in the answers for the three classifier architecture over the baseline system.

While these results are encouraging, there is still much room for more improvement. First, we should probably be a bit more systematic about the definitions and examples for each of these classes. Second, it is unclear to what extent our results would generalize over other data sets and different characters. Finally, it may be the case that other types of classifiers might be appropriate for some of the more specialized tasks.

Acknowledgments

This work 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. The authors would like to thank Jillian Gerten, Ashish Vaswani, Brandon Kennedy and Jaimin Vaidya for useful discussions and their expertise as human coders in mapping of questions and answers set.

References

1. Chronbach, L. J.: Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, (1951) 297–333.
2. Leuski, A., Pair, J., Traum, D., McNerney, P.J., Georgiou, P., Patel, R.: How to talk to a hologram. In Edmonds, E., Riecken, D., Paris, C.L., Sidner, C.L., eds.: Proceedings of the 11th international conference on Intelligent user interfaces (IUI'06), Sydney, Australia, ACM Press New York, NY, USA (2006) 360–362

3. Leuski, A., Patel, R., Traum, D., Kennedy, B.: Building effective question answering characters. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. (2006)
4. Sethy, A., Georgiou, P., Narayanan, S.: Building topic specific language models from webdata using competitive models. In: Proceedings of EUROSPEECH, Lisbon, Portugal (2005)
5. Pellom, B.: Sonic: The University of Colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder, CO (2001)
6. Gandhe, S., Gordon, A.S., Traum, D.: Improving question-answering with linking dialogues. In: Proceedings of the 11th international conference on Intelligent user interfaces (IUI'06), New York, NY, USA, ACM Press (2006) 369–371