Augmenting Conversational Characters with Generated Question-Answer Pairs

Elnaz Nouri and **Ron Artstein** and **Anton Leuski** and **David Traum**
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094-2536, USA

Abstract

We take a conversational character trained on a set of linked question-answer pairs authored by hand, and augment its training data by adding sets of question-answer pairs which are generated automatically from texts on different topics. The augmented characters can answer questions about the new topics, at the cost of some performance loss on questions about the topics that the original character was trained to answer.

Introduction

This paper presents an experiment which takes a conversational character trained on a set of hand-authored, linked question-answer pairs, and adds to the character's knowledge by allowing it to "read" texts about new topics using existing question generation tools. In previous work we have shown how such tools can be used to create a new character from plain text: the tools transform the text into a set of linked question-answer pairs, and these pairs are then imported as a knowledge base into an engine that drives conversational characters (Chen et al. 2011). The present work investigates if the same method can also be used to add knowledge to an existing character. We take an established, hand-authored character and augment its knowledge base with question-answer pairs generated from texts on different topics. The augmented characters are evaluated both on how they answer questions about the original topics, using an extensive test set for the original character, as well as on how they answer questions about the new topics, using small constructed test sets. The results show that the augmented characters can answer questions about the new topics, at the cost of some performance loss on questions about the original topic. The remainder of the paper describes the experiment in detail.

Method

Tools

The characters in our experiments are driven by NPCEditor (Leuski and Traum 2010), a text classification engine for conversational characters which is available for download as part of the ICT Virtual Human Toolkit.[1] NPCEditor is trained on a knowledge base in the form of linked question-answer pairs, and is able to answer novel questions by selecting the most appropriate response from the available answers in the knowledge base. For each new input question, NPCEditor computes a language model for the ideal answer using the linked training data; it then compares the language model of the ideal answer to those of all of the answers in the knowledge base, and selects the closest available answer based on a similarity metric between language models. The use of language models allows NPCEditor to overcome some variation in the phrasing of questions, and retrieve appropriate responses for questions it has not seen in the training data.

To "read" plain text articles into such a conversational character we use two existing question generation tools: Question Transducer (Heilman and Smith 2009), and a reimplementation by Xuchen Yao called OpenAryhpe.[2] Question Transducer identifies sentences in the source text and transforms them into questions in three stages: selecting and forming declarative sentences using paraphrasing rules and word substitutions, syntactically transforming the declarative sentences into questions, and scoring and ranking the resulting questions. OpenAryhpe implements the first two stages (paraphrasing and transformation, but not ranking), and uses additional Named Entity Recognizers in order to identify more phrases that can be questioned. Both tools generate sets of linked question-answer pairs; these can be imported as a training knowledge base into NPCEditor to create a new character, or added to an existing knowledge base consisting of linked questions and answers.

Materials

The base character for the experiment is the twins Ada and Grace, a pair of virtual characters situated in the Museum of Science in Boston where they serve as virtual guides (Swartout et al. 2010). The Twins answer questions from visitors about exhibits in the museum and about science in general; these topics will be referred to as the *original topics*, because these are what the original knowledge base was

[1] http://vhtoolkit.ict.usc.edu

[2] http://code.google.com/p/openaryhpe

Source Article	Length (words)	Extracted Q-A Pairs		
		OA[a]	QT[b]	Tot.[c]
Australia	567	240	308	500
Beer	299	146	231	315
Ludvig_van_Beethoven	765	338	635	889

[a]OpenAryhpe [b]Question Transducer [c]The total is less than the sum of OA and QT due to overlap.

Table 1: Wikipedia text excerpts and question-answer pairs

designed for. All the training data for the Twins were authored by hand.

The base character is successively augmented by adding question-answer pairs generated automatically according to the method described in Chen et al. (2011):

1. Text excerpts are manually selected from articles in Simple English Wikipedia.[3] We started with a sample of 14 texts which were prepared for a pilot experiment, and arbitrarily chose three of the texts for the present experiment (Table 1). These will be referred to as the *new topics*.
2. The texts are transformed into question-answer pairs using the two question generation tools mentioned above, Question Transducer and OpenAryhpe.
3. The question-answer pairs from both question generation tools are imported into NPCEditor and added to the Twins' training data.

We trained a total of five augmented characters in addition to the baseline; Table 2 shows the number of questions, answers and links in each of the sets of training data.

Test set

Original topics. To test performance of the augmented characters on questions from the Twins' original topics we use an extensive test set collected during the initial stages of the Twins' deployment at the Museum of Science, when visitor interaction was done primarily through trained handlers (relying on handlers allowed us to deploy the characters prior to collecting the required amount of visitor speech, mostly from children, necessary to train acoustic models for speech recognition). The handlers relay the visitors' questions through a microphone to be processed by a speech recognizer; they also tend to reformulate user questions to better match the questions in the Twins' knowledge base, and many of their utterances are a precise word for word match of utterances in the Twins' training data. Such utterances are a good test case for the classifier because the intended correct responses are known, but actual performance varies due to speech recognition errors; they thus test the ability of the classifier to overcome variation in the phrasing of questions.

The same test set was used in Wang et al. (2011) to compare various methods of handling speech recognizer output; here we use it to compare different character knowledge bases. The speech recognizer output remains constant in the different testing runs – all characters are tested on exactly the same utterance texts. From the point of view of a classifier for the original topics, question-answer pairs from the new topics can be considered as training noise; what the different characters test, then, is how the addition of knowledge bases for new topics affects the performance of the original, hand-authored part of the character.

[3]http://simple.wikipedia.org

Character	Questions	Answers	Q-A Pairs
Twins	406	148	483
Twins + Australia	652	342	999
Twins + Beer	559	268	807
Twins + Beethoven	849	421	1412
Twins + Australia + Beer	804	462	1323
Twins + Aus. + Beer + Beeth.	1245	735	2252

Table 2: Training data for the various characters

The test set consists of 7690 utterances. These utterances were collected on 56 individual days so they represent several hundred visitors; the majority of the utterances (almost 6000) come from two handlers. Each utterance contains the original speech recognizer output retrieved from the system logs (speech recognition was performed using the SONIC toolkit, Pellom and Hacıoğlu 2001/2005). Some of the utterances are identical – there is a total of 2264 utterance types (speech recognizer output), corresponding to 265 transcribed utterance types (transcriptions were performed manually). The median word error rate for the utterances is 20% (mean 29%, standard deviation 36%). This level of word error rate is acceptable for this application – as we will see below, the original character fails to understand only 10% of the input utterances, and this error rate declines rapidly when the character is allowed to identify its own non-understanding (Figure 1).

New topics. We also test performance of the augmented characters on questions relating to the new topics. Since we do not have an extensive set of spoken utterances as for the Twins' original topics, we use a small, manually authored test set that was developed in a pilot experiment for Chen et al. (2011). The test set includes, for each topic, a small set of questions, with associated answers that are taken from the source text.

Evaluation

Original topics. To evaluate performance on questions from the original topics, we run our test set through each of the characters in Table 2. For each utterance we send the text of the speech recognizer output to NPCEditor, and compare the response to the answers linked to the corresponding manual transcription. A response is scored as correct if it matches one of the linked answers, otherwise it is scored as incorrect. We also collect the confidence scores reported by NPCEditor in order to enable the analysis in Figure 1 below (the confidence score is the inverse of the Kullback-Leibler divergence between the language models of the ideal response and the actual response; see Leuski and Traum 2010).

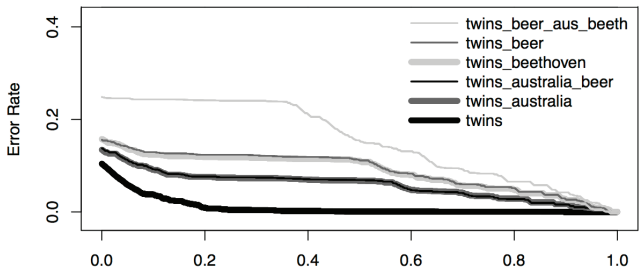

Figure 1: Trade-off between errors and non-returns

New topics. Performance on questions relating to the added knowledge bases is evaluated manually. The text of each question is sent to NPCEditor, and the response is marked as correct, partly correct, or incorrect after comparing it to the predetermined answer key. All the manual ratings were performed by the first author.

Results

Performance on the original topics

Just counting the correct and incorrect responses is not sufficient for evaluating character performance, because NPCEditor employs dialogue management logic designed to avoid the worst outputs. During training, NPCEditor calculates a response threshold based on the classifier's confidence in the appropriateness of selected responses: this threshold finds an optimal balance between false positives (inappropriate responses above threshold) and false negatives (appropriate responses below threshold) on the training data. At runtime, if the confidence for a selected response falls below the predetermined threshold, that response is replaced with an "off-topic" utterance that asks the user to repeat the question or takes initiative and changes the topic (Leuski et al. 2006); such failure to return a response (also called non-understanding, Bohus and Rudnicky 2005) is usually preferred over returning an inappropriate one (misunderstanding).

The capability to not return a response is crucial in keeping conversational characters coherent, but it is not captured by standard classifier evaluation methods such as accuracy, recall (proportion of correct responses that were retrieved), or precision (proportion of retrieved responses that are correct). We cannot use the default threshold calculated by NPCEditor during training, because these default thresholds yield different return rates for different characters. We therefore use a graphical evaluation method which looks at the full trade-off between return levels and error rates (Artstein 2011).

For each test utterance we logged the top-ranked response together with its confidence score, and then we plotted the rate of off-topics against errors at each possible threshold; this was done separately for each character (since confidence scores are based on parameters learned during training, they are not comparable across characters). Figure 1 shows the curves for the baseline character and the 5 augmented characters: non-returns are plotted on the horizontal axis and corresponding error rates on the vertical axis; at the extreme right, where no responses are returned, error rates are necessarily zero for all characters. Lower curves indicate better performance.

Character	Test Set		
	Australia	Beer	Beethoven
N =	9	10	14
Stand-alone (without Twins)	5	8	10
Twins + Australia	5		
Twins + Beer		7	
Twins + Beethoven			9
Twins + Australia + Beer	5	7	
Twins + Aus. + Beer + Beeth.	5	6	9
Twins	0	0	0

Table 3: Correct responses from the augmented characters

The best performer on the test set for the original topics is the original Twins character, with a 10% error rate when all responses are returned, and virtually no errors with a non-return rate above 20%. Performance degrades somewhat with the successive addition of automatically generated questions from the new topics, though the degradation is mitigated to some extent when higher non-return rates are acceptable. In exchange for an increased error rate on questions from the original topics, the augmented characters can now answer questions pertaining to the new topics.

Performance on the new topics

Each new topic has a small set of manually constructed test questions. We first ran each test set through a stand-alone character generated automatically using only the question-answer pairs extracted from the corresponding source text. This represents the ceiling we can expect for performance of the augmented characters. Performance is not perfect, but for each topic the character gives a correct or partly correct answer more than half of the time (Table 3).

Performance of the augmented characters is only slightly degraded compared to that of the stand-alone characters. We tested each question set on those characters that included the relevant knowledge base. The number of correct or partially correct responses is shown in Table 3, and in each case, the correct responses are a subset of the correct responses returned by the corresponding stand-alone character. We also tested the question sets on the original Twins character – as expected, none of the returned responses is a correct answer.

Conclusion

Our experiment demonstrates that automatic question generation is a viable way to add knowledge from sources in plain text to a hand-authored question answering conversational character. Adding such knowledge does affect the character's performance, increasing the error rate on questions from the original topics that the character was designed to answer. In return, however, the augmented character can

also answer questions about the new topics that are covered in the textual sources.

Acknowledgments

Thanks to Michael Heilman and Xuchen Yao, the respective authors of Question Transducer and OpenAryhpe, for giving us access to their code and allowing us to use it in our experiments.

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

Artstein, R. 2011. Error return plots. In *Proceedings of the SIGDIAL 2011 Conference*, 319–324. Portland, Oregon: Association for Computational Linguistics.

Bohus, D., and Rudnicky, A. I. 2005. Sorry, I didn't catch that! – An investigation of non-understanding errors and recovery strategies. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, 128–143.

Chen, G.; Tosch, E.; Artstein, R.; Leuski, A.; and Traum, D. 2011. Evaluating conversational characters created through question generation. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, 343–344. Palm Beach, Florida: AAAI Press.

Heilman, M., and Smith, N. A. 2009. Question generation via overgenerating transformations and ranking. Technical Report CMU-LTI-09-013, Carnegie Mellon University Language Technologies Institute.

Leuski, A., and Traum, D. 2010. Practical language processing for virtual humans. In *Twenty-Second Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-10)*.

Leuski, A.; Patel, R.; Traum, D.; and Kennedy, B. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*.

Pellom, B., and Hacıoğlu, K. 2001/2005. SONIC: The University of Colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder.

Swartout, W.; Traum, D.; Artstein, R.; Noren, D.; Debevec, P.; Bronnenkant, K.; Williams, J.; Leuski, A.; Narayanan, S.; Piepol, D.; Lane, C.; Morie, J.; Aggarwal, P.; Liewer, M.; Chiang, J.-Y.; Gerten, J.; Chu, S.; and White, K. 2010. Ada and Grace: Toward realistic and engaging virtual museum guides. In Allbeck, J.; Badler, N.; Bickmore, T.; and Pelachaud, Catherine Safonova, A., eds., *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20–22, 2010 Proceedings*, volume 6356 of *Lecture Notes in Artificial Intelligence*, 286–300. Heidelberg: Springer.

Wang, W. Y.; Artstein, R.; Leuski, A.; and Traum, D. 2011. Improving spoken dialogue understanding using phonetic mixture models. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, 329–334. Palm Beach, Florida: AAAI Press.