Evaluating Conversational Characters Created through Question Generation

Grace Chen[*] and **Emma Tosch**[†] 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

Question generation tools can be used to extract a question-answer database from text articles. We investigate how suitable this technique is for giving domain-specific knowledge to conversational characters. We tested these characters by collecting questions and answers from naive participants, running the questions through the character, and comparing the system responses to the participant answers. Characters gave a full or partial answer to 53% of the user questions which had an answer available in the source text, and 43% of all questions asked. Performance was better for questions asked after the user had read the source text, and also varied by question type: the best results were answers to *who* questions, while answers to yes/no questions were among the poorer performers. The results show that question generation is a promising method for creating a question answering conversational character from an existing text.

Introduction

For virtual question-answering characters (Leuski et al. 2006), providing knowledge is a major bottleneck. Typically the knowledge needs to be authored manually; acquiring it from corpora requires large amounts of conversational data which are not readily available (Gandhe and Traum 2008; 2010). On the other hand, information is available on many topics in text form. Automatic question answering retrieves answers from both information databases (Katz 1988) as well as unstructured text collections (Voorhees 2003); such online question-answering systems have been incorporated into conversational characters (Mehta and Corradini 2008).

We propose a simple, practical way to allow virtual characters to answer questions about a given text. We use a publicly available toolkit for creating virtual characters – the ICT Virtual Human Toolkit (http://vhtoolkit.ict.usc.edu) – with a knowledge base in the form of linked question-answer pairs (Leuski and Traum 2010). Instead of authoring the knowledge base by hand, we populate it with question-answer pairs derived from a text through the use of question generation tools. This paper describes an experiment which demonstrates the viability of this approach.

Method

Materials

As raw materials we selected three text excerpts from Simple English Wikipedia (http://simple.wikipedia.org). Articles were retrieved on June 10, 2010, and their text was manually copied and pasted from a web browser into a text editor. We conducted a pilot study on 14 articles using a small, manually constructed test set, and chose three of the top five performers for the main experiment: Sword (363 words), River (368 words), and Roman_Empire (466 words).

Question generation

We extracted question-answer pairs from the texts using existing tools: Question Transducer (Heilman and Smith 2009) and OpenAryhpe (http://code.google.com/p/openaryhpe). The generated pairs were imported into NPCEditor, a text classification system that drives virtual characters (Leuski and Traum 2010). NPCEditor learns a mapping between the language models of the questions and answers in the character knowledge base, and then for each new input utterance selects one of the available outputs based on the learned mapping. We trained NPCEditor on knowledge bases formed by pooling the question-answer pairs extracted by the two tools – a total of 407 pairs for the swords topic, 339 for rivers, and 550 pairs for the Roman Empire.

Test set

We collected test questions from 22 participants. Participants first wrote 5 questions about a particular topic, without having read any text materials about it. They then read the source text about the topic, and wrote 5 additional questions about the topic, based on the source text. Finally, each participant provided answers to all of their questions, in the form of a contiguous segment of text from the source. If the participant felt that the text did not contain an answer, they marked the answer as "N/A".

The data were collected using the Qualtrics on-line survey tool (http://qualtrics.com). Each participant provided 30 questions and answers in total – 5 for each topic before reading the text and 5 after the reading – for a total of 660 collected questions and corresponding answers (220 for each topic). Topics were presented to all the participants in the same order: swords, rivers, Roman Empire.

[*] Now at California State University, Long Beach.

[†] Now at Brandeis University.

Evaluation

We presented each of the test questions to NPCEditor and rated the system response against the user-provided answer. Two raters (the first two authors) rated all of the responses independently on a three-point scale: 2 for providing a complete answer, 1 for a partial answer, and 0 for a response that does not answer the question. Agreement between the annotators was high: $\alpha = 0.865$ (Krippendorff 1980).

Results

Question distribution

The most common question type was *what* (55%), followed by yes/no (9%), *who* (8%), *when*, *where* and *how much* (7% each), *how* (6%), and *why* (1%). Three user utterances were classified as "other". The distribution is different for questions produced before and after reading the source text ($\chi^2(8) = 37$, $p < 0.001$): participants produced more *what* and yes/no questions after reading the source text – they had asked more varied question types before. Of the questions asked before reading the text, 46% had no answers; of those asked after, only 5% were without answers. Question types also differed by topic, and again the difference was significant ($\chi^2(16) = 115$, $p < 0.001$). Rivers received no *who* or *when* questions, which together constituted almost a third of the questions about the Roman Empire.

Answer quality

We used the mean of the scores given by the two raters, so each answer received a score between 0 and 2. We ran a 4-way ANOVA of text topic, availability of a response, question authoring time, and question type. Three of the factors came out as highly significant main effects.

Response availability. Responses to questions with available answers ranked higher, with a mean of 0.82 compared to 0.15 ($F(1,592) = 106$, $p < 0.001$). Overall, 53% of the questions with available answers received a full or partial answer (43% of the total questions).

Authoring time. Responses to questions authored after reading the text received higher ratings, with a mean of 0.95 compared to 0.34 ($F(1,592) = 53$, $p < 0.001$). Such questions are more likely to use vocabulary found in the text, with 42% out-of-vocabulary word tokens compared to 52% for questions asked before having read the text (looking only at user questions with available answers). Since training questions are derived from the source text, a better alignment with the user vocabulary should make it easier to map user questions to appropriate answers.

Question type. Ratings were highest for *who* questions (mean 1.15), followed by *when* (0.79), *what* (0.70), *where* (0.54), *how much* (0.53), yes/no (0.27), *how* (0.22), and *why* (0.11) ($F(8,592) = 6.5$, $p < 0.001$). The differences may be due to the abilty of the question generation tools to identify some types of information better than others.

Topic had no significant effect ($F(2,592) = 2.8$, $p = 0.06$). The only significant interactions were between topic and question type ($F(14,592) = 3.6$, $p < 0.001$) and answer availability and question type ($F(7,592) = 2.1$, $p = 0.04$).

Discussion

The experiment demonstrates that our approach is viable – using question generation tools to populate a character knowledge base in question-answer format results in virtual characters that can give appropriate answers to user questions at least some of the time. Some questions do better than others, and *who* questions do particularly well. However, there remain many user questions with an answer in the source text which the character is not able to find, and this is where there is substantial room for improvement. There is a need to bridge the gap between the vocabulary of user questions and extracted questions through improvements to the question generation process and use of lexical resources.

The current work suggests several directions for future research. The experiment only tested characters that answer questions on a single topic from a single source text; we are presently conducting experiments that combine sources on several topics, and add generated question-answer pairs to an existing hand-authored character. We are also looking into using NPCEditor's internal confidence scores to allow the character itself to judge whether an appropriate answer is available. Finally, it would be appropriate to also test the characters in conversation.

Acknowledgments

Thanks to Michael Heilman and Xuchen Yao, the respective authors of Question Transducer and OpenAryhpe, for allowing us to use their code 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

Gandhe, S., and Traum, D. 2008. Evaluation understudy for dialogue coherence models. In *Proc. 9th SIGdial*.

Gandhe, S., and Traum, D. 2010. I've said it before, and I'll say it again: An empirical investigation of the upper bound of the selection approach to dialogue. In *Proc. SIGDIAL 2010*.

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

Katz, B. 1988. Using English for indexing and retrieving. In *Proc. RIAO '88*.

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, California: Sage. 129–154.

Leuski, A., and Traum, D. 2010. Practical language processing for virtual humans. In *Proc. IAAI-10*, 1740–1747.

Leuski, A.; Kennedy, B.; Patel, R.; and Traum, D. 2006. Asking questions to limited domain virtual characters: how good does speech recognition have to be? In *25th Army Science Conference*.

Mehta, M., and Corradini, A. 2008. Handling out of domain topics by a conversational character. In *Proc. DIMEA '08*, 273–280.

Voorhees, E. M. 2003. Overview of the TREC 2003 question answering track. In *Twelfth Text Retrieval Conference*, 54–69.