Evaluation in Natural Language Generation: The Question Generation Task

Vasile Rus
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
vrus@memphis.edu

Zhiqiang Cai
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
zcai@memphis.edu

Arthur C. Graesser
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
a-graesser@memphis.edu

Abstract

Question Generation (QG) is proposed as a shared-task evaluation campaign for evaluating Natural Language Generation (NLG) research. QG is a subclass of NLG that plays an important role in learning environments, information seeking, and other applications. We describe a possible evaluation framework for standardized evaluation of QG that can be used for black-box evaluation, for finer-grained evaluation of QG subcomponents, and for both human and automatic evaluation of performance.

1 Introduction

Natural Language Generation (NLG) is one of the grand challenges of natural language processing and artificial intelligence (Dale et al., 1998). A robust NLG system requires the modeling of speaker's intentions, discourse planning, micro-planning, surface realization, and lexical choices. The complexity of the task presents significant challenges to NLG evaluation, particularly automated evaluation. Major progress towards standardized evaluation exercises of NLG systems will be achieved in shared-task evaluation campaigns (STEC) that are planned over a number of years. They start with simple (sub)tasks in the early years that invite wide participation by various research groups and then gradually increase the difficulty of the problems addressed. The selected shared task should minimize restrictions on alternative approaches. For instance, the test data should not be specified in representations that are favored by particular systems and researchers. The task should also allow evaluation of different aspects of NLG and should be relevant to a variety of applications.

We propose an evaluation framework for the task of Question Generation (QG). QG is defined as a task with simple input and output. The framework accommodates black-box evaluation of alternative approaches and finer-grained evaluation at micro-planning, surface realization, and lexical choice levels. The initial task is extendable to permit evaluation at all levels, including speaker's intentions and discourse planning. QG is an essential component of learning environments, help systems, information seeking systems, and a myriad of other applications (Lauer et al., 1992). A QG system would be useful for building an automated trainer for learners to ask better questions and for building better hint and question asking facilities in intelligent tutoring systems (Graesser et al., 2001). In addition to learning environments, QG facilities could help improve Question Answering systems by launching questions proactively and jumping in with suggested queries when dead-ends in inquiry inevitably occur.

QG as a testbed can benefit from previous experience on standardized evaluations of related shared tasks in Question Answering (TREC-Question Answering track; http://trec.nist.gov) and from evaluations of Intelligent Tutoring Systems such as AutoTutor (Graesser et al., 2001). Data sources from those previous shared tasks can be easily adapted to a QG task with relative efficient costs.

This paper defines the task of QG, briefly describes the QG evaluation framework, and presents evaluation metrics.

2 The Question Generation Task

Our approach to QG assumes that there are one or more sentences (i.e., possible answers to a user question) given as input, whereas the task of a QG approach is to generate questions related to this input. This textual specification of both input and output should encourage wide adoption of the task by many research groups because it does not impose any representational restrictions on the input or output. Various approaches can of course use their own internal representations for input. The input is limited to 1-2 sentences to simplify the task and minimize complexities of discourse level processing. The task can eventually be extended to incorporate discourse by specifying a paragraph as input and asking for a set of related questions as output.

Two data sources are available to extract input and output data. Both consist of a set of sentences and each sentence's associated human-generated questions. The first one is Auto-Tutor (Graesser et al., 2001), an Intelligent Tutoring System that holds dialogues with the learner in natural language. For each input sentence taken from such dialogues, there is an associated set of questions. The second source is the TREC Question Answering track, where thousands of Question-Answer pairs are available from Question Answering evaluations since 1999. In this case, for each sentence (answer) we have a single associated question.

The input (Expectation, Answer) and output data (Questions) are sufficiently well formulated to make the setup of such standardized evaluation quick and easy. The researcher community can target specific feature evaluations of generation systems. For example, by selecting sentences with associated Who? or What person? questions from the TREC QA source, one can focus on testing the capabilities of a system for generating person-related questions. Similarly, one can select sentence-question pairs tailored to the evaluation of lexical choice characteristics of a generation system.

3 Evaluation

The output of a QG system can be evaluated using either automated evaluation or manual evaluation. Automated evaluation can use methods similar to ROUGE in summarization and BLEU/NIST in machine translation which are based on N-gram co-occurrence. An extreme solution is to consider exact question matching in which the generated question and the expected question in the gold standard, containing the ideal/expected questions, have to be identical for a hit. Manual evaluation recruits experts to assess the output of various approaches along different criteria.

The evaluation of any NLG system includes multiple criteria, such as user satisfiability, linguistic well-foundedness, maintainability, cost efficiency, output quality, and variability. Other metrics can serve as proxies for some criteria. For example, precision may be a proxy for user satisfiability. In a recent study (Cai et al., 2006), our group used *precision* and *recall*. *Precision* is the proportion of good questions out of all generated questions. *Recall* or coverage is difficult to objectively compute because the number of questions generated from a sentence is theoretically indeterminate. A recall measure can be observed in specific experiments. In the TREC QA data set, there is only one question for each each answer. *Recall* would be the proportion of those TREC QA questions that are present in the output of a QG system.

References

Z. Cai, V. Rus, H.J. Kim, S. Susarla, P. Karnam, and A.C. Graesser. 2006. NLGML: A natural language generation markup language. In T.C. Reeves and S.F. Yamashita, editors, *Proceedings of E-Learning Conference*, pages 2747–2752, Honolulu, Hawaii. AACE.

Robert Dale, Donia Scot, and Barbara di Eugenio. 1998. Special Issue on Natural Language Generation. *Computational Linguistics*, 24(3):346–353, September.

Arthur C. Graesser, Kurt VanLehn, Carolyn P. Rose, Pamela W. Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22(4):39–52.

T.W. Lauer, E. Peacock, and A.C. Graesser. 1992. *Questions and Information Systems*. Lawrence Erlbaum Associates, Hillsdale, NJ.