Question Generation Shared Task and Evaluation Challenge – Status Report

Vasile Rus
Department of Computer Science
The University of Memphis
Memphis, TN 38152, USA
vrus@memphis.edu

Brendan Wyse
Centre for Research in Computing
The Open University
Walton Hall, Milton Keynes, UK
bjwyse@gmail.com

Paul Piwek
Centre for Research in Computing
The Open University
Walton Hall, Milton Keynes, UK
p.piwek@open.ac.uk

Mihai Lintean
Department of Computer Science
The University of Memphis
Memphis, TN 38152, USA
mclinten@memphis.edu

Svetlana Stoyanchev
Centre for Research in Computing
The Open University
Walton Hall, Milton Keynes, UK
s.stoyanchev@open.ac.uk

Cristian Moldovan
Department of Computer Science
The University of Memphis
Memphis, TN 38152, USA
cmldovan@memphis.edu

Abstract

The First Shared Task Evaluation Challenge on Question Generation took place in 2010 as part of the 3rd workshop on Question Generation. The campaign included two tasks: Question Generation from Sentences and Question Generation from Paragraphs. This status report briefly summarizes the motivation, tasks and results. Lessons learned relevant to future QG-STECs are also offered.

1 Introduction

Automatically generating questions is an important task in many different contexts including dialogue systems, intelligent tutoring systems, automated assessment and search interfaces. Questions are used to express informational needs: when we do not know something, the natural thing to do is to ask about it. As computer systems become more advanced and are expected to be more adaptive and autonomous, their informational needs grow, and being equipped with the ability to ask questions has clear advantages. State-of-the-art spoken dialogue systems are a good case in point: where would they be without the ability to ask questions, for example, about the user's goals (“Where would you like to travel to?”) or about their understanding of the users' utterances (“Did you say ‘London’?”)?

Of course, the purpose of asking questions is not limited to satisfying straightforward informational needs. In a classroom, a teacher may ask a question, not because she doesn't know the answer, but because she wants to know whether the student knows the answer (or perhaps she wants to provide the student with a hint that will help him solve whichever problem he is dealing with). Generating such questions automatically is a central task for intelligent tutoring systems. Exam questions are

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 318–320,
Nancy, France, September 2011. ©2011 Association for Computational Linguistics

another case in point. In the context of automated assessment, generating questions automatically from educational resources is a great challenge, with, potentially, tremendous impact.

2 QGSTEC Input and Output

Question Generation (QG) has recently been defined as the task of automatically generating questions (Piwek et al., 2008; Rus & Graesser, 2009). Whereas this definition more or less fixes the output of QG, it leaves open what the input is, and how the input relates to the output. For the First QGSTEC, the decision on input was aimed at attracting as many participants as possible and promoting a fair comparison environment. Thus, rather than adopting a specific semantic representation as input, the input for both tasks was raw text. Participants were free to (pre)process the text with their own and/or off-the-shelf NLP tools. As for the relation between input and output, the decision was made that the output question should be answered by (part of) the input text – thus the tasks were the inverse of Question Answering. Regarding the output evaluation, again to maximize participation in the tasks, only generic criteria (such as fluency and ambiguity), as opposed to application-specific criteria, were used.

Input data sources for both tasks were Wikipedia, OpenLearn, and Yahoo!Answers.

3 Question Generation from Sentences

Participants were given a set of inputs, with each input consisting of: (A) a single sentence and (B) a specific target question type (e.g., WHO?, WHY?, HOW?, WHEN?).

For each input, the task was to generate 2 questions of the specified target question type. For example, for input instance:

- The poet Rudyard Kipling lost his only son in the trenches in 1915.
- WHO

Two different questions of the specified type that are answered by input sentence were expected, e.g.: 1) “Who lost his only son in the trenches in 1915?” and 2) “Who did Rudyard Kipling lose in the trenches in 1915?”

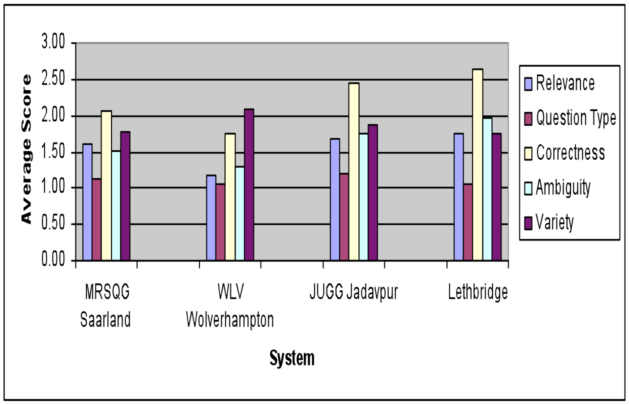

Figure 1: Results for QG from Sentences (without penalty for missing questions)

Five systems entered this task: MRSQG Saarland, WLV Wolverhampton, JUGG Jadavpur and Lethbridge; for descriptions of the systems we refer to Boyer and Piwek (2010). The system-generated questions were scored on five dimensions: Relevance, (Correct) Question Type, (Syntactic) Correctness, Ambiguity and Variety (of generated questions). The averaged results for the systems, based on both peer and independent reviewers, are depicted in Figure 1, with lower values indicating better scores. WLV scores best on all criteria except for “Variety”. The picture changes when systems are penalized for missing questions (Figure 2). Now MRSQG outperforms the other systems on all criteria.

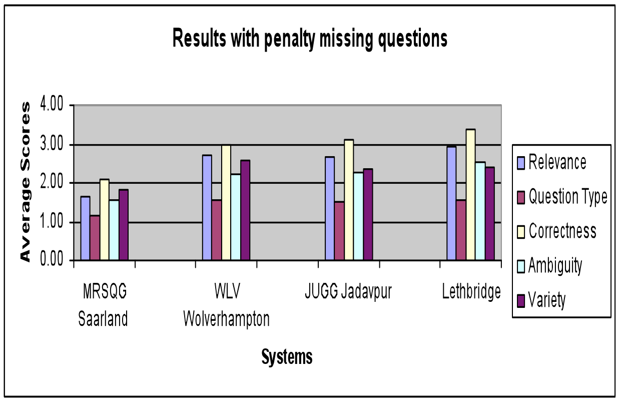

Figure 2: Results for QG from Sentence (with penalty for missing questions)

4 Question Generation from Paragraphs

The inputs for this task were paragraphs such as:

Two-handed backhands have some important advantages over one-handed backhands. Two-handed backhands are generally more accurate because by having two hands on the racquet, this makes it easier to inflict topspin on the ball allowing for more control of the shot. Two-handed backhands are easier to hit for most high balls. Two-handed backhands can be hit with an open stance, whereas one-handers usually have to have a closed stance, which adds further steps (which is a problem at higher levels of play).

For each paragraph, the task was to generate six questions at different levels of specificity: One question that is answered by the paragraph as a whole (e.g. "What are the advantages of two-handed backhands in tennis?"), two medium level questions (e.g., "Why is a two-hand backhand more accurate [when compared to a one-hander]?") asking about major ideas in the paragraphs, e.g. relations among larger chunks of text in the paragraphs such as cause-effect, and three specific question on specific facts (e.g., "What kind of spin does a two-handed backhand inflict on the ball?").

For this task, there was one submission out of five registered participants. The participating team was from University of Pennsylvania (for further details see Boyer & Piwek, 2010). We adopted an independent-judges approach in which two independents human raters judged the submitted questions using five criteria:

Score	**Results/Inter-rater Reliability**
Specificity	General=90%;Medium=121%; Specific=80%; Other = 1.39%/68.76%
Syntactic Correctness	1.82/87.64%
Semantic Correctness	1.97/78.73%
Question Diversity	1.85/100%
Question Type Correctness	83.62%/78.22%

Table 1: Summary of Results for University of Pennsylvania

5 Lessons Learned for Future QG-STECs

The first QG-STEC was a success by many measures including number of participants, results, and resources created. Here we highlight two recommendations for future QG-STECs. Firstly, it is worthwhile considering further fine-tuning of the instructions to judges to improve agreement and possibly replacing rating scales, which we used in evaluating the submissions, with preference judgments as the former seems to pose some challenges such as being unintuitive for raters and the inter-rater agreement tends to be low when using rating scales (Belz & Kow, 2010). Secondly, there is a case for extending the QGSTEC with a task that goes beyond raw text input, given the convergence of semantic representations that is driven by the semantic web.

Acknowledgments

We would like to thank the Engineering and Physical Sciences Research Council who partially supported this work through grant EP/G020981/1 (awarded to Paul Piwek) and the National Science Foundation who partially supported this work through grants RI-0836259 and RI-0938239 (awarded to Vasile Rus).

References

Belz, A. and Kow, E. (2010) Comparing Rating Scales and Preference Judgements in Language Evaluation. In Proceedings of the 6th International Natural Language Generation Conference (INLG'10), pp. 7-15.

Boyer, K.E. and P. Piwek (Eds) (2010). Proceedings of the 3rd Workshop on Question Generation. Carnegie Mellon University, Pittsburgh PA., June 18, 2010.

Piwek, P., H. Prendinger, H. Hernault, and M. Ishizuka (2008). Generating Questions: An Inclusive Characterization and a Dialogue-based Application. In: V. Rus and A. Graesser (eds.), online Proceedings of Workshop on the Question Generation Shared Task and Evaluation Challenge, September 25-26, 2008, NSF, Arlington, VA.

Rus, V. and Graesser, A.C. (2009). Workshop Report: The Question Generation Task and Evaluation Challenge, Institute for Intelligent Systems, Memphis, TN, ISBN: 978-0-615-27428-7.