The First Question Generation Shared Task Evaluation Challenge

Vasile Rus[1], Brendan Wyse[2], Paul Piwek[2], Mihai Lintean[1], Svetlana Stoyanchev[2] and Cristian Moldovan[1]

[1] Department of Computer Science/Institute for Intelligent Systems, The University of Memphis, Memphis, TN, 38152, USA
{vrus,mclinten,cmoldova}@memphis.edu
[2] Centre for Research in Computing, Open University, UK
bjwyse@gmail.com and {p.piwek, s.stoyanchev}@open.ac.uk

Abstract. The paper briefly describes the First Shared Task Evaluation Challenge on Question Generation that took place in Spring 2010. The campaign included two tasks: Task A – Question Generation from Paragraphs and Task B – Question Generation from Sentences. An overview of each of the tasks is provided.

Keywords: question generation, shared task evaluation campaign.

1 Introduction

Question Generation is an essential component of learning environments, help systems, information seeking systems, multi-modal conversations between virtual agents, and a myriad of other applications (Lauer, Peacock, and Graesser, 1992; Piwek et al., 2007).

Question Generation has been recently defined as the task (Rus & Graesser, 2009) of automatically generating questions from some form of input. The input could vary from information in a database to a deep semantic representation to raw text.

The first Shared Task Evaluation Challenge on Question Generation (QG-STEC) follows a long tradition of STECs in Natural Language Processing (see the annual tasks run by the Conference on Natural Language Learning - CoNLL). In particular, the idea of a QG-STEC was inspired by the recent activity in the Natural Language Generation (NLG) community to offer shared task evaluation campaigns as a potential avenue to provide a focus for research in NLG and to increase the visibility of NLG in the wider Natural Language Processing (NLP) community (White and Dale, 2008). It should be noted that the QG is currently perceived as a discourse processing task rather than a traditional NLG task (Rus & Graesser, 2009).

Two core aspects of a question are the goal of the question and its importance. It is difficult to determine whether a particular question is good without knowing the context in which it is posed; ideally one would like to have information about what counts as important and what the goals are in the current context. This suggests that a

STEC on QG should be tied to a particular application, e.g. tutoring systems. However, an application-specific STEC would limit the pool of potential participants to those interested in the target application. Therefore, the challenge was to find a framework in which the goal and importance are intrinsic to the source of questions and less tied to a particular context/application. One possibility was to have the general goal of asking questions about salient items in a source of information, e.g. core ideas in a paragraph of text. Our tasks have been defined with this concept in mind. Adopting the basic principle of application-independence has the advantage of escaping the problem of a limited pool of participants (to those interested in a particular application had that application been chosen as the target for a QG STEC).

Another decision aimed at attracting as many participants as possible and promoting a more fair comparison environment was the input for the QG tasks. Adopting a specific representation for the input would have favored some participants already familiar with such a representation. Therefore, we have adopted as a second guiding principle for the first QG-STEC tasks: no representational commitment. That is, we wanted to have as generic an input as possible. The input to both task A and B in the first QG STEC is raw text.

The First Workshop on Question Generation (www.questiongeneration.org) has identified four categories of QG tasks (Rus & Graesser, 2009): Text-to-Question, Tutorial Dialogue, Assessment, and Query-to-Question. The two tasks in the first QG STEC are part of the Text-to-Question category or part of the Text-to-text Natural Language Generation task categories (Dale & White, 2007). It is important to say that the two tasks offered in the first QG STEC were selected among 5 candidate tasks by the members of the QG community. A preference poll was conducted and the most preferred tasks, Question Generation from Paragraphs (Task A) and Question Generation from Sentences (Task B), were chosen to be offered in the first QG STEC. The other three candidate tasks were: Ranking Automatically Generated Questions (Michael Heilman and Noah Smith), Concept Identification and Ordering (Rodney Nielsen and Lee Becker), and Question Type Identification (Vasile Rus and Arthur Graesser).

There is overlap between Task A and B. This was intentional with the aim of encouraging people preferring one task to participate in the other. The overlap consists of the specific questions in Task A which are more or less similar with the type of questions targeted by Task B.

Overall, we had 1 submission for Task A and 4 submissions for Task B. We also had an additional submission on development data for Task A.

2 TASK A: Question Generation from Paragraphs

1.1 Task Definition

The Question Generation from Paragraphs (QGP) task challenges participants to generate a list of 6 questions from a given input paragraph. The six questions should be at three scope levels: 1 x broad (entire input paragraph), 2 x medium (multiple

sentences), and 3 x specific (sentence or less). The scope is defined by the portion of the paragraph that answers the question.

The Question Generation from Paragraphs (QGP) task has been defined such that it is *application-independent*. *Application-independent* means questions will be judged based on content analysis of the input paragraph; questions whose answers span more input text are ranked higher.

We show next an example paragraph together with six interesting, application-independent questions that could be generated. We will use the paragraph and questions to describe the judging criteria.

Table 1. Example of input paragraph (from http://en.wikipedia.org/wiki/Abraham_lincoln).

Input Paragraph
Abraham Lincoln (February 12, 1809 – April 15, 1865), the 16th President of the United States, successfully led his country through its greatest internal crisis, the American Civil War, preserving the Union and ending slavery. As an outspoken opponent of the expansion of slavery in the United States, Lincoln won the Republican Party nomination in 1860 and was elected president later that year. His tenure in office was occupied primarily with the defeat of the secessionist Confederate States of America in the American Civil War. He introduced measures that resulted in the abolition of slavery, issuing his Emancipation Proclamation in 1863 and promoting the passage of the Thirteenth Amendment to the Constitution. As the civil war was drawing to a close, Lincoln became the first American president to be assassinated.

Table 2. Examples of questions and scores for the paragraph in Table 1.

Questions	Scope
Who is Abraham Lincoln?	*General*
What major measures did President Lincoln introduce?	*Medium*
How did President Lincoln die?	*Medium*
When was Abraham Lincoln elected president?	*Specific*
When was President Lincoln assassinated?	*Specific*
What party did Abraham Lincoln belong to?	*Specific*

A set of five scores, one for each criterion (specificity, syntax, semantics, question type correctness, diversity), and a composite score will be assigned to each question.

Each question at each position will be assigned a composite score ranging from 1 (first/top ranked, best) to 4 (worst rank), 1 meaning the question is at the right level of specificity given its rank (e.g. the broadest question that the whole paragraph answers will get a score of 1 if in the first position) and also it is syntactically and semantically correct as well as unique/diverse from other generated questions in the set.

Ranking of questions based on scope assures a maximum score for the six questions of 1, 2, 2, 3, 3 and 3, respectively. A top-rank score of 1 is assigned to a broad scope question that is also syntactically and semantically correct or acceptable, i.e. if it is semantically ineligible then a decision about its scope cannot be made and thus a worst-rank score of 4 is assigned. A maximum score of 2 is assigned to medium-scope questions while a maximum score of 3 is assigned to specific questions. The best configuration of scores (1, 2, 2, 3, 3, 3) would only be possible for paragraphs that could trigger the required number of questions at each scope level, which may not always be the case.

1.3 Data Sources and Annotation

The primary source of input paragraphs were: Wikipedia, OpenLearn, Yahoo!Answers. We collected 20 paragraphs from each of these three sources. We collected both a development data set (65 paragraphs) and a test data set (60 paragraphs). For the development data set we manually generated and scored 6 questions per paragraph for a total of 6 x 65 = 390 questions.

Paragraphs were selected such that they are self-contained (no need for previous context to be interpreted, e.g. will have no unresolved pronouns) and contain around 5-7 sentences for a total of 100-200 tokens (excluding punctuation). In addition, we aimed for a diversity of topics of general interest.

We also provided discourse relations based on HILDA, a freely available automatic discourse parser (duVerle & Prendinger, 2009).

2 TASK B: Question Generation from Sentences

2.1 Task Definition

Participants were given a set of inputs, with each input consisting of:

- a single sentence and
- a specific target question type (e.g., WHO?, WHY?, HOW?, WHEN?; see below for the complete list of types used in the challenge).

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

Input sentences, 60 in total, were selected from OpenLearn, Wikipedia and Yahoo! Answers (20 inputs from each source). Extremely short or long sentences were not

included. Prior to receiving the actual test data, participants were provided with a development data set consisting of sentences from the aforementioned sources and, for one or more target question types, examples of questions. These questions were manually authored and cross-checked by the team organizing Task B.

The following example is taken from the development data set. Each instance has a unique identifier and information on the source it was extracted from. The <text> element contains the input sentence and the <question> elements contain possible questions. The <question> element has the type attribute for specification of the target question type.

```
<instance id="3">
    <id>OpenLearn</id>
    <source>A103_5</source>
    <text>
        The poet Rudyard Kipling lost his only son
        in the trenches in 1915.
    </text>
    <question type="who">
        Who lost his only son in the trenches in 1915?
    </question>
    <question type="when">
        When did Rudyard Kipling lose his son?
    </question>
    <question type="how many">
        How many sons did Rudyard Kipling have?
    </question>
</instance>
```

Note that input sentences were provided as raw text. Annotations were not provided. There are a variety of NLP open-source tools available to potential participants and the choice of tools and how these tools are used was considered a fundamental part of the challenge.

This task was restricted to the following question types: WHO, WHERE, WHEN, WHICH, WHAT, WHY, HOW MANY/LONG, YES/NO. Participants were provided with this list and definitions of each of the items in it.

2.2 Evaluation criteria for System Outputs and Human Judges

The evaluation criteria fulfilled two roles. Firstly, they were provided to the participants as a specification of the kind of questions that their systems should aim to generate. Secondly, they also played the role of guidelines for the judges of system outputs in the evaluation exercise.

For this task, five criteria were identified: relevance, question type, syntactic correctness and fluency, ambiguity, and variety. All criteria are associated with a scale from 1 to *N* (where *N* is 2, 3 or 4), with 1 being the best score and *N* the worst score.

The procedure for applying these criteria is as follows:

- Each of the criteria is applied *independently* of the other criteria to each of the generated questions (except for the stipulation provided below).

We need some specific stipulations for cases where no question is returned in response to an input. For each target question type, two questions are expected. Consequently, we have the following two possibilities regarding missing questions:

- *No question is returned for a particular target question type*: for each of the missing questions, the worst score is recorded for all criteria.
- *Only one question is returned*: For the missing question, the worst score is assigned on all criteria. The question that is present is scored following the criteria, with the exception of the VARIETY criterion for which the lowest possible score is assigned.

We compute the overall score on a specific criterion. We can also compute a score which aggregates the overall scores for the criteria.

Conclusions

The submissions to the first QG STEC are now being evaluated using peer-review mechanism in which participants blindly evaluate their peers questions. At least two reviews per submissions are performed with the results to be made public at the 3rd Workshop on Question Generation that will take place in June 2010.

Acknowledgments. We are grateful to a number of people who contributed to the success of the First Shared Task Evaluation Challenge on Question Generation: Rodney Nielsen, Amanda Stent, Arthur Graesser, Jose Otero, and James Lester. Also, we would like to thank the National Science Foundation who partially supported this work through grants RI-0836259 and RI-0938239 (awarded to Vasile Rus) and the Engineering and Physical Sciences Research Council who partially supported the effort on Task B through grant EP/G020981/1 (awarded to Paul Piwek). The views expressed in this paper are solely the authors'.

References

1. Lauer, T., Peacock, E., & Graesser, A. C. (1992) (Eds.). *Questions and information systems*. Hillsdale, NJ: Erlbaum.
2. 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.
3. Piwek, P., H. Hernault, H. Prendinger, M. Ishizuka (2007). T2D: Generating Dialogues between Virtual Agents Automatically from Text. In: Intelligent Virtual Agents:

Proceedings of IVA07, LNAI 4722, September 17-19, 2007, Paris, France, (Springer-Verlag, Berlin Heidelberg) pp.161-174
4. Dale, R. & M. White (2007) (Eds.). *Position Papers of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation.*
5. duVerle, D. and Prendinger, H. (2009). A novel discourse parser based on Support Vector Machines. Proc 47th Annual Meeting of the Association for Computational Linguistics and the 4th Int'l Joint Conf on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09), Singapore, Aug 2009 (ACL and AFNLP), pp 665-673.