A New Approach to Ranking Over-Generated Questions

Claire McConnell
University of Pennsylvania, Philadelphia, PA
Prashanth Mannem
International Institute of Information Technology, Hyderabad, India
Rashmi Prasad
University of Wisconsin-Milwaukee, Milwaukee, WI
Aravind Joshi
University of Pennsylvania, Philadelphia, PA

Abstract

We discuss several improvements to the Question Generation Shared Task Evaluation Challenge (QGSTEC) system developed at the University of Pennsylvania in 2010. In addition to enhancing the question generation rules, we have implemented two new components to improve the ranking process. We use topic scoring, a technique developed for summarization, to identify important information for questioning, and language model probabilities to measure grammaticality. Preliminary experiments show that our approach is feasible.

Introduction

Question generation (QG) has important applications in many fields. One useful application of QG is to generate a list of questions about a text focused purely on information presentation, such as a Wikipedia article. The output of such a task would prove valuable in areas such as education, dialog systems and Internet search. Some recent automated QG systems [Heilman and Smith, 2009; Mannem, Prasad, and Joshi, 2010] have approached the task by over-generating questions using transformation rules and selecting the most meaningful ones based on a ranking method. The task of transforming a declarative sentence into a question follows a series of rules that may use preprocessing components, such as part of speech taggers, named entity recognizers, syntactic parsers, and semantic role labelers, to determine the type of question to ask. Ranking the over-generated questions as well as generating and ranking questions whose answers span multiple sentences has proven more difficult.

In this paper, we explore the problems encountered in a system developed for QG on paragraphs and propose several enhancements to the system. We present various rule-based enhancements to the system as well as a ranking system that extends to multi-paragraph QG. The changes to the system focus on three problematic areas in QG: content selection, grammaticality, and Wh word choice. In the first section we give an overview of the existing Penn QGSTEC. Second, we present our evaluation of this system. We then discuss the problem areas and potential solutions in the third section, which include some preliminary experiments. Finally, we conclude with our plans for future work.

Current System Overview

The 2010 QGSTEC System [Mannem, Prasad, and Joshi, 2010], developed at the University of Pennsylvania, uses external NLP tools to parse the input sentences in a paragraph and transform them into general, medium, and specific scope questions. The original release follows the specifications for the 2010 QG Shared Task Evaluation Challenge [Rus et al., 2010]. Given a set of paragraphs, the participants were asked to generate six questions from each paragraph with specific, medium, and general scope levels. The full task description is available at *http://www.questiongeneration.org/QGSTEC2011*.

The following is a sample paragraph from the task description:

> 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... Lincoln won the Republican Party nomination in 1860 and was elected president later that year. 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.

According to the QGSTEC specifications, the "scope" of a question is the amount of the text that the question's answer covers. The general question's scope should be the entire text or most of the text. The medium question's scope should be multiple clauses, while the specific question's scope should be one sentence or less. The QGSTEC description lists the following examples as good general, medium, and specific scope questions for the above paragraph:

General:	Who is Abraham Lincoln?
Medium:	What measures did president Lincoln introduce?
Specific:	What party did Abraham Lincoln belong to?

In the Penn QGSTEC system, the general question is always generated from the paragraph initial sentence, under the assumption that the first sentence gives an overview of what the paragraph is about. The medium and specific questions are generated from the remaining sentences in the paragraph. In order to transform declarative sentences into questions, the system uses the ASSERT semantic role labeling tool [Pradhan et al., 2004] to identify potential answer targets. Some question types are determined purely from the semantic role labels. In other cases, the QGSTEC system uses a named entity tagger [Ratinov and Roth, 2009] to determine if the argument is a person or a location, in which case it assigns a *Who* or *Where* question type. The default question type is *What*.

Consider the following example parsed sentence:

$$[\text{Mary}]_{ARG0} \text{ has } [\text{given}]_{verb} [\text{the book}]_{ARG2} [\text{to John}]_{ARG1}$$
$$[\text{on Friday}]_{ARGM-TMP}$$

In this case, there are four potential target answers, and the system would generate the following questions:

Target	Question
Mary	Who has given the book to John?
the book	What has Mary given to John?
John	Who has Mary given the book to?
on Friday	When has Mary given the book to John?

A more detailed description of the sentence to question transformation process can be found in the system description [Mannem, Prasad, and Joshi, 2010].

After generating an exhaustive list of questions, the medium and specific scope questions are ranked independently based on two criteria. The QGSTEC system first ranks questions according to the depth of the predicate in the dependency parse (obtained with a bidirectional LTAG dependency parser [Shen and Joshi, 2008]), under the assumption that semantic arguments from main clauses contain more important information than those in subordinate clauses. Secondly, questions with more pronouns are given a lower rank based on the intuition that pronouns make questions vague. In the 2010 system, no co-reference resolution was carried out on the input paragraphs.

As also noted in Mannem, Prasad, and Joshi [2010], these two ranking criteria are insufficient for choosing the most meaningful questions. By overlooking information contained in subordinate clauses, the system may exclude some useful questions, and discounting questions based on the number of pronouns used may not be the optimal way to account for vagueness. Finally, the ranking does not take into account grammaticality or information content, which are critical aspects to constructing understandable questions that cover meaningful material.

Evaluation

The first step in the process of improving the current system's QG capability was to evaluate the system. In the evaluation phase, we ran the system on 13 Wikipedia articles of varying length and subject matter. Since the general question does not participate in the ranking process, only the specific and medium questions were included in the evaluation. In total, 97 specific questions and 10 medium questions were generated. These questions were then hand annotated using Heilman and Smith's guidelines for classifying question deficiencies [Heilman and Smith, 2009]. The eight classification categories are: (1) ungrammatical, (2) does not make sense, (3) vague, (4) obvious answer, (5) missing answer, (6) wrong WH word, (7) formatting, and (8) other. It is possible for questions to fall into multiple deficiency categories. Table 1 shows a summary of question deficiencies. The column numbers correspond to the previously mentioned list of deficiency categories.

Table 1: Question Evaluation Results

Qword	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
How	2	4	0	0	9	3	0	0
What	25	24	19	4	7	11	3	1
When	7	7	3	3	2	4	0	1
Where	1	0	1	1	0	2	0	0
Who	1	1	2	0	1	2	0	0
Why	2	0	0	0	0	0	1	0
Total	**38**	**36**	**25**	**8**	**19**	**22**	**4**	**2**

The most common deficiency was "Ungrammatical' (38 questions) followed by "Does not make sense" (36 questions), "Vague" (25 questions), and "Wrong Wh word" (22 questions). We discuss our proposed solutions to these issues in the following section.

Analysis and Solutions

Based on our evaluation of the current system, we have defined three broad categories of problems: *Information Content*, *Grammaticality*, and *Wh word Choice*. We were able to fix many issues by adding hand-written rules, while others were addressed in the ranking mechanism.

Information Content

One of the universal problems with the existing QG system relates to content selection, which involves decisions about whether or not a question is meaningful and worth asking. Methods for content selection may vary depending on the purpose of the questions being generated. For example, one QG task, on the one hand, may require individual questions to be generated, each of which is asked independently of the other. On the other hand, another QG task may require generation of a group of questions to capture the information content of the text as a whole. Finally, it may also be desirable to generate a single general, or topical, question that comprehensively represents the most important information in the text, i.e., what the text is about.

We begin by discussing the issues related to the generation of a list of questions to capture the information content of the input text as a whole. For such a task, there are two problems that we have taken up. First, the existing QGSTEC system can generate multiple indistinguishable questions even when they arise from different target answers. One example might be the following questions:

What are silks produced by?

What are silks mainly produced by?

Such repetitions should be avoided, and in order to address this issue, we compare the words in the set of questions. If two questions share more than 75% of the words, the lower ranked one is removed from the set.

There are also several problems with content selection when the task involves generating individual questions, rather than a group of questions as a whole. One of the more common issues is vagueness. Some examples include the following:

What does it teach?

What does this contrast with?

Previously, the QGSTEC system addressed vagueness by discounting questions based on the number of pronouns in the question. This approach would cause a question such as, "What did Mary buy when she went to the store?" to be unjustly discounted. We have made this rule more specific by excluding questions whose main arguments consist of a single pronoun or a single demonstrative determiner. As a result, questions like "Where did he go?" and "What did John give him?" are excluded, while questions that have pronouns in subordinating clauses are included. It is important to note that there are many other issues that can contribute to vagueness in a question, and some of these problems can be improved using summarization techniques that we will discuss later.

The final problem associated with content selection relates to constructing a general question that alone represents what the text is about. In the system, this is always generated from the paragraph initial sentence. However, while Wikipedia articles typically start with a general topic sentence, the first sentence in a longer text may not always be the best sentence from which to generate a general question. Therefore, methods are needed to find topical sentences that best represent the overall document themes. Such sentences can then be transformed into general questions.

One approach we have incorporated to help identify topical information is a technique used in text summarization. We use TopicS [Louis, 2010], a tool developed at the University of Pennsylvania, to pre-process the input paragraph and generate a list of topic words and their corresponding topic scores. Using the method from Hovy and Lin [2000], the list of words is determined by comparing the word frequencies in the input text to a large background corpus. Next, we score the sentences in the paragraph by the fraction of topic words contained in the sentence. We then use the sentence with the highest topic score to generate the general question.

Our preliminary analysis indicates that topic scoring produces good general questions. Furthermore, even medium and specific scope questions generated from sentences with high topic scores generally have good information content. Our intuition here is that high-scoring sentences tend to be good candidates for questions, because they contain words pertaining to the topic of the text. Other sentences undoubtedly contain worthwhile information; however, the word choice could lead to vagueness when the sentences undergo question transformation. Because of this, we propose to incorporate the topic score as an element of the final ranking for all question scopes.

As a proof of concept, we took a collection of 63 Wikipedia articles, each consisting of one paragraph (on average 8 sentences long), from the QGSTEC2010 test set. Assuming that these paragraphs begin with a general topic sentence, as most Wikipedia articles do, we would expect the initial sentence to have a high topic score. For 38 of the 63 articles, the initial sentence fell into the top quintile of sentences on the topic score. In 53 articles, the first two sentences fell within the top quintile. In this structured domain of Wikipedia articles, the topic score appears to be a very good indicator of candidate sentences for question generation.

Grammaticality

Due to unavoidable errors in the underlying text processing tools, many questions are grammatically indecipherable. Though these sentences may be good candidates for questions, they are useless if not properly generated. In order to alleviate this problem, we incorporated language model probabilities into the ranking process. We use a simple bigram model with Laplace (add-one) smoothing to generate sentence probabilities. The model calculates the question phrase probability based on both token and part of speech bigrams. These probabilities are normalized and averaged to produce a composite score, which is incorporated into the composite ranking. Since questions have different probability distributions than a standard corpus, we used the Microsoft Encarta question database, which consists of about 1,300 well-formed questions, as the background corpus. In general, shorter questions will have a higher language model probability, and we have confronted this issue by multiplying the probability by the length of the question. This ensures that longer questions are not discounted substantially.

Based on preliminary case-by-case analysis, the language model appears to give significant improvements in question grammaticality. Looking at the list of all over-generated questions ranked by language model probability, the more grammatical questions appear to get a higher rank in most cases; however, more extensive evaluation is needed in this area. The evaluation might involve creating a hand annotated data set based on a scale indicating the extent to which questions are grammatically correct. Once we have this, we may be able to better judge how well the language model performs. Furthermore, we believe that the model could be enhanced by incorporating syntactic parse trees as well as a more extensive background corpus.

Wh Word Choice

The current QG system uses five question types: Who, What, Where, When, Why, and How. Often times, the system assigns the wrong question type to the target arguments. This problem is usually associated with errors in the named entity recognizer or the semantic role labeler. There are also many cases where the semantic role labels are correct, but the question type is inappropriate for the target answer.

When Questions *When* questions are generated on targets with temporal semantic roles. We found that there were many instances of bad *When* questions, because the existing

QGSTEC system does not account for all possible scenarios in which temporal relations are expressed.

In many cases, a temporal argument is a bad target entity for a *When* question type. The semantic role labeler labels adverbs such as "often" and "sometimes" as temporal arguments. The *When* questions generated on these target answers are generally confusing or uninformative, so we exclude *When* questions on targets that consist of a single adverb.

Normally, a *When* question is answered with a specific date or time, or a reference to an event that occurred at some time before or after the event in question. In order to ensure an appropriate target answer, we have implemented a simple screening function to search for key words like dates, times, and temporal relations. Arguments that do not pass this screen are excluded from the candidate targets for *When* questions.

A third scenario in temporal arguments is those that express a period of time, such as "for 10 years" or "until September". In these cases, the questions make much more sense when asked as "For how long... ?" or "Until when... ?". We have added several new question types to account for these situations.

Who Questions The current release of the QGSTEC system generates *Who* questions on targets identified by the named entity recognizer as a person, which leaves a limited number of potential answer targets. There are many instances when a *Who* question would be appropriate in the absence of a named person. For example, suppose we have the following input sentence:

> Traditionally, computational linguistics was performed by computer scientists.

If the target answer is "computer scientists", the current system would output the question:

> What was computational linguistics traditionally performed by?

Clearly, it makes more sense to ask a *Who* question in this scenario instead of a *What* question. In order to solve this problem, we need a way of distinguishing between when a target answer refers to a person or an object. One method we have begun to explore is using Wordnet hierarchies [Fellbaum, 1998].

As a preliminary test, we selected 13 Wikipedia articles and extracted all sentences that contained words whose primary synset is a hyponym of the "person" synset. In total, 74 target answers were discovered, and 58 of those targets were good candidates for *Who* questions. There were 16 cases where the target answer was a hyponym of "person", but it was not a good candidate for a *Who* question. In these cases, the target word identified was either not used in the primary sense, or the part of speech tag was incorrect.

Based on these analyses, we believe that using Wordnet hierarchies for classifying question types could prove useful; however, the method would be contingent upon accurate word sense annotation.

Conclusions and Future Work

We have explored three major areas of problems in question generation: *Content Selection*, *Grammaticality*, and *Wh Word Choice*. So far, we have developed several rules-based approaches to improving single-sentence QG, as well as two unconventional methods for ranking questions. Going forward we hope to implement a procedure for evaluating the language model's accuracy in measuring grammaticality. We also plan to further develop the language model implementation by using a larger annotated background corpus and more advanced statistical analysis. Secondly, we would like to incorporate Wordnet hierarchies to improve *Who* questions and explore other potential applications of this database within QG. Finally, we would like to expand the number of question templates. One example includes a template for quantitative questions, which would involve a function to identify candidate numerical targets and the appropriate type of question to ask.

Acknowledgments

We would like to thank Annie Louis for her help with the Topic scoring metric. This work was partially supported by NSF grant IIS-07-05671 (Exploiting and Exploring Discourse Connectivity: Deriving New Technology and Knowledge from the Penn Discourse Treebank.

References

Fellbaum, C. 1998. Wordnet: An electronic lexical database.

Heilman, M., and Smith, N. A. 2009. Question generation via overgenerating transformations and ranking.

Hovy, E., and Lin, C. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics*.

Louis, A. 2010. Topic words tool. http://www.cis.upenn.edu/ $\sim$ lannie/topicS.html.

Mannem, P.; Prasad, R.; and Joshi, A. 2010. Question generation from paragraphs at upenn: Qgstec system description.

Pradhan, S.; Ward, W.; Hacioglu, K.; Martin, J. H.; and Jurafsky, D. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting*.

Ratinov, L., and Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Annual Conference on Computational Natural Language Learning*.

Rus, V.; Wyse, B.; Piwek, P.; Lintean, M.; Stoyanchev, S.; and Moldovan, C. 2010. Overview of the first question generation shared task and evaluation challenge. In *Proceedings of QG2010: The Third Workshop on Question Generation*. http://www.questiongeneration.org.

Shen, L., and Joshi, A. K. 2008. LTAG dependency parsing with bidirectional incremental construction. In *Proceedings of EMNLP*.