School of Computer Science

Doctoral Thesis
in the field of
Language Technologies

Automatic Factual Question Generation from Text

Michael Heilman

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

ACCEPTED:

Thesis Committee Chair — Date 4-26-2011

Department Head — Date 9/23/11

APPROVED:

Dean — Date 9/29/2011

Automatic Factual Question Generation from Text

Michael Heilman

CMU-LTI-11-004

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
Vincent Aleven, Carnegie Mellon University
William W. Cohen, Carnegie Mellon University
Lori Levin, Carnegie Mellon University
Diane J. Litman, University of Pittsburgh
Noah A. Smith (chair), Carnegie Mellon University

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies*

Abstract

Texts with potential educational value are becoming available through the Internet (e.g., Wikipedia, news services). However, using these new texts in classrooms introduces many challenges, one of which is that they usually lack practice exercises and assessments. Here, we address part of this challenge by automating the creation of a specific type of assessment item.

Specifically, we focus on automatically generating factual WH questions. Our goal is to create an automated system that can take as input a text and produce as output questions for assessing a reader's knowledge of the information in the text. The questions could then be presented to a teacher, who could select and revise the ones that he or she judges to be useful.

After introducing the problem, we describe some of the computational and linguistic challenges presented by factual question generation. We then present an implemented system that leverages existing natural language processing techniques to address some of these challenges. The system uses a combination of manually encoded transformation rules and a statistical question ranker trained on a tailored dataset of labeled system output.

We present experiments that evaluate individual components of the system as well as the system as a whole. We found, among other things, that the question ranker roughly doubled the acceptability rate of top-ranked questions.

In a user study, we tested whether K-12 teachers could efficiently create factual questions by selecting and revising suggestions from the system. Offering automatic suggestions reduced the time and effort spent by participants, though it also affected the types of questions that were created.

This research supports the idea that natural language processing can help teachers efficiently create instructional content. It provides solutions to some of the major challenges in question generation and an analysis and better understanding of those that remain.

Acknowledgements

In addition to the dissertation committee, the author would like to acknowledge the following people for helping him conduct this research: the reviewers of the publications related to this work, for their helpful comments; Brendan O'Connor, for developing the treeviz parse visualization tool, for his work on the ARKref tool, and for ideas about information extraction; Nathan Schneider, Dipanjan Das, Kevin Gimpel, Shay Cohen, and other members of the ARK research group, for various helpful discussions; Justin Betteridge for discussions about connections between question generation and information extraction; Nora Presson, Ruth Wylie, and Leigh Ann Sudol, for discussions about the user interface and user study; Maxine Eskenazi and Matthias Scheutz, for being his previous advisors; Howard Seltman, Nora Presson, and Tracy Sweet, for statistical advice; the organizers of the question generation workshops, for enabling many thoughtful discussions about the topic; Jack Mostow, for asking many good questions; and Jill Burstein, for providing links to some of the texts used in this work.

The author would also like to acknowledge partial support from the following organizations: the Siebel Scholars organization; the National Science Foundation, for a Graduate Research Fellowship awarded to the author and for grant IIS-0915187, awarded to Noah Smith; and the U.S. Department of Education's Institute of Education Sciences, for grant R305B040063 to Carnegie Mellon University and its Program for Interdisciplinary Education Research (PIER) directed by David Klahr and Sharon Carver. The views expressed in this work are those of the author and are not necessarily those of the above organizations.

Finally, the author would like to thank his parents and his wife for their unflagging support over many years.

Contents

Chapter 1

Introduction

How would someone tell whether you have read this text? They might ask you to summarize it or describe how it relates to your own research. They might ask you to discuss its strengths and weaknesses or to compare it to another paper.

Or, as a first step, just to see if you bothered to get through it, they might ask you what features we used in our ranking model, or what corpora we tested on. That is, at first, they might just ask you questions to check that you remember the basic facts. Then, if it is clear that you read more than the title and abstract, they might move on to more challenging questions.

Of course, you are probably a highly skilled and motivated reader, and there would be no need to assess whether you read and retained basic factual information. However, that is not the case with all readers. For example, an elementary school teacher might ask his or her students basic questions since they are still learning to read.

Generating such questions, and authoring reading assessments more generally, can be a time-consuming and effortful process. In this research, we work toward automating that process. In particular, we focus on the problem of automatically generating factual questions from individual texts.

We aim to create a system for question generation (QG) that can take as input an article of text (e.g., a web page or encyclopedia article that a teacher might select to supplement the materials in a textbook), and create as output a ranked list of factual questions. A user could then select and revise these questions in order to create practice exercises or part of a quiz to assess whether students read the text and retained knowledge about its topic.

We focus on QG about informational texts—that is, non-fiction texts that convey factual information rather than opinions. While QG about narratives and subjective essays would also be interesting and educationally relevant, we leave these problems to future work. To make our factual QG system generally useful, we avoid the use of domain-specific knowledge (e.g., about historical events or geographical locations) and instead focus on modeling fairly general lexical and syntactic phenomena related to questions and the presentation of factual information.

1.1 Illustrative Example of Factual Question Generation

In this section, we provide examples that illustrate that QG about explicit factual information is a challenging but still feasible task given current natural language processing (NLP) technologies.

We begin with a relatively straightforward example, taken from an *Encyclopedia Britannica Elementary Edition* article about the city of Monrovia.[1]

(1.1) ...Monrovia was named after James Monroe, who was president of the United States in 1822. In that year a group of freed U.S. slaves, sponsored by a U.S. society, started a new settlement on the continent of their ancestors. As more settlers arrived from the United States and from the Caribbean area, the area they controlled grew larger. In 1847 Monrovia became the capital of the new country of Liberia. ...

A number of acceptable factual questions can be generated from this sentence by analyzing its grammatical structure, labeling its lexical items with high-level semantic types (e.g., person, location, time), and then performing syntactic transformations such as subject-auxiliary inversion and WH-movement. From the first sentence, we can extract well-formed and specific questions such as the following:

(1.2) Who was president of the United States in 1822?

(1.3) When was James Monroe president of the United States?

(1.4) Who was Monrovia named after?

[1] We make use of a dataset of Encyclopedia Britannica texts from previous NLP research (Barzilay and Elhadad, 2003). We describe the dataset in more detail in Chapter 4.

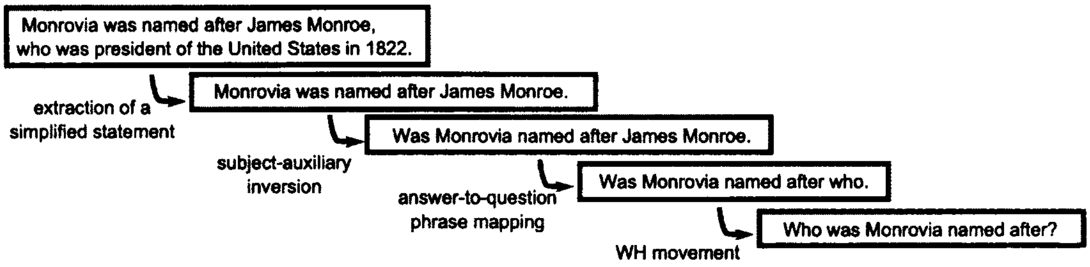
Monrovia was named after James Monroe, who was president of the United States in 1822.
extraction of a simplified statement
Monrovia was named after James Monroe.
subject-auxiliary inversion
Was Monrovia named after James Monroe.
answer-to-question phrase mapping
Was Monrovia named after who.
WH movement
Who was Monrovia named after?

Figure 1.1: An illustration of how simple lexical and syntactic transformations can convert a statement into a question. Chapter 3 describes this process in detail.

(1.5) What was named after James Monroe?

Figure 1.1 is a rough illustration of the sequence of transformations used to generate question 1.4 (we describe each step in detail later in Chapter 3). We can also generate acceptable questions such as the following from the last sentence:

(1.6) When did Monrovia become the capital of the new country of Liberia?

(1.7) What did Monrovia become in 1847?

(1.8) What did Monrovia become the capital of in 1847?

(1.9) What became the capital of the new country of Liberia in 1847?

The second sentence is more challenging, leading to unacceptable outputs such as the following:

(1.10) What did a group of freed U. S. slaves start in that year?

(1.11) What started a new settlement on the continent of their ancestors in that year?

Question 1.10 demonstrates the challenge of asking questions by taking information out of a larger discourse. Specifically, the phrase *that year* cannot be resolved to 1822 in the absence of the preceding sentence. Question 1.11 shows the challenge of creating appropriate question phrases—in this case, identifying that the noun phrase headed by the word *group* should lead to a *who* question, or perhaps a question starting with the phrase *what group*.

We return to these two challenges, along with others, in Chapter 2. Also, in §3.4 of Chapter 3, we present a statistical ranking model, trained from labeled data, that will help us avoid bad questions such as 1.10 while preserving good questions such as 1.4.

1.2 Instructional Content Generation

Many classrooms rely almost exclusively on textbooks as a source of reading materials, either for teaching reading itself or for teaching other topics in areas such as social studies and science. While textbooks have many advantages (e.g., they are well-edited, they contain lesson plans and exercises), they are a relatively small and expensive resource. The texts and exercises in a particular textbook may not match the needs of learners in a particular classroom—either the cognitive needs of learners at various skill levels, or the motivational needs of learners with various interests. For example, a textbook may not provide sufficient materials for themed units or specialized curricula. A teacher in Pittsburgh might want to find outside texts for a unit on the history of the city of Pittsburgh. Or, a teacher at a charter school with an environmentally-themed curriculum might want to find texts about renewable energy and global warming.[2] Also, certain instructional methods, such as the "Reader's Workshop,"[3] make heavy use of outside texts.

In contrast to textbooks, the Internet provides a vast source of texts (e.g., from digital libraries, news sites, Wikipedia, etc.). In theory, this source of texts could provide materials that better match students' skills and interests—and thereby supplement the resources provided by textbooks.

There are two major challenges, however, in making use of new texts:

- First, the Internet is not designed as an educational resource, and so most texts on the Internet are not likely to be useful for particular educational needs. Thus, finding pedagogically relevant texts can be difficult.
- Second, new texts are typically not accompanied by the sorts of useful instructional content that textbooks provide (e.g., practice exercises, assessments, and lesson plans, etc.).

Regarding the first challenge, tailored text retrieval applications could make it easier for teachers

[2] These two examples are taken from some discussions with teachers about situations in which they use outside texts.
[3] See `http://www.readersworkshop.org/` for details on the Reader's Workshop.

to find texts with specific topics, reading levels, etc. Recent research has explored such applications (Brown and Eskenazi, 2004; Miltsakaki and Troutt, 2008; Heilman et al., 2008).

In this work, we focus on the second challenge—or, at least, a small part of it. By developing a tool to automatically generate questions about new texts, we are facilitating the creation of instructional content such as practice exercises and assessments.

1.3 Types of Questions

In this work, we focus primarily on QG about explicit factual information (as in example 1.4). We now more precisely define the types of questions we aim to generate.

There are many types of questions, and researchers have proposed various taxonomies for organizing them. One particularly useful and concise discussion of the dimensions by which questions can be classified is provided by Graesser et al. (2008). They discuss how questions can be organized by the following characteristics: their purpose, the type of information they seek, their sources of information, the length of the expected answer, and the cognitive processes they involve. This work addresses a small area of that space, as follows.

1.3.1 Purpose

Following Graesser and Person (1994), Graesser et al. (2008) enumerate the purposes of questions: the correction of knowledge deficits (e.g., sincere information-seeking questions such as *How do I get to that restaurant?*), the monitoring of common ground (e.g., a science teacher asking, *Do mammals lay eggs?*), the social coordination of action (e.g., *Would you hand me that piece of paper?*), and the control of conversation and attention (e.g., *How are you doing today?*). This work focuses on monitoring common ground by assessing the knowledge that a student possesses about a particular topic.

1.3.2 Type of Information

Graesser et al. (2008) propose a set of 16 categories for questions based on the type of information involved, ranging from simple to complex questions. These categories were derived from previous work by Lehnert (1978) and Graesser and Person (1994). We focus on the simple end of the spectrum, with the goal of achieving a system that is scalable to large data sets and new domains. A system for generating more complex questions would be possible, but it would likely require encoding significant human knowledge about the targeted domain and the types of questions.

Specifically, this work aims to generate two types of questions: concept completion questions and, to a lesser extent, verification questions. Concept completion questions elicit a particular information that completes a given partial concept or proposition (e.g., *Who served as the first Secretary of State under George Washington?*). Verification questions invite a yes-no answer that verifies given information (e.g., *Was Thomas Jefferson secretary of state?*). For comparison, other question types include example questions (e.g., *What is an example of an important document written by Thomas Jefferson?*), goal orientation questions (e.g., *Why did Thomas Jefferson support the Embargo Act of 1807?*), and judgmental questions (e.g., *Were Thomas Jefferson's domestic policies successful?*).

1.3.3 Source of Information

This work aims to generate questions for which the source of answer is the literal information in the text—specifically information that is focused at the sentence or paragraph level rather than spread across the whole document. The questions will not address world knowledge, common sense, opinions of the answering party, etc.

1.3.4 Length of the Expected Answer

This work mostly involves questions whose expected answers are short, usually a single word or short phrase. The expected answers are not essays, for example.

1.3.5 Cognitive Processes

The questions generated by this work mainly assess recognition and recall of information—and, to a much lesser extent, comprehension. Very little inference, application, synthesis, or other complex processing is involved. Thus, in terms of Bloom's (1956) taxonomy of education objectives, we are mainly focused at the "knowledge" level. Higher levels of that taxonomy include "comprehension" (e.g., restating something in one's own words), "application" (e.g., using an equation to solve a practical problem), "analysis" (e.g., recognizing errors), "synthesis" (e.g., writing an essay), and "evaluation" (e.g., selecting the best design for an engineering task).

Clearly, we are focusing on a small part of the space of possible questions. In Chapter 2, we elaborate on why this is still quite a challenging task—and one with connections to existing work in parsing, entity recognition, coreference, and other research areas in computational linguistics. We also discuss some of the challenges of going beyond this small subspace and suggest possible pathways for future QG research.

1.4 Educational Value of Factual Questions

By focusing on explicit factual information, we are restricting the range of useful questions we might be able to produce. Factual questions make up only a small part of the questions used by educators in practice exercises and assessments. For example, educators also ask questions to assess the ability of students to make inferences, as well as their ability to perform various reading strategies such as summarizing or making connections to prior knowledge (National Institute of Child Health and Human Development, 2000, Chapter 4, Part II). However, automatic factual QG may still have the potential to help teachers create instructional content.

Let us first consider the relative benefits of "deeper" inference questions and "shallow" factual questions of the sort we aim to automatically generate. Many studies have explored whether asking higher level, deeper questions leads to better learning than lower level factual questions that focus on recognition and recall. Redfield and Rousseau (1981) and Winne (1979) provide meta-reviews. Most relevant studies focus on primary and secondary education levels. They involve either training teachers to ask high- or low- level questions (and then allowing them to ask what they want to), or explicitly asking teachers to ask certain types of questions. These studies focus primarily on verbal questions asked during class, but are still somewhat relevant to our research on text-specific QG.

While some studies show relatively strong advantages for questions that address higher cognition—several such studies are discussed by Redfield and Rousseau (1981)—many studies find no significant difference between asking shallow and deep questions. In a review of research on question asking, Samson et al. (1987) found that 88% of the 53 studies they reviewed yielded no significant differences in measures of learning between groups of students that were either asked primarily high-level questions or primarily low-level questions. Some studies even find favorable results for shallow, factual questions (Winne, 1979), and Ryan (1973) suggests that lower level questions may be more useful for lower level students.

Deep and shallow questions may complement each other: lower-level questions ensure that students possess the basic knowledge needed to answer higher-level questions. Higher-level questions ask students to manipulate various pieces of previously acquired information in order to formulate an answer. However, such questions implicitly assume that students have already acquired the basic information to be manipulated by higher level cognition—which may not be the case. As stated by Walsh and Sattes (2004), "students must be able to retrieve information if they are to use it in more cognitively complex operations." Initial lower level questions may be useful for ensuring that students grasp necessary factual knowledge before proceeding to higher level questions. Thus, Willen and Clegg (1986) recommend employing both low- and high-level questions.

Another relevant piece of evidence is that teachers tend to ask many lower-level, shallow questions—in fact, some research indicates that as few as 20% of teacher questions require high-level thinking (Gall, 1984). Even if teachers do ask an overly high proportion of many shallow questions, as argued by Graesser and Black (1985), it seems reasonable to infer from their extensive use that shallow questions have considerable educational value.

Factual questions are also prevalent in written tests and assignments. For example, Ozuru et al. (2008) analyzed questions on the grades 7–9 version of the popular Gates-MacGinitie Reading Test.[4] They found that about a quarter were "text-based" questions, defined as questions that match sentences in the source text nearly verbatim and could be achieved by applying simple transformations—broadly similar to the transformations that we discuss in Chapter 3. They also found that another quarter of the questions involved somewhat more complex transformations such as rewording and syntactic restructuring. A few of these types of transformations, such as the extraction of information from embedded clauses and the resolution of pronouns, are captured by our approach (§3.2 of Chapter 3). Further extensions to our approach could improve its coverage of such transformations. Of

[4] See `http://www.riversidepublishing.com/products/gmrt/` for more details on the Gates-MacGinitie Reading Test.

course, Ozuru et al. (2008) did find that about half of the Gates-MacGinitie questions were bridging or inferential questions—the sorts of questions that are beyond the scope of this work, and perhaps beyond the reach of current NLP techniques. Ozuru et al. (2008) observe that many different levels of questions are needed when working with students of a wide range of ability levels. They also analyzed a version of the test for grades 10–12 and found fewer text-based questions (about 5%) and more bridging and inferential questions (about 75%), suggesting that as grade levels increase, deeper questions become more prevalent.

Asking shallow questions and focusing on factual information may be useful when deeper questions would be too difficult, particularly for struggling students dealing with challenging new material. Research suggests that teachers should pose questions that students have a good chance of answering correctly, in order to avoid discouragement and other negative motivational effects (del Soldato and du Boulay, 1995; Malone and Lepper, 1987), and when students are struggling with the text, simpler questions focusing on explicit factual information may be challenging enough.

An important point is that while human teachers are capable of generating either "deep" questions involving complex inference or "shallow" factual questions, automatic techniques are much more likely to be error prone when complex inference is involved than when it is not. On the other hand, automated QG tools may be capable of generating large sets of shallow questions very quickly and could help teachers to focus on generating good deep questions. Another important point is that exercises and assessments often consist of many types of questions: automatically generated factual questions could be complemented by manually generated deeper questions. Thus, our goal is not to fully automate the process of creating questions, but rather to assist teachers by freeing up their time and reducing their cognitive load. In Chapter 5, we test whether we can achieve this goal; specifically, we develop a QG tool and conduct a user study to test whether teachers can use the tool to create factual questions more efficiently than when they have to create them on their own.

1.5 Comparison to the Cloze Procedure

When automatically generating factual WH questions, there are a variety of challenges, as discussed in Chapter 2. For example, WH questions may be ungrammatical if the automatic parser used in a

QG system provides an incorrect syntactic analysis (§2.2.1). There exists at least one alternative type of reading assessment that is easier to automate and perhaps more effective for certain applications: namely, the cloze procedure (Taylor, 1953).[5]

To generate a cloze test, one takes a passage and replaces some of the words in the passage with blanks. There are various ways of choosing words to replace, the simplest being to choose every Nth word. The reader's task is then to identify the original words that fill in the blanks. Researchers have found that such cloze tests are effective for measuring first language reading comprehension (Bormuth, 1967; Rankin and Culhane, 1969) as well as second language ability (Oller, 1972). Cloze tests can also be automated without introducing many errors (see, e.g., Mostow et al., 2004) since they only require tokenization of an input text into words.[6]

One important difference between cloze questions and the WH questions we focus on is that a cloze question (i.e., a sentence with a blank in it) is usually presented to the student either in the context of the passage from which it came or immediately after the passage. That is, cloze questions for reading assessment are typically "open book" questions. In contrast, we aim to generate WH questions that could be used to assess recall of specific information at some point after the student read the text (i.e., in a "closed book" scenario).[7]

Of course, single sentence cloze questions could be presented outside of a source passage, but then automation becomes more challenging. Taking information out of context can lead to various issues such as unresolved pronouns or other types of reference. For example, consider the following cloze question.

(1.12) He then became the _____ President of the United States.

For a text where multiple U.S. Presidents are mentioned, a question like this would be vague, permitting multiple answers. Thus, for "closed book" applications, cloze questions, like factual WH questions, become challenging to generate automatically. Similar discourse-related issues are a ma-

[5] Cloze items go by various names such as "gap-fill," "fill-in-the-blank," and "sentence completion" questions.
[6] A free online tool for generating cloze questions is provided at `http://www.lextutor.ca/cloze/`.
[7] Factual WH questions could be used for "open book" applications as well, as in the work of Gates (2008).

jor challenge in generating factual WH questions, as discussed in §2.3 of Chapter 2.[8] Context also affects cloze questions for vocabulary assessment (Pino et al., 2008; Skory and Eskenazi, 2010) since a particular target vocabulary word (which would be blanked out) can only be identified by a student if a sufficiently informative context is provided.

A distinct advantage of WH questions over the cloze procedure is the naturalness of WH questions. WH questions are a type of language that people encounter in both spoken and written language, whereas filling in blanks in incomplete sentences is an artificial task. As such, WH questions can be used more readily in a wider range of applications. For example, WH questions could be used in text-based or spoken tutorial dialogue applications, where maintaining natural interactions with users is an important desideratum (Johnson et al., 2000)

Since WH questions and the cloze procedure are in a somewhat different space, and since their utilities may vary considerably depending on the application, we leave direct experimental comparisons to future work.

As a more general point, it seems that research on WH questions is more likely to lead to better understanding of how to automatically generate a wider variety of questions, including deeper ones and those involving other types of structural transformations (e.g., questions involving paraphrasing or textual inference). We discuss the potential for such extensions later in §6.2.2 of Chapter 6.

1.6 Prior Work on Intelligent Tools for Education

A driving motivation for this work is to create tools to help teachers generate instructional content. In this section, we discuss connections to related research on the use of technology and artificial intelligence in education. Later, §1.9 addresses related work on the specific problem of QG.

Many applications of computer technology for assisting teachers involve hardware, such as clicker devices for gathering student responses (Trees and Jackson, 2007) and smartboards (Smith et al., 2005), or standard productivity software such as Microsoft PowerPoint (Szabo and Hastings, 2000). In this work, we focus on creating a tool using techniques from artificial intelligence—specifically,

[8] In Chapter 3, we present an overgenerate-and-rank framework for QG. Such a framework could be adapted for cloze questions or other types of questions. For example, for cloze questions, one could replace the second stage for converting statements into questions with a stage for replacing words with blanks. One could also adapt the features set used in the statistical ranker in stage 3 to address characteristics of cloze questions.

from natural language processing.

The study of artificial intelligence in educational technologies, particularly of intelligent tutoring systems (Koedinger et al., 1997; Vanlehn et al., 2005; Woolf, 2008), is a small but growing field. Many intelligent tutoring systems can be seen as a tool for teachers in that they provide guidance and feedback while students work through practice exercises. In facilitating practice, tutoring systems allow teachers to focus their effort on other issues such as planning curricula and delivering instruction about new concepts. It is worth noting that much of the work on tutoring systems has focused on interactions between an individual student and the computer, and less research on how tutoring systems affect teachers—though there is definitely research on such issues, particularly work by Feng and Heffernan (2006) and Ainsworth (2004).

With respect to intelligent tutoring systems, research on authoring tools for tutoring systems is particularly relevant to this work (e.g., Koedinger et al., 2004; Razzaq et al., 2009). For example, Ritter (1998) describes an authoring tool for automatically parsing the text of an algebra word problem into a formal semantic representation that could be loaded into a cognitive tutor. While their task and techniques differ from what we explore in this work, they have the same high-level goal of generating instructional content from text. A major issue in the research on and development of intelligent tutoring systems—and instructional technology more generally—is the efficiency of content generation and development. For example, it has been estimated that for intelligent tutoring systems, approximately 200 hours of development are required per hour of instruction (Anderson et al., 1995). For tutoring systems with substantial amounts of text content, opportunities may exist for NLP applications like the one we describe to increase the efficiency of system development.

Some of the challenges of authoring content can be addressed by intelligently re-using work by human teachers. For example, Aleahmad et al. (2008) describe a crowd-sourcing approach to the problem of generating content. Focusing on math exercises, they propose creating an online repository of materials by eliciting contributions from teachers and other Internet users. Such content could potentially be used within a tutoring system, or perhaps more directly by classroom teachers. Note that in such an online system for sharing or creating content, automated techniques such as the QG system we describe here could provide "seed" content for contributors to select and revise.

There has also been work on automatically creating content in the area of computer-assisted language learning. For example, Meurers et al. (2010) describe a system that takes arbitrary texts as input and, with NLP technologies, highlights specific grammatical constructions and automatically creates grammar practice exercises. Also, Heilman et al. (2008) describe a system that uses natural language processing and text retrieval technologies to help English as a Second Language teachers find pedagogically appropriate reading practice materials (e.g., texts at an appropriate reading level) for intermediate and advanced language learners.

There has been considerable work on applying NLP to educational problems, but most applications deal with the analysis of student responses rather than the generation of instructional content. For example, tutorial dialogue systems (Litman and Silliman, 2004; Graesser et al., 2005; Boyer et al., 2009) use NLP to analyze students' dialogue moves and respond in pedagogically appropriate ways. Automated scoring technologies (Shermis and Burstein, 2003; Mohler and Mihalcea, 2009; Nielsen et al., 2008) grade student responses to essays or short answer questions. Systems for grammatical error detection (Leacock et al., 2010) analyze student writing to find errors involving prepositions, determiners, etc. And, finally, tools for analyzing discussion boards (McLaren et al., 2009) use NLP to provide teachers with efficient ways of monitoring discussions among large groups of students.

This work on automatic QG contributes to the literature on educational applications of artificial intelligence techniques by exploring a problem that combines various aspects of the related work described above. Relatively little past work has focused on NLP or AI tools for automatically generating instructional content and, in particular, how such tools would be used by educators. We explore such an application and, in Chapter 5, present a user study involving potential users (K-12 teachers).

1.7 Prior Work on Overgeneration-and-Ranking

Our approach to QG, described in more detail in Chapter 3, is based on an "overgenerate-and-rank" strategy. The system generates a large set of candidate questions using existing NLP tools and manually written rules. It then ranks candidate questions using a statistical model of question quality.

The general idea of ranking a system's output using statistical methods is nothing new and has been explored in various previous research. For example, Collins (2000) describes methods for re-