THE QUESTION GENERATION SHARED TASK AND EVALUATION CHALLENGE

Workshop Report

Sponsored by the National Science Foundation

EDITED BY:

VASILE RUS AND ARTHUR C. GRAESSER

The University of Memphis

TABLE OF CONTENTS

PREFACE

The *Workshop on the Question Generation Shared Task and Evaluation Challenge* (www.questiongeneration.org) has been a successful attempt to bring together researchers from various disciplines to discuss the fundamental aspects of Question Generation and to set the stage for future developments in this emerging area. The idea of the workshop was inspired by two previous meetings that were dedicated to shared task evaluations in natural language generation. One meeting was the NSF/SIGGEN Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, which was held in April 2007. The other meeting was Generation Challenges 2009, the umbrella event designed to bring together a variety of shared-task evaluation efforts that involve the generation of natural language. Shared tasks are systematic research exercises in which researchers or entire research groups collectively address one or more well-defined research tasks. The advantages of shared tasks are manifold: better monitoring of progress, shared resources, and community strengthening.

The Question Generation (QG) workshop had two major goals: (1) identify potential shared QG tasks and produce a 5-year plan of such tasks and (2) create a community of researchers who investigate QG mechanisms and shared tasks. While the first goal proved to be too ambitious, the attempt provided the necessary inspiration for participants to reflect on proposed specific QG tasks. The second goal was clearly achieved by bringing together 28 participants of diverse backgrounds from both academia and industry (Microsoft Research Asia and Yahoo! Research United Kingdom). The workshop was a world-class event attended by researchers from all over the world: Asia (China), Americas (Brazil, Canada, The United States), and Europe (Germany, Romania, Spain, UK). Importantly, this was an interdisciplinary enterprise with attendees coming from the fields of Cognitive Science, Computational Linguistics, Computer Science, Discourse Processing, Educational Technologies, and Language Generation. There was a keynote presentation by Marilyn Walker, followed by 19 paper presentations. Walker's illuminating keynote presentation advanced the idea of Question Generation being more broadly construed as a discourse processing task rather than being limited to a language generation task, per se. The subsequent talks were divided into four major groups: systems, tasks, resources, and guidelines. The *systems* talks presented existing systems that generate questions whereas the *tasks* talks proposed concrete QG shared tasks and subtasks. *Resources* talks explored the resources that are already available or could be created for use in future QG shared tasks. The *guidelines* talks proposed general principles that should shape the development of a long-term plan on QG shared tasks and evaluation.

The afternoon of the first day had breakout sessions for brainstorming and summarizing ideas along four dimensions: guiding principles for long-term plans on QG shared tasks and evaluation (*longitudinal* group), concrete and generic QG tasks (*tasks* group), data collection and annotation (*data* group), and evaluation methodologies and metrics (*evaluation* group). On the second day the groups presented their recommendations. Post-workshop plans were announced, which included the suggestion to have a follow-up workshop and a workshop report. The attendees decided that a follow-up workshop would be held in conjunction with the International Conference on Artificial Intelligence in Education (AIED 2009).

This workshop report contains one chapter from each of the four working groups: longitudinal, tasks, data, and evaluation. Due to the early stages of development of this area of Question Generation, the general feeling at the workshop was that the proposal of concrete shared tasks (and a corresponding long-term 5-year plan) should be postponed until the follow-up workshop at AIED 2009. The follow-up workshop has the goals of identifying specific shared tasks and the resources that are needed for running these tasks.

Vasile Rus and Arthur Graesser, Editors

Institute for Intelligent Systems, University of Memphis

February 1st, 2009 - Memphis, TN

ACKNOWLEDGMENTS

We would like to thank all the participants for their contributions to the workshop and to our sponsor Tanya Korelsky from the National Science Foundation for enthusiastically supporting our Question Generation enterprise. We appreciate the hard work of the workshop steering committee, which included James Lester, Jose Otero, Paul Piwek, and Amanda Stent. The steering committee members led the groups during the workshop and the writing of the chapters in this report. We are extremely grateful to Marilyn Walker, our keynote speaker and co-leader of the group concentrating on tasks. We also appreciate Jack Mostow and Rodney Nielsen who took on the roles of co-leaders of groups during the workshop and the writing of chapters in this report. Finally, thanks go to Donia Scott and Anja Belz from the language generation community for their support in the early stages of planning the workshop.

PARTICIPANTS

Kristy Elizabeth Boyer, North Carolina State University

Yllias Chali, University of Lerthbridge, Canada

Albert Corbett, Carnegie Mellon University

Daniel Flickinger, Stanford University

Corina Forascu, The Al. I. Cuza University, Romania

Donna Gates, Carnegie Mellon University

Art Graesser, The University of Memphis

Michael Heilman, Carnegie Mellon University

Kateryna Ignatova, Technical University of Darmstadt, Germany

Aravind Joshi, University of Pennsylvania

Tatiana D. Korelsky, National Science Foundation

James Lester, North Carolina State University

Chin-Yew Lin, Microsoft Research Asia

Mihai Cosmin Lintean, The University of Memphis

Jack Mostow, Carnegie Mellon University

Rodney D Nielsen, University of Colorado

Jose Otero, Universidad de Alcala, Spain

Juan Pino, Carnegie Mellon University

Paul Piwek, The Open University, UK

Rashmi Prasad, University of Pennsylvania

Vasile Rus, The University of Memphis

Natália Giordani Silveira, Federal University of Rio de Janeiro, Brasil

Amanda Stent, Stony Brook University

Lucy Vanderwende, Microsoft Research

Marilyn Walker, University of Sheffield, UK

CHAPTER 1: GUIDELINES FOR QUESTION GENERATION SHARED TASK EVALUATION CAMPAIGNS

LONGITUDINAL GROUP CHAPTER

ART GRAESSER (UNIVERSITY OF MEMPHIS)
JOSE OTERO (UNIVERSITY OF ALCALA)
ALBERT CORBETT (CARNEGIE MELLON UNIVERSITY)
DANIEL FLICKINGER (STANFORD UNIVERSITY)
ARAVIND JOSHI (UNIVERSITY OF PENNSYLVANIA)
LUCY VANDERWENDE (MICROSOFT RESEARCH)

ABSTRACT

The objective of this chapter is to provide a set of recommendations for initiating a multi-year program of research in Question Generation (QG). The chapter proposes some guiding principles for a 5-year campaign of shared tasks in QG, identifies specific research groups that could participate in the shared tasks, and lists possible sources of support for multiyear evaluations in QG.

1.1. WHY QUESTION GENERATION?

For the first time in history, a person can ask a question on the web and receive answers in a few seconds. Twenty years ago it would take hours or weeks to receive answers to the same questions as a person hunted through documents in a library. In the future, electronic textbooks and information sources will be mainstream and they will be accompanied by sophisticated question asking and answering facilities. As a result, we believe that the Google generation is destined to have a much more inquisitive mind than the generations that relied on passive reading and libraries. The new technologies will radically transform how we think and behave.

Applications of automated QG facilities are endless and far reaching. Below are listed a small sample, some of which are addressed in this report:

1. Suggested good questions that learners might ask while reading documents and other media.
2. Questions that human and computer tutors might ask to promote and assess deeper learning.
3. Suggested questions for patients and caretakers in medicine.
4. Suggested questions that might be asked in legal contexts by litigants or in security contexts by interrogators.
5. Questions automatically generated from information repositories as candidates for Frequently Asked Question (FAQ) facilities.

The time is ripe for a coordinated effort to tackle QG in the field of computational linguistics and to launch a multi-year campaign of shared tasks in Question Generation (QG). We can build on the disciplinary and interdisciplinary work on QG that has been evolving in the fields of education, the social sciences (psychology, linguistics,

anthropology, sociology), and computer science (artificial intelligence, human-computer interaction). The structure of any shared task on QG is also technically and logistically feasible, and would contain these components:

1. **Information sources**. There is a body of text information sources that may or may not be structured and may or may not be theoretically annotated by humans.
2. **Input text**. The input to the QG system may be a word, a set of words, a single sentence, a text, a set of texts, a stretch of conversational discourse, an inadequate question, and so on.
3. **Question Generation system**. The QG system operates directly on the input text, executes implemented QG algorithms, and consults relevant information sources. Very often there are specific goals that constrain the QG system.
4. **Processing goals**: There are specific goals that constrain the QG system: questions are generated in dependence of the system's goals. Also, questions' quality is directly dependent on the extent to which they fulfill these goals.
5. **Output questions**. These are the questions that the QG system generates.
6. **Evaluation of questions**. The quality of the generated questions is evaluated by multiple criteria, including the extent to which they meet purported goals.

One of the early tasks for a multi-year program of QG research is to define the structure of the shared tasks (Nielsen R. , 2008; Piwek, Prendinger, Hernault, & Ishuzuka, 2008; Rus, Cai, & Graesser, 2008).

1.2. Limitations in Human Question Asking Motivate Automated Question Generation

Available research has repeatedly confirmed that humans are not very skilled in asking good questions. Therefore they would benefit from automated QG systems to assist them in meeting their inquiry needs. This section reports some of the research that supports our claim that human question asking is extremely limited in both quantity and quality.

There is an idealistic vision that learners are curious question generators who actively self-regulate their learning. That is, they identify their own knowledge deficits, ask questions that focus on these deficits, and answer the questions by exploring reliable information sources. Unfortunately, this idealistic vision of intelligent inquiry is rarely met, except for the most skilled learners. Most learners have trouble identifying their own knowledge deficits (Hacker, Dunlosky, & Graesser (Eds), 1998) and ask very few questions (Dillon, 1990; Good, Slavings, Harel, & Emerson, 1987; Graesser & Person, 1994). Graesser and Person's (1994) estimate from available studies revealed that the typical student asks less than 0.2 questions per hour in a classroom and that the poverty of classroom questions is a general phenomenon across cultures. The fact that it takes several hours for a typical student to ask one question in a classroom is perhaps not surprising because it would be impossible for a teacher to accommodate 25-30 curious students. The rate of question asking is higher in other learning environments (Graesser, McNamara, & VanLehn, 2005). For example, an average student asks over 26 questions per hour in one-on-one human tutoring sessions (Graesser & Person, 1994) and 120 questions per hour in a learning environment that forces students to ask questions in order to access any and all information (Graesser, McNamara, & VanLehn, 2005; Graesser, Langston, & Baggett, 1993). Computer-based information systems therefore can accommodate an increase in user questions by 2-3 orders of magnitude compared with classroom settings. This can only occur, of course, if the computer environments have adequate QG and question answering facilities.

The quality of the questions is important in addition to quantity (Graesser, Ozuru, Y., & Sullins, 2009; Scardamalia & Bereiter, 1992; Vanderwende, 2008). For example, training learners to ask *deep* questions (such as *why, why not, how, what-if, what-if-not*) is desired if we want the learner to acquire difficult scientific and technical material that taps causal mechanisms. The comparatively *shallow* questions (*who, what, when, where*) are often asked by students and instructors, but these shallow questions do not tap causal structures. It is somehow easy to generate shallow questions with current NLG models in computational linguistics (Sag & Flickinger, 2008). The automated generation of deep questions (such as why questions) might also be within the grasp of NLG research if the information sources are adequately annotated by causal relations (Prasad & Joshi, 2008). An indiscriminate generation of why-questions from text would presumably not be sufficient because some why questions do not elicit illuminating content. In fact, questions' quality is ultimately dependent on questioner's goals. A multi-year QG research plan may start with a tentative, absolute measure of questions' quality, such as the causal or non-causal character of questions. Then, a relative measure of quality, dependent on the extent to which the questions help in achieving the systems' goals, may be established. *The generation of the full landscape of question categories (both shallow and deep) should be part of a multi-year QG research plan*.

One of the key predictors of deep questions during inquiry is the existence of goals, tasks, or challenges that place someone in *cognitive disequilibrium*. Learners face cognitive disequilibrium when they encounter obstacles to goals, anomalies, contradictions, disputes, incompatibilities with prior knowledge, salient contrasts, obvious gaps in knowledge, and uncertainty in the face of decisions (Chinn & Brewer, 1993; Collins, 1988; Festinger, 1957; Flammer, 1981; Graesser & McMahen, 1993) (Otero, in press; Otero & Graesser, 2001; Otero, Ishiwa, & Vicente, 2008; Schank R. , 1999). Graesser and his colleagues have developed a cognitive model of question asking called PREG (Graesser & Olde, 2003; Otero & Graesser, 2001; Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005) that embraces cognitive disequilibrium in its foundation. The PREG model has a set of rules that predict the particular questions that readers should ask on the basis of the characteristics of the text, the type of disequilibrium, the reader's background knowledge, and metacognitive standards of comprehension (Otero & Graesser, 2001). *The generation of the good questions from texts in academic settings should be part of a multi-year QG research plan*.

Most students and adults have a long way to go before they acquire the skills of asking good questions. They can learn how to ask good questions through direct training, by observing outstanding inquiry that is modeled by experts, or by observing high-quality output from automated QG facilities. Given the poverty of human question asking, researchers in cognitive science and education have often advocated learning environments that encourage students to generate questions and that model good questions (Beck, McKeown, Hamilton, & Kucan, 1997; Collins, 1988; Edelson, Gordin, & Pea, 1999; Palinscar & Brown, 1984; Schank R. , 1999; Pressley & Forrest-Pressley, 1985). Moreover, improvements in the comprehension, learning, and memory of technical material can be achieved by training students to ask good questions during comprehension (Rosenshine, Meister, & Chapman, 1996). The training can be provided by either humans or by intelligent tutoring systems that model good questions (Graesser, McNamara, & VanLehn, 2005).

Animated pedagogical agents provide a promising learning environment for modelling good question asking and inquiry skills. These agents have become increasingly popular in recent advanced learning environments (Atkinson, 2002; Baylor & Kim, 2005; Graesser, et al., 2004; McNamara, Levinstein, & Boonthum, 2004; Reeves & Nass, 1996; Johnson, Rickel, & Lester, 2000). These agents could help the students learn either by modelling good inquiry by having two or more agents interacting with each other (Craig, Gholson, Ventura, & Graesser, 2000) or by holding a conversation directly with the student (Graesser, et al., 2004; Graesser, McNamara, & VanLehn, 2005; VanLehn, Graesser, Jackson, Jordan, Olney, & Rose, 2007). The agents may take on different roles: mentors, tutors, peers, players in multiparty games, or avatars in the virtual worlds. One recent system with agents, called iDRIVE (*Instruction with Deep-level Reasoning questions In Vicarious Environments* (Craig, Sullins, Witherspoon, & Gholson, 2006), was

designed with the explicit goal of modeling the asking of deep questions during learning. Student agents ask deep questions and the tutor agent answers them. The iDRIVE system has been shown to dramatically improve learning on science and technology topics, as well as improving student questions. However, the designers of iDRIVE had to handcraft the questions and the agent interactions. This technology could scale up tremendously if the questions were automatically generated.

Learners are not the only ones who experience limitations in QG skills. Other examples include:

1. Teachers in classrooms ask shallow questions over 90% of the time (Dillon, 1990; Graesser & Person, 1994).
2. Tutors have trouble generating good hints and prompts to get the student think, talk, and act in productive learning trajectories (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Corbett & Mostow, 2008; DiPaolo, Graesser, White, & Hacker, 2004). Tutors also need to ask good questions to assess how well the students learned and to troubleshoot specific deficits in knowledge and skill (Corbett & Mostow, 2008).
3. Questions on exams need to be tailored for deeper learning and more discriminating assessments of learning (Corbett & Mostow, 2008; Graesser, Ozuru, Y., & Sullins, 2009; Leacock & Chodorow, 2003).
4. Users of Google and other information retrieval systems tend to input only a few words rather than full sentences, queries, or sets of diagnostic words (Lin C. Y., 2008; Marciniak, 2008). Questions are often vague or underspecified. There is a need for guided question reformulation to more quickly retrieve answers.
5. Frequently Asked Question (FAQ) facilities are usually not frequently asked questions by clients, but rather they are a set of questions developed by the designers of the facility and only questions for which the designers have answers.

In conclusion, the quality of questions needs to improve for teachers, instructors, intelligent tutoring systems, information systems, and help facilities. *Therefore, methods of training individuals to improve the quality of questions through agents and automated facilities that generate questions should be part of a multi-year QG research plan*.

1.3. QUESTION QUALITY, COMPLEXITY, AND TAXONOMIES

An important initial step in a QG campaign is to take stock of the landscape of question categories so that researchers can specify what types of questions they have in mind, as well as the educational context (Rus, Cai, & Graesser, 2007). This section identifies some QG categories, taxonomies, and dimensions that might be considered in the QG campaign. The complexity and quality of the questions systematically vary across the broad landscape of questions. Finding criteria of question quality is a key requirement for good performance of QG systems. What we proposed in this section is merely a start. A theoretical analysis of a broad landscape of question categories and of the defining characteristics of quality questions should be part of a multi-year QG research plan.

Researchers in several fields have proposed schemes for classifying questions. Question taxonomies have been proposed by researchers who have developed models of question asking and answering in the fields of artificial intelligence (Lehnert, 1978; Schank R. C., 1986), computational linguistics (Harabagiu, Maiorano, & Pasca, 2002; Voorhees, 2001), discourse processing (Graesser & Person, 1994; Graesser, Person, & Huber, 1992), education (Beck, McKeown, Hamilton, & Kucan, 1997; Mosenthal, 1996) and a number of other fields in the cognitive sciences (for a review, see (Graesser, Ozuru, Y., & Sullins, 2009)).

SINCERE-INFORMATION SEEKING (SIS) QUESTIONS VERSUS OTHER TYPES OF QUESTIONS. Questions are not always generated by a person's knowledge deficits and cognitive disequilibrium. Graesser, Person, and Huber (1992)

identified four very different types of question generation mechanisms that occur in naturalistic settings. Whereas SIS questions are bona fide *knowledge deficit* questions, other types of questions address communication and social interaction processes. *Common ground* questions are asked when the questioner wants to establish or confirm whether knowledge is shared between participants in the conversation ("Did you say/mean oxygen?", "Are you understanding this?"). *Social coordination* questions are indirect requests for the addressee to perform an action or for the questioner to have permission to perform an action in a collaborative activity (e.g., "Could you graph these numbers?", "Can we take a break now?"). *Conversation-control* questions are asked to manipulate the flow of conversation or the attention of the speech participants (e.g., "Can I ask you a question?").

ASSUMPTIONS BEHIND QUESTIONS. Most questions posed by students and teachers are not SIS questions. Van der Meij (1987) identified 11 assumptions that need to be in place in order for a question to qualify as a SIS question.

1. The questioner does not know the information he asks for with the question.
2. The question specifies the information sought after.
3. The questioner believes that the presuppositions to the question are true.
4. The questioner believes that an answer exists.
5. The questioner wants to know the answer.
6. The questioner can assess whether a reply constitutes an answer.
7. The questioner poses the question only if the benefits exceed the costs.
8. The questioner believes that the respondent knows the answer.
9. The questioner believes that the respondent will not give the answer in absence of a question.
10. The questioner believes that the respondent will supply the answer.
11. A question solicits a reply.

A question is a non-SIS question if one or more of these assumptions are not met. For example, when a physics teacher grills students with a series of questions in a classroom (e.g., *What forces are acting on the vehicle in the collision?, What are the directions of the forces?, What is the mass of the vehicle?*), they are not SIS questions because they violate assumptions 1, 5, 8, and 10. Teachers know the answers to most questions they ask during these grilling sessions, so they are not modeling bona fide inquiry. Similarly, assumptions are violated when there are rhetorical questions (*When does a person know when he or she is happy?*), gripes (*When is it going to stop raining?*), greetings (*How are you?*), and attempts to redirect the flow of conversation in a group (a hostess asks a silent guest: *So when is your next vacation?*). In contrast, a question is a SIS question when a person's computer is malfunctioning and the person asks a technical assistant the following questions: *What's wrong with my computer? How can I get it fixed? How much will it cost?*

QUESTION CATEGORIES. The following 16 question categories were either proposed by Lehnert (1978) or by Graesser and Person (1994) in their analysis of tutoring. It should be noted that sometimes a question can be a hybrid between two categories.

1. *Verification*: invites a yes or no answer.
2. *Disjunctive*: Is X, Y, or Z the case?
3. *Concept completion*: Who? What? When? Where?
4. *Example*: What is an example of X?
5. *Feature specification*: What are the properties of X?
6. *Quantification*: How much? How many?
7. *Definition*: What does X mean?
8. *Comparison*: How is X similar to Y?

9. *Interpretation*: What is the significance of X?
10. *Causal antecedent*: Why/how did X occur?
11. *Causal consequence*: What next? What if?
12. *Goal orientation*: Why did an agent do X?
13. *Instrumental/procedural*: How did an agent do X?
14. *Enablement*: What enabled X to occur?
15. *Expectation*: Why didn't X occur?
16. *Judgmental*: What do you think of X?

Categories 1-4 were classified as simple/shallow, 5-8 as intermediate, and 9-16 as complex/deep questions in Graesser and Person's empirical analyses of questions in educational settings. This scale of depth was validated to the extent that it correlated significantly (r = .60 ± .05) with both Mosenthal's (1996) scale of question depth and the original Bloom's taxonomy of cognitive difficulty (1956). Although the Graesser-Person scheme has some degree of validity, it is an imperfect scale for depth and quality. For example, one can readily identify *disjunctive* questions that require considerable thought and reasoning, as in the case of the difficult physics question: *When the passenger is rear-ended, does the head initially (a) go forward, (b) go backwards, or (c) stay the same?* Generating an answer to this question requires a causal analysis, which corresponds to question categories 10 and 11, so this question may functionally be a hybrid question. But hybrid questions present a problem if we are trying to create a unidimensional scale of depth. *One task for the QG challenge is to formulate and test a categorization scheme that scales questions on depth as well as other dimensions of quality*.

OTHER DIMENSIONS OF QUESTIONS. Some other dimensions of questions are frequently addressed in classification schemes (Flammer, 1981; Graesser, Ozuru, Y., & Sullins, 2009; Nielsen R. , 2008; Voorhees, 2001).

1. *Information sources.* Does the answer come from a text, world knowledge, both, elsewhere?
2. *Length of answer*: Is the answer a single word, a phrase, a sentence, or a paragraph?
3. *Type of knowledge*: Is the knowledge organized as a semantic network, plan, causal structure, spatial layout, rule set, list of facts, etc.?
4. *Cognitive process*: What cognitive processes are involved with asking and answering the question? For example, using Bloom's taxonomy, do the cognitive processes involve recognition memory, recall of information, deep comprehension, inference, application of ideas, synthesis of information from multiple sources, comparison, or evaluation?

One early task in a multi-year plan for QG research is to settle on a theoretical analysis of questions that a community of researchers can work with.

1.4. CORPORA AVAILABLE FOR ANALYSIS

There are a number of corpora available in different research communities to conduct analyses of QG. These corpora can be analyzed in early years of the multi-year QG research plan. New data on QG can build on the available corpus base in order to concentrate on research questions that are targeted by the QG plan. We briefly enumerate some of these existing corpora, but this issue will be addressed in more detail in Chapters 2 and 3.

1. Naturalistic discourse samples in diverse conversational contexts that have been collected and distributed, such as the Linguistic Data Consortium, ARDA AQUAINT, Association for Computational Linguistics, and NIST. Many of these corpora are annotated by human experts or by computer systems that implement advances in computational linguistics.

2. Naturalistic corpora of human tutoring that have been collected and transcribed by researchers at a number of universities, including Carnegie Mellon University, University of Colorado, Illinois Institute of Technology, University of Memphis, University of Pittsburgh, Rhodes College, Stanford University, Vanderbilt University, and University of Wisconsin.
3. Student-computer tutorial interactions with intelligent tutoring systems, animated conversational agents, multiparty games, and other advanced learning environments with natural language conversation. There are computer logs with these interactions, as well theoretical annotations supplied by the computer. The institutions with these NLP-based learning environments are Carnegie Mellon University, University of Colorado, University of Edinburgh, Illinois Institute of Technology, University of Memphis, North Carolina State University, University of Pittsburgh, Stanford University, and University of Wisconsin.
4. Samples of questions that learners generate from texts in published studies in education, cognition, and instruction. Authors can be contacted for materials and distributions of questions.
5. QG corpora from commercial information retrieval facilities, such as Google and Yahoo.
6. Samples of questions on FAQ facilities in commercial applications (such as Microsoft), government agencies, universities, hospitals, and foundations. The questions and answers change over time, presumably because of feedback from users.
7. Multiple-choice and open-ended questions collected for examinations, including those from Educational Testing Service and the College Board. There are archives with questions that may be available to the public.

These do not exhaust the corpora available for analyses in the multi-year QG plan. Further examples of corpora can be found in the Chapter 3: Data.

1.5. Five-year Annual Selection and Sequencing of Shared Tasks

A 5-year plan for QG research would have a concrete roadmap of what tasks to concentrate on each year. There would be details about the task goals, corpora for training and testing, timetables, and criteria for evaluating QG performance. This section will not provide a concrete QG plan, but it will identify some of the considerations that will be addressed in subsequent chapters.

1. A theoretical analysis of questions, question generation, and answers. This includes a taxonomy, dimensions, quality criteria, and evaluation metrics that are theoretically justified. This should be completed in early years of the 5-year plan.
2. There will be an increase in the complexity and constraints of the questions selected in tasks over the years. For example, in early years there may be an inspection of any question generated in an environment. Then there would be a progression from shallow questions to causal questions; from questions about information in the explicit text to those that require inferences; from open-ended questions to multiple choice questions with good distracters; and from single texts to multiple texts as information sources.
3. There are many different dimensions of question complexity and quality to consider. These would need to be well specified in evaluations.
4. The plan in early years would capitalize on annotated corpora that exist in the NLP field.
5. Multiple languages and cultures should be accommodated in the plan.
6. Human experts and sometimes non-experts will need to evaluate the output of QG analyses. Advice will be needed from experts in application fields (e.g., medicine, law).
7. Students and postdocs will need to be involved in all stages of the QG plan.

8. There will be an evolution from in-house evaluations, to grassroots involvement from more than one research center, to formal evaluations by NIST.

1.6. FUNDING SOURCES

A number of funding sources have been identified to support the QG research plan.

1. National Science Foundation funds research in advanced learning environments, computational linguistics, and artificial intelligence.
2. Institute of Education Sciences funds education research, including computer technologies. There is an emphasis on assessing learning gains.
3. National Institutes of Health funds research on learning in children (NICHD), doctor-patient interaction, and information technologies to promote health.
4. The Department of Defense funds research on conventional computer systems, intelligent tutoring systems, multiparty games, communication technologies, interrogation techniques, and language translation.
5. Promising foundations are Spencer Foundation, Gates Foundation, James McDonnell Foundation, and the William and Flora Hewlett Foundation.
6. Relevant large corporations are Google, Yahoo, Microsoft, and Cisco.

1.7. CLOSING COMMENTS

Question generation is expected to play a more prominent role during learning in the age of Google, self-regulated learning, and complex electronic learning environments. However, inquiry learning requires a cooperative learning environment that exposes students to good questions because the quality and quantity of student questions is unspectacular at this point in history. There needs to be automated systems of QG that produce families of good questions in specific domains of knowledge so these questions can be modeled and guide the students. The field of computational linguistics, as well as companion sciences, is well positioned to build computational models of question asking and answering. Such efforts will only be realized when the researchers have a broad perspective on the landscape of possible questions and QG mechanisms. Chapter 1 provides a succinct sketch of the landscape.

CHAPTER 2: QUESTION GENERATION TASKS AND SUBTASKS

TASK GROUP CHAPTER

NATÁLIA GIORDANI SILVEIRA (FEDERAL UNIVERSITY OF RIO DE JANEIRO, BRAZIL)
JAMES LESTER (NORTH CAROLINA STATE UNIVERSITY, USA)
MIHAI LINTEAN (UNIVERSITY OF MEMPHIS, USA)
RASHMI PRASAD (UNIVERSITY OF PENNSYLVANIA, USA)
MARILYN WALKER (UNIVERSITY OF SHEFFIELD, UK)

ABSTRACT

The Text-to-Question task and the Tutorial Dialog task were identified promising candidates for shared tasks, along with an associated Evaluation Track. In the Text-to-Question task, a question generation (QG) system is given a text, and its goal would be to generate a set of questions for which the text contains answers. In the Tutorial Dialogue task, a QG system would be given a tutorial dialogue history and a target set of propositions, and its goal would be to generate a question that would elicit from the student an answer containing the propositions. These shared tasks would attract considerable interest from the NLP, NLG, Intelligent Tutoring System, and IR communities, and offer a much needed focus for synergistic interactions.

2.1. INTRODUCTION

Experience in the NLP community has shown that by identifying a common task (or set of tasks) and committing to a shared evaluation framework for gauging the performance of alternate approaches to that task, significant advances can be made that otherwise are difficult to achieve. A key component of designing a Shared-Task Evaluation Campaign (STEC) is selecting the task around which competitions will be organized: *defining a task and its associated evaluation metric in effect defines a STEC*.

At the workshop, a working group (the *Task Group*) was charged with proposing a set of tasks that could contribute to the formulation of a Question Generation STEC. In the course of identifying a set of tasks, the Task Group considered candidate tasks that were proposed in position papers (Corbett & Mostow, 2008; Gates, 2008; Lin C. Y., 2008; Nielsen R. , 2008; Ignatova, Bernhard, & Gurevych, 2008; Prasad & Joshi, 2008; Rus, Cai, & Graesser, 2008) (Smith, Heilman, & Hwa, 2008) and that emerged in workshop discussions. The resulting tasks were introduced to the workshop participants as a whole and there was an attempt to refine the task definitions. Ultimately, four question generation tasks were discussed: Text-to-Question, Tutorial Dialogue, Assessment, and Query-to-Question:

- *Text-to-Question*: In the Text-to-Question task, a question generation (QG) system is given a text, and its goal would be to generate a set of questions for which the text contains, implies, or needs answers.
- *Tutorial Dialogue*: In the Tutorial Dialogue task, a QG system would be given a tutorial dialogue history and a target set of propositions, and its goal would be to generate a question that would elicit from the student an answer containing the propositions.

- *Assessment*: In the Assessment task, a QG system would be given a text and, optionally, a dialogue, and its goal would be to select a concept, determine a question type, and generate a textual "assessment" question.
- *Query-to-Question:* In the Query-to-Question task, a Q-Gen system would be given a formal query, and its goal would be to translate the query into a canonical form of a natural language question.

Each of the tasks was viewed as a promising candidate for inclusion in a Question Generation STEC. It appeared that some of the tasks could be viewed as variants of one another, and that some can be used in the same families of applications. The Text-to-Question and the Assessment tasks are similar in that both take a text as input and yield a question (or set of questions) as output; both would be of interest to the NLU and NLG communities. The Text-to-Question and Query-to-Question tasks are similar in that both could be of interest to the IR community as pre-processing steps of an online Question Answering system. The most enthusiasm was expressed for the Text-to-Question and Tutorial Dialogue tasks, so these are elaborated in this chapter.

2.2. THE TEXT-TO-QUESTION TASK

The *Text-to-Question Question Generation Task* can be characterized as follows: given a text, the goal of a QG system performing the Text-to-Question Question Generation task would be to exhaustively create a set of Text-Question pairs, such that every possible question that could be generated would be included in the set (Figure 2.1).

Given:

- Text ***T***

Create:

- Text-Questions pairs {$\boldsymbol{P_1 \ldots P_n}$} each represented as a ($\boldsymbol{K_i}$, $\boldsymbol{Q_i}$) pair, where $\boldsymbol{K_i}$, the *target text*, indicates which text segment from ***T*** represents the answer and the $\boldsymbol{Q_i}$ represents a question that would elicit $\boldsymbol{K_i}$

FIGURE 2.1. TEXT-TO-QUESTION GENERATION TASK

Two alternative formulations of the task could be specified. The first alternative is that the text would be raw. The *raw* formulation of the task would require participants to submit QG systems that provided with end-to-end solutions. Such systems would by necessity include a full complement of NLP modules (e.g., tokenizers, POS taggers, parsers). These systems would first conduct whatever analyses its designers believed was appropriate before question generation proper was undertaken. The raw formulation would likely be preferred by teams that included significant NLP expertise, especially those who would not be supportive of particular representational commitments.

The second alternative is that the text would be annotated. In the *annotated* formulation of the task, QG systems would be given texts annotated with linguistic information. Annotations could range from low level information (e.g., POS, syntax) to higher level information (semantic role labels, discourse relations such as the temporal, conditional, causal, and contrastive relations of the Penn Discourse TreeBank). The annotated formulation would

likely be preferred by teams that endorsed (or at least accepted) the particular formalisms chosen. It would also benefit those whose teams might include the NLP expertise but nevertheless wished to focus on particular QG tasks downstream from the syntax and semantic analysis modules. For the annotated formulation, annotations could either be created manually or generated automatically by existing tools.

The performance of these QG systems could be evaluated either manually or automatically. Manual evaluation would make use of a panel of human judges who would compare system-generated pools of questions against human-generated pools of questions. Automated evaluation would require the creation of tools, which would be similar to Machine Translation evaluation tools such as ROUGE (Lin C. Y., 2004) and BLEU (Papineni, Ward, Roukos, & Zhu, 2002). The automated evaluation tools would be developed in later years of the competitions. Evaluation metrics would include IR-style measures such as precision and recall, as well as NLG metrics such as fluidity.

The Text-to-Question Question Generation task is promising for several reasons. It would likely attract the interest of a broad range of NLP researchers. The NLU community would be drawn (at least) to the text analysis problems, whereas the NLG community would be drawn (at least) to the text generation problems. The task is also of interest because it is "application neutral." The resulting systems could be applied to a multitude of problems ranging from online question answering to tutorial dialogue. For example, websites such as Yahoo!, Google, MSN Encarta, or Wikipedia could use the techniques as the basis for off-line preprocessors to question answering systems. Moreover, sub-tasks could range in complexity, with some addressing specific issues, e.g., one year the task might focus on causal consequence question generation. By preserving the language independence of the task, later sub-tasks could explore multi-lingual instantiations of the problem. Furthermore, several corpora that are appropriate for the task are already available (see Chapter 3 in the report), so start-up time would be minimal.

2.3. THE TUTORIAL DIALOGUE TASK

The *Tutorial Dialogue Question Generation Task* can be characterized as follows: given a dialogue history between a tutor and a student, the goal of a QG system performing the Tutorial Dialogue Question Generation task would be to create a question that would cause a student to produce an answer with information that would "cover" the propositions in the specified target set (Figure 2.2). The dialogue history could either be comprehensive, i.e., it could provide the full history from the beginning of the tutorial session, or perhaps even across multiple sessions, or it could be represented as a sliding window containing a small number of turns. Each turn would be marked with the appropriate speaker (tutor or student). Furthermore, like the text in the Text-to-Question task, the dialogue history could be raw, or it could be annotated. In the annotated formulation of the task, annotations could range from low level syntactic annotations to tutorial dialogue act annotations.

The performance of Tutorial Dialogue QG systems could be evaluated against different classes of individuals that range from the students themselves to expert tutors. We assume in this chapter, for purposes of illustration, that the expert tutors would serve as the gold standard. Given the same dialogue history and the same set of target propositions, a panel of expert tutors would generate answer-eliciting questions. The responses of Tutorial Dialogue QG systems would then be graded with respect to precision and recall on words, noun phrases, and question words used by the expert tutor panel. An alternative to the panel of expert tutors would be a panel of domain experts, which could be used either separately or in conjunction with the expert tutors.

The Tutorial Dialogue Question Generation task holds much appeal. *First*, it will be of significant interest to intelligent tutoring system researchers, which includes a thriving tutorial dialogue community. It embodies a problem that is solved daily by tutors and, consequently, that must be solved to develop natural language intelligent tutoring systems. Tutors regularly generate questions, so it is a reasonable hypothesis that intelligent

tutoring systems that can perform this task will be more effective. *Second*, it is well grounded. In contrast to some question generation tasks, the Tutorial Dialogue Question Generation task is grounded in the concrete interactions of naturally occurring dialogues. *Third*, it offers a means for systematically considering student characteristics. The expertise of the student could be included in the tutorial dialogue or provided as a parameter (a student model) in the task definition. *Fourth*, many tutorial dialogue corpora have been acquired. Some of these corpora could form the basis for the proposed task to the extent that they are amenable to precise formulation, whereas others might be used in an actual STEC. *Lastly*, the task as described here is analogous to the DARPA Communicator task (http://xml.coverpages.org/darpaCommunicator.html; accessed on March 30, 2009), which proved to be a landmark competition in the evolution of dialogue management research.

Given:

- Tutorial dialogue history ***H***
- Set of expected propositions ***P*** to be covered

Create:

- Question ***Q*** such that if answered by the student, it would induce the student to state ***P*** in the context of ***H***

FIGURE 2.2. TUTORIAL DIALOGUE QUESTION GENERATION TASK

Several issues would need to be considered in a more precise formulation. It would require an explicit representation of the knowledge to be communicated. Even if the requirement was not for a formal representation of the knowledge, it would nevertheless need to be encoded in structured units such as propositions, which can be a challenging task for ill-structured domains. It would also require a reference text that described the reference knowledge to be communicated. For some domains and tasks, such a reference text, particularly one geared to novices, might require effort to construct if it did not already exist. Questions arose as to whether there would in fact be a gold standard with respect to the generated questions. Would multiple tutors readily agree about the best questions to generate?

Several alternative formulations of the task should be entertained. The dialogue for the QG system could either be raw or annotated. Raw dialogue would be more challenging and would more accurately reflect "real-world" conditions, whereas dialogue annotated with tutorial dialogue acts could offer a more realistic first step in an evolving research program. As another alternative, the knowledge level of students could be provided as one of the inputs to the task. Novice students would be asked one type of question, intermediate students would be asked more difficult questions, and the most advanced students could be asked the most challenging questions. Finally, the generation of hints would be important. Generating customized scaffolding could constitute a subtask, with the task of faded scaffolding generation posing an even more challenging task.

2.4. DISCUSSION

Selecting appropriate tasks for a STEC is a non-trivial undertaking. The tasks should satisfy at least five criteria. *First*, they must be sufficiently challenging that they advance the state of the art, but they should not be too

difficult. Tasks that are too simple will not yield progress, and tasks that are too difficult will not produce advances in the field. *Second*, tasks must target core functionalities. Choosing a problem whose solution is critical to the advancement of the field will attract the interest of the greatest number of participants. *Third*, tasks should be decomposable. It should be possible to consider individual subtasks in isolation. *Fourth*, it should be possible to temporally distribute subtasks in a progression. Progressions should transition from simple to complex, with competitions incrementally introducing increasing complexity and requiring increasingly *broader* solutions. This resonates with one of the guiding principles outlined in the Chapter 1 by the *Longitudinal* group. *Lastly*, they should be readily evaluated. Evaluation should be cost effective. That is, it should be possible to devise or adapt automated evaluation tools. Alternatively, if human evaluators are used, it should not be too labor intensive to obtain their judgments). Evaluation metrics should be objective. Thus, if human evaluators are used, it should be possible to establish inter-judge reliability). The evaluation framework should not permit gaming the system, sophisticated guessing, and other design flaws that would compromise the quality of the analyses.

Both the Text-to-Question and the Tutorial Dialog tasks satisfy each of these criteria:

- **Complexity**. The Text-to-Question task is sufficiently challenging. In the *raw* formulation, it potentially involves myriad fundamental NLU and NLG problems. However, it is not overly difficult because of the simplifications afforded by the annotations. The Tutorial Dialogue task is sufficiently challenging because it too potentially involves the full gamut of NLU problems, although it can be specified in an annotated formulation that might be more appropriate for a team with less NLP expertise.
- **Core functionalities.** The Text-to-Question task would focus on key NLU and NLG problems, the solution of which would carry important implications for the field of computational linguistics and a broad range of applications. The Tutorial Dialogue task would also focus on key NLU and NLG problems, the solutions of which would significantly advance the state-of-the-art in intelligent tutoring systems.
- **Decomposability**. The Text-to-Question task offers decomposability with respect to an optional NLU set of tasks, whose decomposition is well understood, and a question-generation-specific set of tasks that could include target concept text span identification, question type selection, and realization. The Tutorial Dialog task is also decomposable. That is, like the NLU aspects of the Text-to-Question task, the NLU dialogue history analysis tasks offer a relatively standard decomposition, whereas the question-generation proper tasks could include topic identification, question type selection, and realization (see Nielsen, 2008).
- **Progression of tasks**. Both the Text-to-Question and the Tutorial Dialog tasks offer a progression of tasks with increasing complexity. Early competitions could focus on shorter, simpler texts (or shorter, simpler dialogues), a narrow range of question types, and more direct, literal questions. Subsequent competitions could address longer, more complex texts (or longer, more complex dialogue histories), a broader range of question types, and emphasize entailment and inference. Similarly, early competitions could target a *generic* reader (or student), while later competitions could target specific reader populations (or specific, e.g., novice or expert, students).
- **Evaluation**. Devising appropriate evaluation methodologies always poses significant challenges, but both the Text-to-Question and the Tutorial Dialog tasks appear to be candidates for cost-effective evaluation. Given the discussions at the workshop, perhaps some combination of manual and semi-automated methods can be devised, potentially by adopting (and adapting) evaluation techniques from the computational linguistics community. One possibility is introducing an Evaluation Track at the STEC, and using the best evaluation tools developed in one year to evaluate systems submitted in the subsequent year.

2.5. CONCLUDING REMARKS

Although much remains to be done to flesh out tasks for a Question Generation STEC, both the Text-to-Question task and the Tutorial Dialog task appear to be promising candidates. A multi-year parallel campaign with a Text-to-Question Track, a Tutorial Dialogue Track, and perhaps an Evaluation Track, could lead to significant theoretical and practical advances in question generation technologies. Such a STEC would attract considerable interest from the NLP, NLG, Intelligent Tutoring System, and IR communities, and offer a much needed focus for synergistic interactions.

CHAPTER 3: DATA REQUIREMENTS, SOURCES, AND ANNOTATION SCHEMES FOR QUESTION GENERATION SHARED TASKS

DATA GROUP CHAPTER

PAUL PIWEK (THE OPEN UNIVERSITY, UK)
JACK MOSTOW (CARNEGIE MELLON UNIVERSITY)
YLLIAS CHALI (UNIVERSITY OF LERTHBRIDGE, CANADA)
CORINA FORASCU (AL. I. CUZA UNIVERSITY, ROMANIA)
DONNA GATES (CARNEGIE MELLON UNIVERSITY)

ABSTRACT

This chapter introduces a framework for characterizing data for Question Generation as a basis for QG STECs. It offers an inventory of types of resources and some examples.

3.1. INTRODUCTION

This chapter summarizes and elaborates on the discussions in the Data group at the 2008 Workshop on the Question Generation Shared Task and Evaluation Challenge. The Data group set out to identify the requirements on data for Question Generation Shared Tasks. The group formulated a framework for capturing such requirements and came up with a preliminary inventory of sources for QG data. One of the principal concerns that we aimed to address involved taking into consideration the often divergent demands of the communities present at the workshop, in particular, Natural Language Understanding, Natural Language Generation, and Intelligent Tutoring Systems.

It became very clear at the workshop that Question Generation (henceforth QG) cannot be boiled down to a single task. Rather, there is a range of QG tasks, varying from generation of questions from plain text (Text-to-Question Generation) to generation of questions from user search queries (Query-to-Question Generation). Additionally, each task can typically be divided into a number of subtasks. It is likely that the diversity of tasks and subtasks is also reflected by the range of data and resources that will be required. In this chapter, we begin by setting out a generic framework for QG task and data representations (Section 3.2). We subsequently describe how this framework can help with identifying QG data. The focus will be on one specific type of QG, namely Text-to-Question generation. Our discussion of this task and its data will be based on a decomposition of the task into three subtasks (Section 3.3). The chapter ends with a number of concluding remarks (Section 3.4).

3.2. FRAMEWORK FOR TASK AND DATA REPRESENTATION

The overall challenge of question generation can be posed abstractly as follows: given a body of information, and criteria for good questions, generate questions that ask about the information and satisfy the criteria. For example, the information might include a source text, and the criteria might include how well the questions assess (or assist) the reader's comprehension of that text. Or the information might include the entire Web, and the criteria might include how well the questions represent the intent of users' queries. Or the information might include domain knowledge plus tutorial dialogue history, and the criteria might include how successfully the questions get the student to articulate particular concepts the tutor wants them to learn.

What constitutes useful training and evaluation data for QG tasks? Such data includes questions along with the information they ask about, their answers, and various attributes of the source, the question, and its answer. The following two definitions state more precisely how we view QG Data by introducing the notion of a QG Data Instance and a QG Data Representation. The definitions should be read as specifying logical models, not implementations (which can take many different shapes and forms).

> DEFINITION QG Data Instance: Given a Question Generation task, a data instance is a 6-tuple (S, A, Q, SA, AA, QA), where:
>
> - S is the source,
> - A is the target answer,
> - Q is the question, and
> - SA, AA and QA are attributes/annotations on the source, answer and question, respectively.
>
> DEFINITION QG Data Representation: Given a question generation task, the data for that task can be represented as a set of data instances:
>
> - $(S_1, A_1, Q_1, SA_1, AA_1, QA_1), \ldots, (S_k, A_k, Q_k, SA_k, AA_k, QA_k)$

Although each instance relates a question to a single answer, the framework does permit linking questions with several answers. This is achieved by having several instances for the same question, with each instance containing a different answer.

The task criteria can be viewed as predicates on the attributes. For example, some criteria for assessment could be based on attributes regarding reliability and validity of the questions. The criteria could state that the value for these attributes should be maximized. Another, much less demanding criterion could be that the question should be grammatical. Furthermore, criteria could look beyond a single generated question by requiring the generation of many different questions for a particular answer.

It is beyond the scope of this chapter to provide an exhaustive list of attributes for QG data. The list will vary depending on the specific type of QG. Here we provide some examples of possibly useful attributes, starting with attributes for the source (SA):

- input format (e.g. raw text, queries, tutorial dialogues, blogs or chats),
- parts of speech, syntax, word senses (according to a given classification, like WordNet),
- semantic labels (e.g., named entities, thematic roles),
- discourse structures (e.g., coherence relations/rhetorical structure), and
- language

Attributes of a target answer (**AA**) include:

- format of the expected answer: a snippet of text, a set of such snippets, an inference from the source text, or not available because the question is raised (Piwek, Prendinger, Hernault, & Ishuzuka, 2008) rather than answered by the source text, or the question has no answer in the source text,
- language, and
- syntax, semantics and discourse structure of the answer.

Attributes of a question (**QA**) include:

- length
- format (such as multiple choice, fill-in-the-blank, or free-form),
- topic, depth (such as literal or inferential),
- language,
- type (see Chapter 1 of this report), and
- style (such as its clarity, grammaticality, formality, and tone).

Other examples of attributes include the difficulty of a question for a given population, as well as the distribution of answers and response times, and the scores awarded to questions or answers, whether manually or automatically.

Various question generation processes may input some of these attributes and output others. For example, evaluation, whether manual or automatic, inputs the source text, target answer, question, and criteria, with outputs on how well the question satisfies the criteria.

3.3. TEXT-TO-QUESTION GENERATION

As explained in Section 3.2 above and in Chapter 2, a common Question Generation task is to input a given source text and output questions about it. The task of generating a question about a given text can be decomposed into three subtasks. First, given the source text, a content selection step is necessary to select a target to ask about, such as the desired answer. Second, given a target answer, select question type, i.e., the form of question to ask, such as a cloze or why question. Third, given the content, and question type, construct the actual question in a question construction step. These steps correspond roughly with what Nielsen (2008) calls Concept Selection, Question Type Determination, and Question Construction.

3.3.1. REQUIREMENTS THE TASK PUTS ON DATA

An important requirement for sharing this task is that researchers be free to focus on the task as a whole or any one of the subtasks with minimal or no commitment to a particular representation (e.g., syntax or semantics). This requirement dictates that we specify each subtask in terms of a lowest common denominator representation, namely natural language. Thus, instead of specifying the target for a question in some semantic representation that not everyone uses, we can just specify a correct answer in natural language. The simplest case is that the target is just a sentence in the text. In harder cases, the target answer is only implicit, and must be inferred from the text. We might even specify a target not given in the text, for example if we want to generate the sorts of questions we want graduate students to ask when they read a research paper.

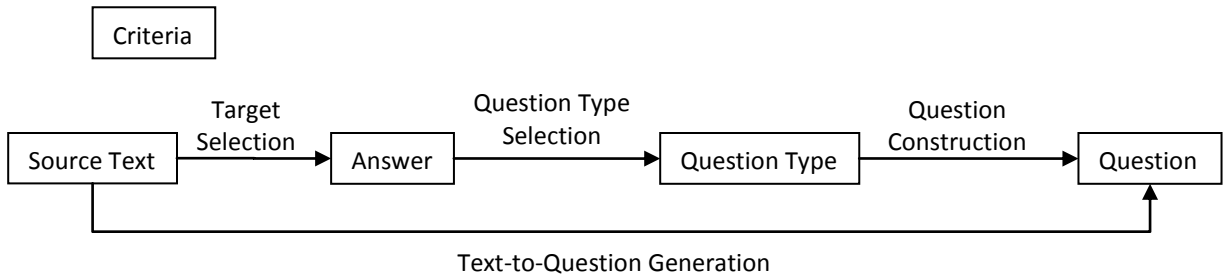

FIGURE 3.1: TEXT-TO-QUESTION GENERATION: THE TASK, SUBTASKS AND RESOURCE (IN BOXES)

	Text-to-Question (Full)	Content Selection	Q-Type Selection	Question Construction
Source	+	+	+	+/-
Answer	+/-	+	+	+
Question	+	-	-	-
Source Attributes	+/-	+/-	+/-	+/-
Answer Attributes	+/-	+/-	+/-	+/-
Question Attributes	+/-	+/-	+	+

TABLE 3.1: OVERVIEW OF RELEVANT PARTS AND ATTRIBUTES OF DATA INSTANCES FOR DIFFERENT SUBTASKS OF THE TEXT-TO-QUESTION GENERATION TASK. THIS TABLE CONCERNS THE REQUIREMENTS FOR *TRAINING DATA*. KEY: + = REQUIRED, +/– = OPTIONAL, – SUPERFLUOUS.

Although we should not force researchers to use a particular representation, we should allow and enable for them to do so. Indeed, many if not most relevant data resources employ particular annotation schemes, some of them quite widespread (though not necessarily universal) within the natural language research community. Thus, the shared task might include PDTB annotations (Prasad & Joshi, 2008) of the source text, but its use would be optional.

Figure 3.1 shows a schematic overview of the Text-to-Question generation task and its three subtasks. Let us assume that we have QG Data as described above consisting of instances of the form (S, A, Q, SA, AA, QA). Now, depending on whether we want to use the data as training or test data and depending on what task/subtask we want to focus on, we select specific parts and attributes of each of the instances. For instance, if we require training data for the question type selection subtask, we need the source text and answer (and optionally annotations on these), and also the question annotation, in particular, the bit which provides the question type. Note that we do not need the question itself for the training (and also test) data for this particular subtask. In contrast, if we're dealing with question construction, we primarily need the answer and question annotation (specifically, the question type information), which are the inputs to this task, and also the question itself, which is the output. In Table 3.1 we summarize which parts of the data instances are required (+), optional (+/–), or superfluous (–) for training of the various subtasks of Text-to-Question generation.

3.3.2. RESOURCES SUITABLE FOR THE TEXT-TO-QUESTION GENERATION TASK

An ideal resource for question generation would provide us with data that has information for all the parts and attributes of our data instance representations (S, A, Q, SA, AA, QA). In reality, such resources are sparse, and

typically we will need to supplement the information present in the resource, either through handcrafted or automated additions. For instance, generic Computational Linguistics resources will normally provide us with annotated text, but the questions that this text is about will need to be added manually. Here we will distinguish five broad categories of resources:

1. **Generic Computational Linguistics Resources**. Such resources will typically consist of considerable amounts of text and corresponding annotations. The domain may be restricted, e.g., many annotations in Computational Linguistics work with the Wall Street Journal (WSJ). This includes, for example, the Penn Discourse TreeBank (PDTB) which is a subset of the WSJ annotated with discourse relations. Another issue with this type of resource is that they typically lack answers and questions. Thus, they could be used as source text, but they would need to be enriched with questions and answers. An exception is the PDTB; a subset of the PDTB has already been enriched with handcrafted why-questions (Verberne, Boves, Coppen, & Oostdijk, 2007).
2. **(Open) Learning Resources**. These include all the parts that we require for a data instance (Source, Answer and Question), but may lack linguistic annotations. Examples are items on standardized tests, test item banks, and end-of-chapter questions. Several universities have made their teaching and learning resources available for open access. These resources constitute a potentially rich source of data. A prominent example in the United States is the MIT OpenCourseWare site (http://ocw.mit.edu/), whereas in the United Kingdom the Open University has released a significant portion of its content together with authoring tools for collaborative learning environments through the OpenLearn initiative (http://openlearn.open.ac.uk/). An overview of further initiatives can be found in (Yuan, MacNeill, & Kraan, 2008).
3. **Data from related STECs**. A number of existing and past STECs have used data that are similar to those required for QG. In particular, STECs for Question-Answering are relevant for Text-to-Question generation data, and summarization exercises are pertinent specifically to the Target Selection subtask:
    a. DUC/TAC (Document Understanding Conference/Text Analysis Conference): complex questions, summaries as answers,
    b. TREC (Text REtrieval Conference): question taxonomy, target domain, questions in domain,
    c. CLEF QA (Cross Language Evaluation Forum – Question Answering) campaign data: articles from Wikipedia frozen versions, newspapers (for English: LA Times and Glasgow Herald) and questions marked with Question topic, type (factoid/def/list), expected answer type (org/person/...), correct answer, text snippets (Forascu, 2008). For 20091 the JRC-Acquis2, containing European legislative texts, will be also used as source text, and a larger taxonomy of question types is envisaged.
4. **Data from community-driven Question Answering portals**. These will typically consist of a user question with additional detail and multiple answers from various sources. Examples are:
    a. Yahoo!Answer (Marciniak, 2008),
    b. WikiAnswers,
    c. FAQ lists.
5. **Annotation and representation tools**. In addition to static data, relevant resources for QG tasks also include any tools for annotation of text with attributes relevant to the task and reusable generation resources, such as the LinGO/DELPHIN grammars with HPSG reversible syntactic frames (Sag & Flickinger, 2008).

[1] http://nlp.uned.es/clef-qa/
[2] http://langtech.jrc.it/JRC-Acquis.html

6. **QG STEC evaluation data**. It is quite likely that the evaluation of system output in the first years of the QG STEC will involve some manual effort of scoring automatically generated output. When designing the evaluation scheme, it might be worthwhile trying to do this in such a way that the evaluation data are reusable as test data in subsequent years.

3.4. Concluding Remarks

Data resources are a fundamental aspect of any Shared Task Evaluation Campaign (STEC). For Question Generation STECs, this chapter identified existing data resources that are readily available or that could be further processed to help running such QG STECs. The chapter outlined desiderata for data resources with respect to the major phases of the QG process: content selection, question type selection, and question construction. It is feasible and advisable to start an effort to enrich existing resources at least in the near future before QG-targeted data sets could be developed as a first step towards the creation of QG targeted data sets.

CHAPTER 4: METHODS AND METRICS IN EVALUATION OF QUESTION GENERATION

EVALUATION GROUP CHAPTER

RODNEY NIELSEN, BOULDER LANGUAGE TECHNOLOGIES, UNIVERSITY OF COLORADO, BOULDER
KRISTY ELIZABETH BOYER, NORTH CAROLINA STATE UNIVERSITY
MICHAEL HEILMAN, CARNEGIE MELLON UNIVERSITY
CHIN-YEW LIN, MICROSOFT RESEARCH ASIA
JUAN PINO, CARNEGIE MELLON UNIVERSITY
AMANDA STENT, STONY BROOK UNIVERSITY

Abstract

This chapter discusses some of the issues involved in evaluating automated Question Generation (QG) systems. It looks at the pros and cons of various techniques and provides related recommendations, particularly within the context of a multi-year shared task campaign for QG. We discuss how these tasks are affected by the target application, discuss human evaluation techniques, and propose application-independent methods to evaluate system performance automatically.

4.1. INTRODUCTION

The nature of automatic question generation is different depending on the application within which it is embedded. If the purpose is educational assessment, the questions are intended to evaluate the respondent's knowledge, understanding and skills in a subject area. If the intent of the questions is to facilitate learning, such as in a Socratic tutoring environment, then the questions may serve to lead students to an "aha" moment, where they understand a concept that they previously did not. Whereas, if the questions are being generated by an automated help desk system, the goal is for the system to learn what circumstances resulted in the user's request for help and to generate questions that will help ascertain an appropriate solution to the user's problem.

Question generation is ideally defined and evaluated in the context of the application requiring it. For example, in Intelligent Tutoring Systems, given the learner model, learner goals, and a context of prior interactions, the objective is to choose the next topic, question type, and surface form in a way that maximizes some aspect of learning, which should then be evaluated based on learning gains. In contrast, questions in an educational assessment application should be evaluated based on their ability to discriminate between student proficiency levels, whereas questions generated by an automated help desk should be evaluated based on their effectiveness in resolving the customer's problem. The typical types and goals of questions differ based on the application and these differences must be considered when designing a question generation task and appropriate evaluation techniques.

An early goal for the QG research community is to define the nature of a shared task and perhaps its application context (Nielsen R. , 2008; Piwek, Prendinger, Hernault, & Ishuzuka, 2008; Rus, Cai, & Graesser, 2008). One outcome of the Workshop on the Question Generation Shared Task and Evaluation Challenge was that a large majority of the participants were interested in participating in a QG challenge in the context of Intelligent Tutoring Systems (ITS), which may be explained by the large number of workshop participants from the ITS community. This

chapter focuses primarily on the ITS domain. However, there is an attempt to be application independent where possible and to discuss some of the pros and cons of the evaluation strategies relative to other QG contexts. Another outcome of the workshop was general consensus on the three-step QG structure proposed in Nielsen (2008). This chapter follows that structure and discusses evaluation relevant to each of these steps, which are reviewed in the following section.

4.2. QUESTION GENERATION AS A THREE STEP PROCESS

Most applications utilizing question generation can be conceived of as dialogue systems, where the question generated will depend not only on the subject material (e.g., text or knowledge base), but also the context of all previous interactions. Even assessment should ultimately adapt to the student's performance on previous questions. Given a dialogue context, question generation is viewed as a three step process. In the first step, *Content Selection*, the topic from which a question is to be generated is identified. In the second (not necessarily subsequent) step, *Question Type Selection,* a decision is made about the type of question to be asked. In some applications, e.g. Reading Comprehension, the Question Type Selection step may precede the Content Selection step. For instance, a Reading Comprehension system may set the goal of asking a *Why* question and then try to identify causal content to ask about. In the final step, *Question Realization*, the surface form of the question is created based on the prior steps. For the purposes of evaluation, we will consider these to be separate challenge tasks.

Identifying the most appropriate concept from which to construct the *next* question in a dialogue and deciding the question type are possibly the most important goals of question generation (Nielsen R. , 2008; Vanderwende, 2008). While these are very difficult, context sensitive tasks, it is reasonable to identify a priori the set of key concepts from which questions are the most likely to be generated. However, even if the question types are severely constrained, the concepts selected are application dependent, since what is important to one application may not be to another, necessitating distinct tracks for these tasks in the later years of the challenge.

Questions can be generated from a variety of knowledge sources, such as: plain text, linguistically annotated text, structured databases, and knowledge bases with formal semantic or logic representations such as might be output by a natural language understanding system. However, the majority of participants at the workshop preferred that questions be generated from text, possibly with various forms of linguistic annotation. Therefore, this chapter assumes natural language text as the starting point for the Key Concept Identification task. *Starting with the text and the application track, the objective of Key Concept Identification is to output an annotation to identify key spans of text, or snippets, for which questions are likely to be generated*.

The goal of question type selection is to specify a set of characteristics that the question should manifest (Graesser, Rus, & Cai, 2008; Nielsen, Buckingham, Knoll, Marsh, & Palen, 2008). This is a somewhat subjective decision, which is based on a dialogue or pedagogical theory and, ideally, on the dialogue context, user model and goals. However, because considering all of these factors results in a substantial barrier to participate in a STEC, we focus here on the task of determining reasonable question characteristics independent of context. That is, *given a source text and a target concept, the goal of Question Type Selection is to determine the set of reasonable question types*.

The most appropriate type of question does not depend on the text alone, but also on the application-specific context, so we propose the question type be an input to the realization task. Finally, the text itself is a common part of the context across all applications, so it too should be an input. Combined, this leads to the proposal that

the *Question Realization task consists of creating a natural language question of a given type, from specified snippets, given the full text as context*.

4.3. EVALUATION

The discussion begins with some of the general issues that must be considered in designing effective evaluation metrics for QG. We subsequently discuss the evaluation of each of the three proposed shared challenge tasks in the following subsections. These evaluation methods and metrics are largely adopted from Nielsen (2008). For each task, we start by discussing the more conceptual issues and then describe one or more possible specific evaluation strategies, both human and automated.

It is critical to define well-founded evaluation metrics to ensure that research progresses in important directions. A common complaint is that shared tasks often result in all research being tuned to the challenge and to its particular evaluation metric(s), rather than the qualities these metrics were really intended to measure. We attempted to ensure the evaluations described in this section avoid such an undesirable situation and focused on the most important aspects of QG, while not overly constraining research. Additionally, the final two subsections discuss an open track, which moves beyond a shared task and standard evaluation metrics.

4.3.1. EVALUATION DESIDERATA

Human versus Automated Evaluation. Human evaluation is the preferred method for the QG challenge because it should provide more accurate results and facilitate recognizing the nuances of what comprises a good question. Human evaluation sometimes does not imply direct assessment of the question. For instance in ITSs, the ultimate criteria is learning gains. On the other hand, automated evaluation is preferable during system development to facilitate timelier, less expensive, and more consistent assessment of system modifications. Ultimately, a combination of manual and automatic evaluation is likely to provide the most consistent and accurate results, while reducing the costs and time associated with a purely manual approach.

Application-Independent Metrics. The focus of this chapter is on evaluating ITS-related QG, since the workshop discussions suggested this as the first target application. However, the evaluation metrics should, wherever possible, strive for application independence.

End-to-End versus Fine-Grained Evaluation. In assessing the overall value of systems, it is beneficial to have a single metric that can differentiate between levels of performance. Some of the metrics discussed in the Question Realization section should fulfill this purpose. However, for the purposes of assessing why one system outperformed another, it is beneficial to have finer-grained measures. Here we have broken the end-to-end QG process down into the three major tasks described previously and defined individual evaluation metrics for each task. This approach simultaneously facilitates the assessment of subsystem performance on the overall task performance and allows groups to participate and contribute in specific areas without having to build end-to-end QG systems.

We have listed a subset of possible secondary metrics for the final Question Realization task. These metrics should provide additional insight into why one system outperformed another in the overall QG task. We believe it would also be worthwhile to evaluate each subsystem on subcategories of questions (e.g., at minimum based on the major and minor subcategories of question types in the primary taxonomy described in Nielsen et al. (2008) and possibly based on some of the secondary question classifications. This too should provide an effective means of illuminating the contributions of various research groups and their system designs.

Many of the metrics defined in this chapter are averages over performance evaluations of individual questions or equivalent units. Making these lower-level scores available to the research community should facilitate training machine learning algorithms to detect the attributes of high quality questions and to generate such questions. The availability of detailed scoring data will also enable researchers to perform their own analyses of the effects of various system components on the subsets of questions on which their research focuses.

Context Sensitivity and Extrinsic versus Intrinsic Metrics. In the context of an ITS, Content Selection and Question Type Selection should ultimately be performed concurrently and should be based on the specific context of the tutoring dialogue. For example, if human raters evaluate questions independent of the context, we would hypothesize that they would virtually always rate interesting causal consequence questions higher than simple verification questions, despite the fact that in certain contexts, such as for a struggling student, the verification question might be much more appropriate. In fact, for ITS QG, all three subtasks would best be evaluated based on some category of student learning gains, the real goal of an ITS.

However, there are a number of downsides to evaluation metrics based on learning gains in the context of a QG shared task, particularly in the early years of such a task. First and foremost, it would likely limit participation in the challenge. This is due to the fact that (a) participating groups would have to invest deeply in understanding all of the diverse pedagogical factors involved in QG for an ITS, (b) it would require participants to implement all of the associated subsystems regardless of their particular interests, and (c) it would suggest that, for fair comparison, participants should use a common infrastructure, which may not be appropriate for their pedagogical strategy.

Second, such a metric does not facilitate assessing the quality or importance of the main subsystems involved. For comparison, consider the Recognizing Textual Entailment (RTE) challenge. Utilizing a discourse commitment-based strategy, the best performing system (Bensley & Hickl, 2008) outperformed the median accuracy by 15%, while another team purportedly using the same basic strategy only achieved results roughly at the median. Without having common subsystem performance analysis, it is difficult to determine exactly why one system excelled, while the other did not.

Third, a learning gain-based metric would require numerous human subjects to act as students. This would almost certainly extend the timeframe of the evaluation significantly and complicate the administration of the challenge.

Using Common Corpus Annotations versus Providing only the Raw Source Text. From the evaluation perspective, there are pros and cons to providing or enforcing the use of a common annotated input dataset. In many shared tasks, only the raw input text is provided and research groups with very similar approaches sometimes achieve very different results. It is often difficult to tell what resulted in the performance difference, because the discrepancies might have been the result of any subsystem. Requiring the use of a common annotation would ensure that these factors do not cloud the true source of the performance differences. Providing common linguistic and other annotations would also allow research groups, especially students, to avoid the overhead investment in establishing and learning infrastructures and methods that are not central to their research goals.

On the other hand, it is possible that the contribution of a particular research group lies primarily in advances in these ancillary subsystems. The alternative annotation schemes might be central to other aspects of the group's research or philosophy. Hence, requiring the use of a particular annotation scheme could limit participation in the challenge and advances in novel research.

Our recommendation is that most of the common linguistic annotation layers (e.g., sentence segmentation, tokenization, part of speech tagging, constituency and dependency parsing, semantic role labeling, and rhetorical structure annotation) be provided with the input data and that groups be encouraged to use the annotation

provided unless other schemes are central to their research. This should reduce the obfuscating evaluation factors, while allowing groups to spend more time on the aspects of the task they care most about, and still allowing researchers to take novel directions and follow their own philosophies.

There are good reasons for providing both automatically generated linguistic annotations and gold-standard human annotations. In order to assess the state of the art, systems must use annotations output by automated systems, which are relatively easy to supply. However, it is also reasonable to assess the impact of a given approach assuming it had nearly perfect input, since an approach might provide no value or even have a negative impact when utilizing noisy inputs, whereas the same system could impart significant benefits if its input data were of high quality. Therefore, to the extent feasible from a time and cost perspective, we propose providing gold-standard human annotated text in addition to automated annotations, which are relatively easy to supply for large quantities of source text or other input.

Other Obfuscating Factors. In defining the QG shared tasks it is also important to avoid factors that might result in an inability to differentiate between systems or to determine the reason for the systems' performance differences. For example, allowing some groups to incorporate spoken dialogue while others use strictly text would affect the perception of system output, even if the underlying QG systems were equivalent. However, such systems should still be allowed to submit results under an open track that is not included in the main evaluation.

Additionally, it is important that the test data cover a broad enough spectrum of input cases and be large enough to ensure that differences can be detected in the participating systems. Furthermore, the evaluation results should include measures of statistical significance to elucidate the likelihood that performance differences between two systems are greater than what would be expected by chance.

4.3.2. Content Selection

Content Selection and Question Type Selection should be based on the specific context of the dialogue, and for an ITS, should consider the learner model and goals. This is an ambitious task and will likely not be considered until later years of the challenge. In the early years, it may be better to perform and evaluate the tasks separately and independent of context. In this case, Key Concept Identification can be framed as identifying the set of concepts for which questions are the most likely to be generated at some point in a dialogue. These key concepts are assumed to be typical of what might be found in a good summary, which could encourage Automatic Summarization researchers to participate in this subtask.

Conceptually what one would like for Key Concept Identification is an evaluation metric that scores systems that find all of the key concepts (high Recall) and only the key concepts (high Precision), penalizing systems that over-generate or classify less important concepts as being key. This suggests the use of the F-measure (van Rijsbergen, 1979). However, the standard F-measure would assume that all concepts could be classified in a binary manner, as key or not key, whereas concept importance actually varies along a scale. This would imply the use of a ranking metric or correlation with a scalar importance judgment. However, having human assessors rank or even rate the importance of a large number of the concepts in a text would be a subjective and time-intensive task. Therefore, we propose a metric, detailed below, that is similar in spirit to the modified F-measure developed to evaluate question answering described by Lin and Demner-Fushman (2005). In addition, other metrics, such as a variant of average precision (Dagan, Glickman, & Magnini, 2006), commonly used in Information Retrieval could be utilized to assess a system's confidence estimates.

The evaluation methods and metrics described here assume that for a given application track (e.g., ITS), at least two experts in that area annotate a set of test documents, tagging the spans of text (*snippets*) representing key

concepts, eventually implied concepts, from which the experts feel questions should be generated (these spans need not be contiguous). The experts also label the snippets as vital or optional depending on the perceived significance of the concepts. *Vital* snippets represent those concepts that must be identified by a high quality QG system, and *optional* snippets represent concepts that it is reasonable for the QG system to identify. Alternatively, the annotators could extract and rewrite the key concepts more succinctly. A third expert would then adjudicate these snippets.[3] A system's job is to find all of the vital concepts in the text, while not tagging any false positives (concepts that are neither vital nor optional). It should be noted that a simple identification of key snippets in the source text biases towards literal questions as opposed to deeper inferential questions. Annotating protocols should be designed such that inferential questions could be evaluated.

Since human assessors would provide the most accurate assessment of whether a gold-standard concept is identified by a system, they should be utilized in the organized challenge. This would also involve determining the reliability of human annotation, by computing inter-annotator agreement between human annotators. Human judgments could either be binary decisions, (i.e., identified versus not identified), or could be ratings of the extent to which a concept was covered by the system. The binary decision is more practical from a time and cost perspective, while the scalar judgment might allow more sensitive measurements of QG system performance.

Given an annotated test set, the Key Concept Identification task can also be evaluated using a fully automatic method that is completely independent of the application for which the QG is being performed. This F-measure-based evaluation weights each vital concept equally, independent of its length, and bases recall on the coverage of vital snippets and precision on the extent to which a system tagged snippet is covered by a single human annotated snippet, vital or optional.

It is important to articulate a formal definition of recall scores, precision scores, and F-measures. These are provided below. The scores are defined with respect to *facets,* which are any fine-grained component of the semantics of an utterance or text; however other underlying units of meaning could be used. Facets could be extracted from the relations in a syntactic dependency parse (Nielsen R. , Ward, Martin, & Palmer, 2008).

Let *k* be the number of vital snippets, *m* be the total number of annotated snippets across all human annotators, *n* be the total number of system-tagged snippets, $\boldsymbol{V}_i$, $\boldsymbol{A}_i$, and $\boldsymbol{S}_i$ be the set of semantic facets in the i^{th} vital, human-annotated (vital and optional), and system-tagged snippets respectively, and $|\boldsymbol{X}_i|$ be the number of semantic facets in the specified set Calculate the *Instance Recall* (*IR*) for each vital snippet and *Instance Precision* (*IP*) for each system-tagged snippet as:

$$IR_i = \max_{j=1..n} \left|\mathbf{V}_i \cap \mathbf{S}_j\right| \Big/ \left|\mathbf{V}_i\right|$$

$$IP_j = \max_{i=1..m} \left|\mathbf{S}_j \cap \mathbf{A}_i\right| \Big/ \left|\mathbf{S}_j\right|$$

Let the overall recall and precision equal the average instance recall and precision and calculate the F-measure as usual, where β assigns a relative importance to precision and recall:

$$R = 1 \Big/ k \sum_{i=1}^{k} IR_i$$

[3] For efficiency, much of this adjudication might be skipped by, for example, considering all spans that were included within at least two annotators' snippets to be vital, if one marked it as such, with most other spans optional, and by only adjudicating spans with particular word overlaps.

$$P = 1/n \sum_{j=1}^{n} IP_j \quad F_\beta = \left(1+\beta^2\right) P \cdot R \Big/ \left(\beta^2 P + R\right).$$

The procedure described allows different snippet alignments when calculating *IR* versus *IP* and in some cases, multiple alignments for a single snippet. It could be revised to find the single alignment that maximizes the overall F-measure, but this is probably not worth the effort, as it would probably only have a significant effect on the metric for extreme cases. It is worth considering these extreme cases in the next paragraph as baseline measures of performance.

A system that tags the full text as a single snippet would achieve perfect recall, $R = 1.0$, (it will have tagged all of the vital concept snippets) and its precision would be $P = \max_{i=1..m} |\boldsymbol{A}_i|/|\boldsymbol{T}|$, where $\boldsymbol{T}$ is the set of all semantic facets in the source text. If less than 10% of the document's semantic facets were tagged by the annotators and the largest snippet included less than 1/3 of the total tagged content words, then $P < 0.1/3$ and the balanced F-measure, $F_1 < 0.065$. If annotator snippets all corresponded to sentences, an approach that tagged each sentence separately would also achieve perfect recall, $R = 1.0$, (again, it will have tagged all of the vital concept snippets) and $P = m/n$. If annotators tagged less than 10% of the document sentences, then $P < 0.1$ and $F_1 < 0.182$. An approach that tags only the single most probable sentence, assuming that sentence is a vital snippet, would achieve perfect precision, $P = 1.0$, (everything the system tagged would also be identified in the gold standard) and $R = 1/k$. If there were $k \geq 10$ vital snippets in the gold standard, then $R \leq 0.1$ and $F_1 \leq 0.182$. As is desired, these F-measures are all sufficiently low that one would hope that essentially all systems would easily exceed them.

We further propose that *IR* and *IP* be calculated based on whether the facets in a gold-standard concept are entailed by the corresponding system-tagged snippet, or based on the probability of this entailment. Nielsen, Ward and Martin (2008) describe a facet-based entailment system that outperforms the simpler lexical overlap based equivalence assessment and should provide more accurate evaluation results. This hypothesis will be tested in the first QG challenge by comparing each to human judgments.

Key Concept Identification is similar to Automatic Summarization (AS) in that both seek to identify critical information in the source text. The goal of AS is to provide a fluent summary of the key noteworthy concepts in the text, which systems often generate by extracting sentences verbatim from the original document. As described above, the goal of Key Concept Identification is to identify key question-worthy concepts by selecting spans of the original text. However, adopting AS evaluation methods, such as ROUGE (Lin C. Y., 2004), results in a few shortcomings when applied to Key Concept Identification. First, AS compares the full summaries rather than the individual key concepts. This would effectively lead to weighting longer question concepts more heavily, rather than treating all questions as equally important. Second, and more importantly, ROUGE has no means of differentiating between vital concepts and concepts that it is reasonable to consider question-worthy, but that are not critically important.

PREG is a conceptual model that predicts the questions that will be asked by students (Otero & Graesser, 2001). The model is evaluated based on signal detection theory, using a metric called d' (Green & Swets, 1966). This metric, like the F-measure, combines two lower-level measures. The first measure is the same recall value that is part of the F-measure, but is called the *hit rate* in signal detection theory and is called the *sufficiency score* by Otero and Graesser. The second value, rather than measuring the precision of identifying question-worthy topics, as in the F-measure, instead measures the *false alarm rate* from signal detection theory. The false alarm rate is the fraction of all negative examples that the system labels as positive examples, (i.e., given all of the questions that could be generated, but that are not worthy of generation, it measures the fraction of these questions that the model erroneously predicted would or should be asked). In the context of the QG task, the false alarm rate would

be a strange measure, since theoretically, for any text, there are an infinite number of possible frivolous questions that could be asked, effectively resulting in a constant false alarm rate of 0.0. Precision, on the other hand, is clearly interpretable (what is the system's accuracy when it labels a concept as being key), and is easily quantifiable (the number of key concepts labeled by the system is easily counted, as is the number which were system errors).

Rather than using semantic facets, the evaluation method defined here could just as easily be adapted to use surface text content words, word stems, n-grams, or combinations of n-grams, varying *n* as in the BLEU score (Papineni, Ward, Roukos, & Zhu, 2002). However, the hypothesis is that semantic facets abstract away from the syntax and surface form of the text and focus more on the semantics involved in the underlying concepts and would, therefore, provide a higher correlation with human judgments of concept coverage. This hypothesis will be tested in the first QG challenge. It should be noted that utilizing semantic facets in the evaluation method does not imply that QG systems must use the representation internally; it need only be implemented in the evaluation tools, which can be made publicly available. For detail regarding the automatic generation of the facet representation, see (Nielsen R., 2008).

4.3.3. Question Type Selection

The goal of question type selection is to specify a set of characteristics that the question should manifest (c.f., Graesser et al., 2008; Nielsen et al., 2008b). This is a somewhat subjective decision, which is based on a dialogue or pedagogical theory and, optimally, on the dialogue context, user model and goals. However, since considering all of these factors results in a substantial barrier to challenge entrance, we focus here on the task of simply determining question characteristics independent of context. That is, given just a source text (or knowledge base) and a target concept (output from the previous subtask), the goal is to determine the *set* of reasonable question types.

There are numerous characteristics that can be specified for a question. Here, we primarily focus on selecting the question's basic type (e.g., Concept Completion, Procedural, Causal Consequence, Justification). However, many of the issues in this section apply equally well to a number of the other question characteristics. For this subtask, it is unclear how much value would be added by human evaluation; therefore, we focus strictly on presenting one possible automated evaluation method.

Given the source text and target concept, annotators would enumerate the types of questions that would be appropriate to generate for a given target concept. These lists would then be adjudicated, and the gold-standard question types would be labeled as being vital for a system to identify or optional. System output would then be evaluated based on an average F-measure over all target concepts. The F-measure for a given target concept is calculated in the usual manner. Its recall is computed as the fraction of vital question types included in the system's output for that target concept and the precision is the fraction of all of the question types selected by the system that were identified as either vital or optional in the gold-standard annotation. Again, the final evaluation metric would then be the average over all of the target concept F-measures.

Essentially, no change is required for this evaluation when dialogue context is considered and question type selection is performed concurrently with target concept selection. However, in this case it is likely that the lists of vital and optional question types identified by the annotators would be constrained to a smaller set by the context.

4.3.4. Question Realization

The output of the Question Realization task is the final surface form of the question. Longer-term considerations might include gestures, facial expressions, intonation, etc., (c.f., (Piwek, Prendinger, Hernault, & Ishuzuka, 2008),

for a related discussion), but, if mandatory, these would likely severely restrict participation in the challenge, and thus are not explicitly considered in this section. Inputs to the Question Realization task include the target concept and question type described in the subtasks above, as well as one or more source texts and or knowledge bases. Longer-term considerations might include a user model, goals, and dialogue context, among other things.

To be able to fairly compare Question Realization subcomponents and to facilitate participation by researchers who are interested only in the realization subtask (e.g., some NLG groups), the target concept and question type inputs should be gold-standard data. This is the assumption throughout this section. However, to evaluate end-to-end system performance, the target concept and question type output by the same team in previous subtasks would be used. Even in this case, many of the evaluation metrics described in this section would still apply.

Optimally, the Question Realization task would be evaluated differently depending on the application. Questions for educational assessment might be evaluated according to their discriminating power (Lin and Miller, 2005), tutoring questions for their effect on learning gains, automated help desk questions for their efficacy in resolving a customer's problem, etc. This would effectively ensure that research focuses on questions and aspects of QG that are important to the application (see (Vanderwende, 2008) for a discussion of question importance).

In the educational assessment track, where most prior work has taken place, one possibility would be to have a two-part human evaluation. First, judges filter out questions that do not match the specified type or target content. Then, the remaining questions are distributed across tests, and the final evaluation is based on the average discriminating power of the questions, assigning questions filtered out due to type or content errors the worst power possible, -1.0.

It is probably impractical to evaluate ITS QG optimally, within the framework of a short challenge, since one really cares about longer-term learning gains, preparation for future learning (Bransford & Schwartz, 1999), accelerated future learning (Chi & VanLehn, 2007) or a similar effect, none of which are likely to be achieved in a single or a few tutoring sessions. On the other hand, it might be reasonable to run the challenge over an extended period of time or simply to evaluate based on short-term learning. An evaluation of short-term learning might be conducted over the Internet, for example, via a system like Amazon's Mechanical Turk , which allows anyone with Internet access to sign up to perform human intelligence tasks such as annotating text or answering questions.

Another option to consider is evaluating the systems participating in the challenge based on average ratings (or rankings) from appropriate experts. This could be done independent of context in the early years of the challenge and could be enhanced to include a consideration of the context in the later years. However, one difficulty with such an evaluation is that two experts who subscribe to different pedagogical theories might provide radically different ratings for the questions that systems generate. A possible solution is to require that all systems be based on a common pedagogical theory, but this restriction would negatively curb research in a way similar to utilizing an inappropriate evaluation metric (Vanderwende, 2008).

All of the above are effectively extrinsic evaluation metrics, focusing on the target application. There are also numerous intrinsic metrics that might be of more interest to particular research communities, (e.g., adequacy, grammaticality, fluency, succinctness, clarity, appropriate use of anaphora, interestingness, and difficulty). Examining the correlation between these metrics and the extrinsic metrics associated with an application's ultimate goal might help elucidate the important factors in system performance along with the differences in and contributions of participating systems. While some of the factors listed above might be part of the question type specification (e.g., difficulty), their impact on the extrinsic metrics might still be illuminating.

Automated system evaluation faces many of the same challenges as human evaluation. Nevertheless, if the question type and source text snippet are provided as input, then the questions generated are likely to look similar frequently, regardless of the application, particularly in the early years of the shared task. Therefore, we propose an automatic evaluation technique that compares the system-constructed question to one or more gold-standard questions written by application experts. This form of evaluation, which is consistent with the proposal of Rus, Cai and Graesser (2007) and common in other areas such as Machine Translation (MT) and Automatic Summarization (AS), generally involves comparing overlap in n-grams.

(Soricut & Brill, 2004) present a unified framework for automatic evaluation using n-gram co-occurrence statistics, which in part relates evaluation factors (faithfulness, compactness, precision and recall) to the size of n-grams. MT typically utilizes up to 4-grams to ensure fluency; whereas, AS, which often comprises selecting already syntactically sound key sentences is frequently evaluated strictly by unigrams, since fluency is essentially guaranteed. It remains an open question what would be the appropriate size of context to evaluation Question Generation using n-grams.

Additionally, we propose to assess an evaluation method that utilizes the facet-based representation discussed above in the evaluation of Key Concept Identification. This representation effectively factors out much valid syntactic alternation and focuses near the bigram level. We propose the use of this representation and the corresponding entailment system (Nielsen R. , Ward, Martin, & Palmer, 2008) to automatically evaluate the extent to which a system question is a paraphrase of a gold-standard question. Specifically, we propose to use an average F-measure over questions, where a constructed question's F-measure is based on the most similar expert question. The question's recall would be calculated as the fraction of facets in the expert question that are entailed by the system question and, inversely, its precision would be the fraction of facets in the system question that are entailed by the expert question. As above, it is also worth testing the effect of using the probability of entailment, rather than a strict binary entailment decision.

Precedents for this facet-based evaluation include: Owczarzak, Genabith and Way (2007), which found that a dependency-based metric correlates higher with human judgment on fluency than n-gram metrics; Turian, Shen and Melamed (2003), which found that an F-measure can outperform current precision-focused metrics in similar evaluations; Perez and Alfonseca (2005), which indicated that MT n-gram-based metrics fall far short in recognizing textual entailment; and Lin and Demner-Fushman (2005), which found that macro-averaging over answers is more appropriate than micro-averaging over answer nuggets in Question Answering evaluation.

Within the ITS or educational assessment settings, the automated evaluation metric, whether based on facets or n-grams or other units of text or meaning, must also appropriately penalize questions that include facets of the reference answer (or targeted content/key concepts) that are not included in the expert question. Otherwise, questions that give away the answer or that simply repeat the source text might result in an undesirably high score. A downside to automatic evaluation is that it will inappropriately penalize the best problem solving questions, most of which are unique, having extremely little overlap on the surface. However, this can be addressed in the future, when such question generation becomes more feasible.

A possible compromise between human and automatic evaluation during system development, external to the QG challenge, is to have the gold-standard or contrast questions evaluated by humans and then weight automatic metrics by the quality of the contrast question involved in entailing the system question. These contrast questions could be gold-standard questions generated by experts, they could be questions generated by novices, or they could be generated by QG systems. In fact, for use during system development and tuning, the evaluated contrast

questions could be a natural byproduct of previous human evaluations of questions generated during the workshop challenge.

4.3.5. Open QG Track

Forcing all research groups interested in contributing to the QG workshop to perform the shared task and be evaluated according to the metrics described in this chapter could result in groups not pursuing novel related research and advancing the broader state of the art in QG. To alleviate this problem, we propose that the challenge include an open track, where researchers can contribute according to their own goals and priorities. These research groups can submit papers for peer review under the open track, utilizing evaluation metrics appropriate to their particular QG task or application area.

4.3.6. Evaluation Track, Tools, and Participant Results

We recognize that the metrics described here are not the best possible long-term means of evaluating QG systems. A common problem seen in other areas such as Machine Translation is that an evaluation metric that is adopted early might remain in use indefinitely even when other metrics show improvements. One of the key reasons for this is that, when writing a paper, researchers must compare their results to prior work and the only available evaluation results for that prior work are based on the earlier evaluation metrics. Another reason is that there is generally not a clear established means of vetting new evaluation methods.

We recommend that these issues be addressed in several ways. First, we propose establishing an evaluation track in later years of the QG challenge and developing a formal process for vetting submitted evaluation tools and metrics. Second, we propose requiring that any users of the challenge datasets must submit their system output to a central archive for the purposes of direct comparison in the future. Similarly, we propose that participants in the evaluation track who utilize novel evaluation tools must provide these tools for distribution under an open source agreement. This would allow future researchers to evaluate prior work utilizing the latest accepted evaluation metrics and then compare those results to their own system's performance using the new metrics.

An additional benefit of incorporating an evaluation track in future challenges is that it would encourage the participation of other research groups. For example, given a number of the metrics described above, it is likely that members of the Recognizing Textual Entailment research community would contribute to this track.

4.4. Conclusion

In short, we believe it is feasible to implement effective evaluation metrics for shared tasks in a Question Generation challenge. Furthermore, we believe metrics and strategies can be designed in a way that encourages broad participation and novel research. The evaluation methods and metrics described in this chapter should fulfill these and other goals of a QG challenge.

REFERENCES

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94* , 416-427.

Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D. L. Forrest-Presley, G. E. Mackinnon, & T. G. Waller (Eds), *Metacognition, cognition and human performance* (pp. 155-205). New York: Academic Press.

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15* , 95-115.

Beck, I., McKeown, M., Hamilton, R., & Kucan, L. (1997). *Questioning the Author: An approach for enhancing student engagement with text.* Delaware: International Reading Association.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine, & P. C. Kyllonen (Eds), *Generating items from cognitive tests: Theory and practice* (pp. 199- 217). Mahwah, NJ: Lawrence Erlbaum.

Bensley, G., & Hickl, A. (2008). Application of lccs groundhog system for rte-4. *Proceedings of the Text Analysis Conference (TAC). National Institute of Standards and Technology (NIST).*

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain.* New York: McKay.

Bransford, J. D., & Schwartz, D. L. (1999). *Rethinking transfer: a simple proposal with multiple implications.* Washington, D.C.: American Educational Research Association.

Burstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis, & J. Burstein (Eds), *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25* , 471-533.

Chi, M., & VanLehn, K. (2007). Accelerated future learning via explicit instruction of a problem solving strategy. *Artificial Intelligence in Education.* Amsterdam, Netherlands: IOS Press.

Chinn, C., & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63* , 1-49.

Ciardiello, A. V. (1998). Did you ask a good question today? Alternative cognitive and metacognitive strategies. *Journal of Adolescent & Adult Literacy, 42* , 210-219.

Collins, A. (1988). Different goals of inquiry teaching. *Questioning Exchange, 2* , 39-45.

Corbett, A., & Mostow, J. (2008). Automating comprehension questions: Lessons from a reading tutor. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

References

Craig, S. D., Gholson, B., Ventura, M., & Graesser, A. C. (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11* , 242-253.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). Deep-level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. *Cognition and Instruction, 24(4)* , 563-589.

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, & F. d'Alché-Buc, *Machine Learning Challenges. Lecture Notes in Computer Science , Vol. 3944* (pp. 177-190). Springer.

Deane, P., & Sheehan, K. (2003). *Automatic item generation via frame semantics.* Retrieved from Education Testing Service: http://www.ets.org/research/dload/ncme03-deane.pdf

Dillon, J. (1990). *The practice of questioning.* New York: Routledge.

DiPaolo, R. E., Graesser, A. C., White, H. A., & Hacker, D. J. (2004). Hints in human and computer tutoring. In M. Rabinowitz (Ed.), *The design of instruction and evaluation* (pp. 155–182). Mahwah, NJ: Erlbaum.

Edelson, D. C., Gordin, D. N., & Pea, R. D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *The Journal of the Learning Sciences, 8* , 391-450.

Festinger, L. (1957). *A theory of cognitive dissonance.* Evanston, IL: Row, Peterson.

Flammer, A. (1981). Towards a theory of question asking. *Psychological Research, 43* , 407-420.

Forascu, C. (2008). Generating Questions in the CLEF taxonomy. *Workshop on Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Gates, D. (2008). Generating Look-Back Strategy Questions from Expository Texts. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Good, T., Slavings, R., Harel, K., & Emerson, M. (1987). Students' passivity: A study of question asking in K-12 classrooms. *Sociology of Education, 60* , 181-199.

Graesser, A. C., & McMahen, C. L. (1993). Anomalous information triggers questions when adults solve problems and comprehend stories. *Journal of Educational Psychology, 85* , 136-151.

Graesser, A. C., & Olde, B. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology, 95* , 524-536.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31* , 104-137.

Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In G. V. Nakamura, R. M. Taraban, & D. Medin (Eds.), *The psychology of learning and motivation: Vol. 29. Categorization by humans and machines* (pp. 411-436). Orlando, FL: Academic Press.

Graesser, A. C., Lu, S., Jackson, G., Mitchell, H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36* , 180-193.

References

Graesser, A. C., Lu, S., Olde, B. A., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory and Cognition, 33* , 1235-1247.

Graesser, A. C., McNamara, D., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist, 40* , 225-234.

Graesser, A. C., Ozuru, Y., & Sullins, J. (2009). What is a good question. In M. McKeown (Eds), *Festscrift for Isabel Beck.* Mahwah, NJ: Erlbaum.

Graesser, A. C., Person, N., & Huber, J. (1992). Mechanisms that generate questions. In T. Lauer, E. Peacock, & A. C. Graesser (Eds), *Questions and information systems.* Hillsdale, NJ: Erlbaum.

Graesser, A. C., Rus, V., & Cai, Z. (2008). Question classification schemes. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Guthrie, J. T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly, 23* , 178-199.

Hacker, D. J., Dunlosky, J., & Graesser (Eds), A. C. (1998). *Metacognition in educational theory and practice.* Mahwah, NJ: Erlbaum.

Harabagiu, S. M., Maiorano, S. J., & Pasca, M. A. (2002). Open-domain question answering techniques. *Natural Language Engineering, 1* , 1-38.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30* , 141-158.

Ignatova, K., Bernhard, D., & Gurevych, I. (2008). Generating High Quality Questions from Low Quality Questions. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Johnson, W. L., Rickel, J., & Lester, J. (2000). Animated Pedagogical Agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 11* , 47-78.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Prentice-Hall.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31* , 338-368.

Landauer, T., McNamara, D., Dennis, S., & Kintsch (Eds), W. (2007). *Handbook on Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37* , 389-405.

Lehnert, W. G. (1978). *The Process of Question Answering: a computer simulation of cognition.* Lawrence Erlbaum Associates.

References

Lin, C. Y. (2008). Automatic question generation from queries. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Lin, C. Y. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. *Workshop on Text Summarization Branches Out.*

Lin, J., & Demner-Fushman, D. (2005). Automatically Evaluating Answers to Definition Questions. *In Proc. HLT/EMNLP.*

Lin, R. L., & Miller, M. D. (2005). *Measurement and Assessment in Teaching.* Prentice Hall.

Marciniak, T. (2008). Language generation in the context of Yahoo! answers. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

McNamara, D., Levinstein, I., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers, 36* , 222-233.

Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88* , 314-332.

Nielsen, R. D., Buckingham, J., Knoll, G., Marsh, B., & Palen, L. (2008). A taxonomy of questions for question generation. *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge.* Arlington, Va.

Nielsen, R. D., Ward, W., Martin, J., & Palmer, M. (2008). Annotating students' understanding of science concepts. *Proceedings of the Language Resources and Evaluation Conference.*

Nielsen, R. (2008). Question Generation: Proposed challenge tasks and their evaluation. *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Nielsen, R., Ward, W., Martin, J., & Palmer, M. (2008). Automatic Generation of Fine-Grained Representations of Learner Response Semantics. *In Proc. ITS.*

Otero, J. (in press). Question Generation and Anomaly Detection in Texts. In D. Hacker, J. Dunlosky, & A. Graesser (Eds), *Handbook of Metacognition in Education.* Routledge.

Otero, J., & Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction, 19* , 143-175.

Otero, J., Ishiwa, K., & Vicente, S. (2008). Readers' questioning: Some hints for automated question generation. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Owczarzak, K., Van Genabith, J., & Way, A. (2007). Dependency-Based Automatic Evaluation for Machine Translation. *In Proc. NAACL/HLT Workshop on Syntax and Structure in Statistical Translation.*

Palinscar, A. S., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1* , 117-175.

Papineni, K., Ward, T., Roukos, S., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Philadelphia: Association for Computational Linguistics.

Perez, D., & Alfonseca, E. (2005). Application of the Bleu algorithm for recognising textual entailments. *PASCAL WS Recognizing Textual Entailment.*

Piwek, P., Prendinger, H., Hernault, H., & Ishuzuka, M. (2008). Generating questions: An inclusive characterization and a dialogue-based application. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Prasad, R., & Joshi, A. (2008). A discourse-based approach to generating why-questions from text. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., et al. (2008). The Penn Discourse TreeBank 2.0. *In Proceedings of LREC.*

Pressley, M., & Forrest-Pressley, D. (1985). Questions and children's cognitive processing. In A. C. Graesser, & J. B. Black (Eds), *The psychology of questions* (pp. 277-296). Hillsdale, NJ: Erlbaum.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, televisions, and new media like real people and places.* Cambridge, U.K.: Cambridge University Press.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research, 66* , 181-221.

Rouet, J. (2006). *The skills of document use: From text comprehension to web-based learning.* Mahwah, NJ: Erlbaum.

Rus, V., Cai, Z., & Graesser, A. C. (2007). Evaluation in Natural Language Generation: The Question Generation Task. *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation.*

Rus, V., Cai, Z., & Graesser, A. C. (2008). Question generation: A multiyear evaluation campaign. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Sag, I. A., & Flickinger, D. (2008). Generating questions with deep reversible grammars. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Scardamalia, M., & Bereiter, C. (1992). Text-based and knowldege-based questioning by children. *Cognition and Instruction, 9* , 177-199.

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively.* Hillsdale, NJ: Erlbaum.

Schank, R. (1999). *Dynamic memory revisited.* New York: Cambridge University Press.

Smith, N., Heilman, M., & Hwa, R. (2008). Question Generation as a Competitive Undergraduate Course Project. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

Soricut, R., & Brill, E. (2004). A unified framework for automatic evaluation using n-gram cooccurence statistics. *Proceedings of ACL-2004: 42nd Annual meeting of the Association for Computational Linguistics.*

Turian, J., Shen, L., & Melamed, I. (2003). Evaluation of Machine Translation and its Evaluation. *In Proc. of the MT Summit IX.*

Van der Meij, H. (1987). Assumptions if information-seeking questions. *Questioning Exchange, 1* , 111-118.

Van der Meij, H. (1994). Student questioning: A componential analysis. *Learning and Individual Differences, 6* , 137-161.

van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition.* Butterworths.

Vanderwende, L. (2007). Answering and Questioning for Machine Reading. *In Proc of the 2007 AAAI Spring Symposium on Machine Reading.*

Vanderwende, L. (2008). The importance of being important. *Workshop on the Question Generation Shared Task and Evaluation Challenge.* NSF, Arlington, VA.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31* , 3-62.

Verberne, S., Boves, L., Coppen, P. A., & Oostdijk, N. (2007). Discourse-based answering of why-questions. Traitement Automatique des Langues. *Special Issue on Computational Approaches to Document and Discourse, 47(2)* , 21–41.

Voorhees, E. (2001). The TREC Question Answering Track. *Natural Language Engineering, 7* , 361-378.

Wisher, R. A., & Graesser, A. C. (2007). Question asking in advanced distributed learning environments. In S. M. Fiore, & E. Salas (Eds), *Toward a science of distributed learning and training* (pp. 209-234). Washington, D.C.: American Psychological Association.

Yuan, L., MacNeill, S., & Kraan, W. (2008). *Open Educational Resources – Opportunities and Challenges for Higher Education. Technical report.* Retrieved December 12, 2008, from JISC CETIS: http://wiki.cetis.ac.uk/images/0/0b/OER_Briefing_Paper.pdf

ISBN: 978-0-615-27428-7