Question Answering and Generation

Arthur C. Graesser[1], Vasile Rus[2], Zhiqiang Cai[1], and Xiangen Hu[1]
[1] *Department of Psychology*
[2] *Department of Computer Science*

Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
USA

ABSTRACT

Automated Question Answering and Asking are two active areas of Natural Language Processing with the former dominating the past decade and the latter most likely to dominate the next one. Due to the vast amounts of information available electronically in the Internet-era, automated Question Answering is needed to fulfill information needs in an efficient and effective manner. Automated Question Answering is the task of providing answers automatically to questions asked in natural language. Typically, the answers are retrieved from large collections of documents. While answering any question is difficult, successful automated solutions to answer some type of questions, so called factoid questions, have been developed recently culminating with the just announced Watson Question Answering system developed by I.B.M. to compete in Jeopardy-like games. The flip process, automated Question Asking or Generation is about generating questions from some form of input such as text, meaning representation, or database. Question Asking/Generation is an important component in learning technologies such as tutoring systems. Advances in Question Asking/Generation will revolutionize learning and dialogue systems. This chapter presents an overview of recent developments in Question Answering and Generation starting with a description of the question landscape.

INTRODUCTION

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 answering and generation facilities. As a result, we believe that the Google generation is destined to have a much more inquisitive mind than the generations who relied on passive reading and libraries. The new technologies will radically transform how we think and behave.

BACKGROUND

Automatic Question Answering is the task of providing meaningful answers to questions in natural language. There are two features that makes automatic Question Answering attractive: (1) it keeps the user-system interface natural as users can ask questions the way they ask other humans therefore eliminating the need to train users on specific query languages and (2) it can provide effective access for everyone to the huge online repository of knowledge which is the Internet. Advanced online Question

Answering services can provide effective access to information to everyone, computer-savvy or not, as interface barriers are eliminated.

While early explorations of automated Question Answering have been attempted since the beginning of the computing era, the advent of the Internet in the 1990s greatly stimulated research on Question Answering in order to provide effective access to information to all users. In particular, research during the last decade has focused on building Question Answering technologies that can successfully answer one type of questions, factoid questions which have well-defined answers. This chapter emphasizes these recent developments on factoid Question Answering.

The reverse process of Question Generation (QG) or asking is a fundamental human capacity that is present in childhood as a primary form of learning, curiosity, and discovery. Students in K12, college, and adult populations are known to improve their learning after they learn how to acquire improved skills of QG. QG is an essential component of learning environments, help systems, information seeking systems, and a myriad of other applications. Mechanisms of QG have been less explored in the Computational Linguistics and Text Retrieval community, the two communities that led the recent efforts on Question Answering processes. We know that language generation is a more difficult task to take on, as all natural language generation tasks are, but do not believe that the inherent difficulty should prevent the exploration of automated QG. Recent efforts by Rus and colleagues (2007, 2009a, 2009b) led to the creation for the first time of a coherent and strong QG research community, which has grand research plans for the next decade.

QUESTION QUALITY, COMPLEXITY, AND TAXONOMIES

An important initial step in a Question Answering or Generation project 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 present in this section is merely the tip of the iceberg.

Question taxonomies have been proposed by researchers who have developed models of Question Answering and Generation in the fields of artificial intelligence, computational linguistics (Voorhees, 2001), discourse processing, education and a number of other fields in the cognitive sciences (for a review, see Graesser, Ozuru, Y., & Sullins, 2009).

Sincere-information seeking (SIS) versus other types of questions. Questions are not always generated by a person's knowledge deficits and cognitive disequilibrium, which occurs when there are obstacles to goals, contradictions, impasses during problem solving, anomalous information, and uncertainty. 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 (1977) 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 (Graesser, Ozuru, Y., & Sullins, 2009).

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?

Given the diversity of questions that can be asked, building systems that can automatically answer them can be challenging. We present next an overview of automated approaches to the task of Question Answering.

QUESTION ANSWERING

Building computer systems that could answer natural language questions has been puzzling the research community since the early days of computing. Early automated Question Answering systems have focused on providing quick and natural access to expert knowledge stored in some computational form (formal knowledge bases, structured databases, etc.). A good example is Woods and colleagues' LUNAR system (Woods, Kaplan, & Webber, 1972). LUNAR was created to answer questions about the Apollo 11 moon rocks for the NASA Manned Spacecraft Center. In particular, the goal was to built a system "sufficiently natural and complete" that the wording of a question would require negligible effort for the user. An example of questions that the LUNAR system was supposed to answer is "What is the average concentration of aluminum in high-alkali rocks?"

The PLANES (Waltz, 1978) Question Answering system was intended to offer a natural language interface to access information in a structured database (aircraft maintenance domain). The basic idea was to allow a non-programmer to retrieve information from a structured database with minimal prior training. One interesting feature of the system was educating the user in formulating questions that the system could understand.

As we already learned, answering any type of questions is very challenging as it requires knowledge about the world, user task, inference capabilities, user modeling, linguistics knowledge, and knowledge about the pragmatics of discourse and dialogue. To make the task more palatable given the current technological advancements, the research community decided in late 1990s to focus on one type of questions, i.e. factoid questions whose answers are relatively short and well-defined. The National Institute for Standards and Technology (NIST) initiated in 1999 the Question Answering challenge, the first large-scale evaluation of domain-independent Question Answering systems (Voorhees & Tice, 2000). The Question Answering challenge was one of the many challenges NIST proposed as part of its Text REtrieval Conference (TREC; trec.nist.gov), which promotes advances in text retrieval technologies.

Probably the most important factor that prompted the NIST's Question Answering challenge was the advent of the Internet, in particular the World Wide Web, in the 1990s which led to an explosion of information available online, mainly textual information (as opposed to formal knowledge or structured databases). As a first response to the information retrieval challenge posed by the Internet, commercial services in the form of search engines were created to provide information search functions to users looking for information on the web. Nevertheless, traditional search engines of that era were classical information retrieval systems lacking two important features that were needed by the community of

Internet users, most of them mundane users. The two missing features were: (1) the user queries were not natural language questions but rather sets of keywords and (2) the output of these systems was a ranked list of documents (web pages); users had to look up the answer in the list of documents in order to fulfill his particular information need, a time-consuming task. NIST's Question Answering challenge fostered the development of large-scale Question Answering systems that could retrieve answers to users' questions from large collections of documents (millions or billions of documents). NIST provided the framework to evaluate and monitor progress of developed large-scale, open-domain Question Answering systems. The open-domain label referred to the generality of the topics covered by the collections of documents, typically topics of general interest such as news articles or government documents, and less to the ability of systems to answer any questions in any domain.

The Question Answering track focused solely on factual questions of the form *Who is the voice of Miss Piggy?* (Answer: Frank Oz) or *How much could you rent a Volkswagen bug for in 1966?* (for the curious reader the answer is $1 per day). Participants in the QA track were given a set of test questions (200 at TREC-8 in 1999) and a large collection of documents (~5 GB of text). They were supposed to provide a ranked list of 5 answers in the form of generated answers or simply excerpts from documents where the answers are located. Answers were of two types: short (up to 50 characters) and long (up to 250 characters). An answer was considered correct if a person reading it would consider it so. The output of all participants was automatically compared to a gold standard of correct answers (collected and/or checked manually by human judges) using regular-expressions-based matching. The scoring rubric was so designed to reward system's ability to provide correct answers in the first positions of the ranked list of answers. The score assigned to each test question was the reciprocal of the rank of the first correct answer in the list of 5 answers. For instance, a correct answer at the top of the list was given a perfect score of 1 while a correct answer in the second position would be assigned a score of 0.5, and so on. An overall mean reciprocal rank (MRR) score was computed for the set of test questions by averaging the ranking scores of the individual questions. In the first TREC Question Answering challenge, the best systems obtained MRR scores ~0.5-0.6 for the short answer category (50 bytes at most per answer), meaning they provide the first correct answer, on average, in the second position in the list of 5 ranked answers.

Since the first Question Answering challenge was offered in 1999 at TREC-8, many systems have been developed and much research has been conducted exploring various steps (e.g. question classification or answer justification) of the Question Answering process. The major approaches to answering factual questions range from shallow (Moldovan et al., 1999) to deep (Rus, 2002). Shallow approaches use shallow linguistic methods combined with heuristic-based scoring techniques to locate and rank answers. Deep approaches rely on world knowledge and inference mechanisms to retrieve correct answers. Deep approaches can offer justify the correctness of the answers using logical explanations (Rus, 2002).

There are several important outcomes resulting from NIST's decade-long QA track. First, there is a better understanding of factoid questions and factoid Question Answering processes. Second, many successful Question Answering systems were developed and are now available in many languages. Third, Question Answering research led to better search engines, which now can take queries in the form of natural language questions and provide a ranked list of short answers instead of just documents. Fourth, high-performance Question Answering systems are being built, culminating with the recent announcement by I.B.M. of its DeepQA system, codename Watson. Watson is intended to be precise and fast enough to compete in real-time with top human contestants of the Jeopardy! game.

While a lot of progress has been made in the area of automated Question Answering that focuses on factoid questions (with all its variants such as simple factoid questions, list factoid questions where the answer is a list of items such as names, multilingual Question Answering, etc.), there is a long road ahead to develop Question Answering systems that could answer any type of questions in any domain for every user. In particular, there is need to develop methods to automatically handle deep questions or to provide answers tailored to a particular user's background. Answers delivered to an expert should be a finer-grain level of detail and using domain-specific language while answers provided to novices should be more general using plain language.

QUESTION ASKING

While research on Question Answering has been flourishing over the last decade, there was less attention paid to automatic Question Asking/Generation despite the fact that applications of automated QG facilities will be far reaching. Sample applications of automated QG facilities include the following:

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.

To enable such applications, QG technologies must be developed in a more systematic and large-scale manner, building 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 (computational linguistics, artificial intelligence, human-computer interaction, information retrieval).

Early Question Generation and Question Asking Research. Early explorations on QG were sporadic and less systematic. Cognitive science and education researchers paid more attention to Question Generation, which they called *Question Asking*, than other communities because Question Asking has frequently been considered a fundamental cognitive process. The ideal learner is an active, self-motivated, creative, inquisitive person who asks deep questions and searches for answers to such thought-provoking questions. There is a long history of researchers who have advocated learning environments that support inquiry learning and question asking. Question Asking is one of the processing components that underlies higher level cognitive activities, such as comprehension, problem solving, and reasoning.

Existing research on Question Asking has frequently embraced the notion that clashes between stimulus input and world knowledge are very much at the essence of Question Asking/Generation. Thus, questions are asked when there are contradictions, anomalous information, obstacles to goals, uncertainty, and obvious gaps in knowledge. Although it is widely acknowledged that discrepancies between input and knowledge trigger questions, the precise mechanisms need to be specified in more detail than has been achieved in psychology and education.

The field of Artificial Intelligence (AI) has offered computational models that make some attempt to specify the knowledge representations and knowledge discrepancies that underlie Question Asking/Generation. According to Schank's SWALE model (Schank, 1986), for example, questions are asked when we observe anomalous events and request explanations for such events (e.g., *Why did the event occur?*). Long-term memory is viewed as a large inventory of cases that record anomalous events and their associated explanations, which are driven by *why*, *what-if*, and other deep questions.

Models of Question Asking/Generation in AI have excelled in analytical detail and computational precision, but the next step to evaluate whether the models explain Question Asking in humans has never been taken. In contrast, the fields of psychology and education have empirically tested general theoretical claims about Question Asking, but they have underachieved in formulating the precise conditions, knowledge representations, and computational mechanisms that generate the questions.

The PREG model of Question Asking (Otero & Graesser, 2001) reduced this gap between the two enterprises. It is a comprehensive analytical model of Question Asking/Generation that incorporates the mechanisms that have been identified in Education, Psychology, Discourse Processing, and Artificial Intelligence. The PREG model was used to predict the particular questions that children and adults would ask when they read expository texts on scientific phenomena. The predicted questions are sensitive to four information sources or processing components: (1) the explicit text, (2) the reader's world knowledge about the topics in the text, (3) the reader's metacognitive skills, and (4) the reader's knowledge about the pragmatics of communication. The PREG model has received some support in psychological investigations, but more research is needed to determine the precise conditions under which particular questions are generated. This step can only be achieved by interdisciplinary efforts that include the fields of Psychology, Education, Artificial Intelligence, and Computational Linguistics. Such efforts would be greatly facilitated by large-scale investments in research infrastructure that would allow researchers to focus on the real issues of Question Generation/Asking rather than on building supporting resources. This project is an important step toward building much-needed QG research infrastructure.

Attempts at QG in Natural Language Processing/Computational Linguistics can be categorized into three groups: *Query/Question Reformulation*, *pseudo Question Generation*, and *simple Question Generation*. In *Query/Question Reformulation* the task is to generate queries/questions given a question as input. A representative system for Query/Question Reformulation is the CO-OP system (McKeown, 1983), which implements question reformulation through paraphrasing in the context of database access. CO-OP's goal was to allow non-experts to send queries to a database management system using natural language–based questions. Hoshino and Hiroshi (2005) use the term Question Generation to refer to sentences with gaps in multiple-choice language tests. In these tests the student is asked to choose, from a list of options, the word that best fills the gap in a given sentence so that the most appropriate sentence is formed. Since the generated output is not a question, we consider this work to be *pseudo Question Generation* rather than true Question Generation. Mitkov and Ha (2003) developed a computer-aided procedure for generating multiple-choice questions, given a set of documents on a topic. The procedure first identifies candidate concepts and sentences in which they appear. Given the sentence, it uses a shallow parser, transformational rules, and WordNet (a lexical database of English; Miller, 1995) to map the sentence into its interrogative form. The system can only handle sentences whose structure is SVO or SV (S-Subject; V-Verb; O-Object). For example, an SVO sentence is transformed into *Which HVO? question*, where H is a *hypernym* (a more general term) of the S term (e.g., *vehicle* is a hypernym of *car*). Due to its limited scope, we call Mitkov and Ha's work *simple Question Generation.*

Given the relative isolation of researchers from the various communities working on Question Generation/Asking, three years ago, Rus, Cai, and Graesser (2007) advanced the idea of a workshop to bring Question Generation/Asking researchers together to form a community that will further the field in a more coherent and systematic way.

Bringing Together the Question Generation Research Community (2008–2010). With NSF support, the 1st Workshop on Question Generation took place in September 2008 in Arlington, VA. The workshop was attended by 29 participants from all over the world, from both academia and industry (e.g., Microsoft, Yahoo). Since then, two other workshops were organized in 2009 and 2010 with an ever-growing number of participants (33 participants in 2009 and 35 participants in 2010). Also, in 2010 the first Question Generation Shared Task Evaluation Campaign (QG-STEC) was organized by The University of Memphis and Open University (United Kingdom). Two tasks were offered: Task A, Question Generation from Paragraphs, and Task B, Question Generation from Sentences. These tasks evolved naturally from the 1st Workshop on Question Generation that identified four categories of QG tasks (Rus & Graesser, 2009): Text-to-Question, Tutorial Dialogue, Assessment, and Query-to-Question. Tasks A and B in the first QG-STEC are part of the Text-to-Text category of Natural Language Generation tasks identified by the NLG community (Rus et al., 2007; Dale & White, 2007).

Current Question Generation Research. Automated QG is currently seen as a discourse task (Rus & Graesser, 2009) that is beyond a traditional natural language generation task. Automated QG consists of three major steps: (1) target concept identification, (2) question type selection, and (3) question construction. Whether the three main steps in the QG processing pipeline are independent and follow each other sequentially or can be executed in parallel is still an open question. It may be the case that they can be executed both sequentially and in parallel, in which case the exact conditions under which one type of execution is more appropriate are yet to be identified. Importantly, the primary automated QG model can be informed by studies on human question asking processes and refined and validated through further research, which will be greatly facilitated by infrastructure investments such as the one proposed here. For instance, it is documented that for human readers, identification of target content and the subsequent decision on question type do not appear to be primary processes, carried out in an independent, sequential manner. Rather, they seem to be directly dependent on cognitive disequilibrium (Dillon, 1990; Graesser & McMahen, 1993). Indeed, deep question generation in human readers is essentially triggered by the existence of an obstacle that prevents a reader with particular domain knowledge from achieving a discourse representation goal (Otero, 2009; Otero & Graesser, 2001; Sullins et al., 2010). Therefore, reading goals, and the obstacles found by a reader in achieving them, play a primary role in the definition of target content and type of questions asked.

The Question Generation Shared Task Evaluation Campaign. An important activity of the QG research community was the organization during Spring 2010 of the first Question Generation Shared Task Evaluation Challenge (QG-STEC). The QG-STEC followed a long tradition of STECs in Natural Language Processing: see various tracks at the Text REtrieval Conference (TREC; trec.nist.gov), e.g., the Question Answering track mentioned earlier. In particular, the idea of a QG-STEC was inspired by the NLG community's goal to offer shared task evaluation campaigns as a potential avenue to encourage focused research efforts (White & Dale, 2008). Along the same lines, a QG-STEC should provide an exciting opportunity for the QG research community to work together toward a clear, well-defined research goal. The tasks in the first QG-STEC were defined to attract as many participants as possible. To achieve this goal, two principles were followed: application neutrality and no representational commitment (see Rus et al., 2010 for more details). While application neutrality makes it more difficult (but not impossible) to judge the importance of generated questions, the alternative of choosing one application, e.g. tutoring systems, and running the QG-STEC in that context would have limited the pool of participants to those working in that application area. It should be noted that application-independent STECs have been offered in the past in other fields, e.g., generating generic summaries (as opposed to query-specific summaries).

Regarding input representation, both Tasks A and B in the first QG-STEC used raw text to avoid representational commitments. Future QG-STECs will aim to provide annotated texts so that participants can focus on the fundamental targeted task of QG and do minimal extra work such as text understanding on the input. Deciding what annotation(s) will be offered in future QG-STECs is one key issue to be discussed during the QG research infrastructure planning process.

The two tasks offered in the first QG-STEC were selected by the members of the QG community among five candidate tasks. A consensus-reaching process was implemented to decide the two tasks. Community members were invited to proposed tasks and then a preference poll was conducted to select the two most preferred tasks. Question Generation from Paragraphs (Task A; The University of Memphis) and Question Generation from Sentences (Task B; Open University) were chosen to be offered in the first QG-STEC. The other three candidate tasks were Ranking Automatically Generated Questions (Michael Heilman and Noah Smith, Carnegie Mellon University), Concept Identification and Ordering (Rodney Nielsen and Lee Becker, University of Colorado), and Question Type Identification (Vasile Rus and Arthur Graesser, The University of Memphis). These five candidate QG tasks will form a main thread of discussions during the infrastructure specification process that we will undertake in this project because they are all candidates for future QG-STECs. Notably, all teams who proposed the five tasks (University of Memphis, Open University, University of Colorado, and Carnegie Mellon University) have agreed to play leadership roles in this infrastructure planning project.

There was some overlap between Tasks A and B in the first QG-STEC. The overlap was intentional, with the aim of encouraging people preferring one task to participate in the other too. The overlap consisted of one type of questions in Task A, the *specific questions*, being somehow similar to the type of questions targeted by Task B. The difference was the fact that in Task B, participants were asked to generate specific questions of a certain type, e.g., *Who*, while in Task A there was no such constraint.

The young QG research community is very active with many activities being undertaken by various research groups. We describe next research conducted by leading and active members of the QG community. Other QG research exists, but it is beyond the scope of this proposal to describe it all (Ureel II et al., 2005; Hallett, Scott, & Power, 2007; Lee & Seneff, 2007).

The University of Memphis. The University of Memphis has been involved in QG research for several decades (Graesser, Person, & Huber, 1992; Graesser & Person, 1994; Otero & Graesser, 2001; Rus et al., 2007; Rus & Graesser, 2009; Rus et al., 2010). 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, as described earlier.

More recently, we developed an automated QG system originally motivated by work in building AutoTutor (Graesser et al., 2004, 2005), an intelligent tutoring system that helps students learn by holding a conversation in natural language. The core pedagogical strategy of AutoTutor is called Expectation and Misconception-Tailored Dialogue. The dialogue manager tries to get the learner to articulate a number of sentence-like answers (called expectations). AutoTutor gets the student to do the talking by giving generic pumps ("tell me more"), hints (e.g., "What about X?", "What is the relation between X and Y?"), and question prompts ("After release, the horizontal force on the packet is what?") for the student to fill in specific words. A QG Mark-up Language (QGML) was also developed that can be used to automatically generate questions from sentence-like expectations (Cai et al., 2006; Rus, Cai, & Graesser, 2007).

The University of Memphis has played a leadership role in building the QG research community (Rus et al., 2007; Rus & Graesser, 2009; Rus & Lester, 2009) as well as in organizing and running the first QG-STEC (Rus et al., 2010). The University of Memphis ran Task A, Question Generation from Paragraphs, as part of the first QG-STEC.

Open University. Research on QG at the Open University (OU) is carried out in the Natural Language Generation Group of the Centre for Research in Computing (one of the top 20 Computing Research Centres in the U.K. in the 2008 Research Assessment Exercise).

Dr. Piwek pioneered research on Text-to-Dialogue (T2D) generation based on discourse relations (Piwek et al., 2007). He currently holds two grants in which QG plays a central role. The CODA project (Piwek & Stoyanchev, 2010; Stoyanchev & Piwek 2010a, b) investigates the automatic generation of dialogue text (akin to FAQs) from monologue. This includes the automatic generation of questions from text in monologue form. This effort includes collaboration with the National Institute of Informatics in Tokyo, which focuses on creating computer-animated movies from dialogue text.

The DataMIX project brings together U.K. researchers from seven different disciplines (including computational linguistics, design, theoretical physics, chemistry, and music computing) to investigate more inclusive ways of presenting data and information. The OU contribution includes an investigation into the usefulness of presenting data in question-answer form to different users (including lay persons versus experts, users with visual impairments, and users affected by the digital divide).

Members of the OU team have participated and presented work at all QG workshops to date, including a paper on an open-source QG system (Wyse & Piwek, 2009). In 2010, the OU team was involved in the organization of the first QG-STEC (Task B - Generating Questions from Sentences). Dr. Piwek was co-chair of the Third Workshop on QG and is currently guest-editing a special issue on QG.

Carnegie Mellon University. Heilman and Smith (2010b, 2009) have been focusing on QG for the creation of educational materials for reading practice and assessment. They aim to create an automated system that can take as input a text (e.g., an article that a student might read for homework or during class) and produce as output a ranked list of questions for assessing students' knowledge of information in the text. Their QG system leverages existing NLP tools and formalisms to solve various linguistic challenges. Using manually written rules that encode linguistic knowledge about question formation, it generates a large set of candidate questions. These candidate questions are then ranked by a statistical model of question quality to address semantic and pragmatic issues that are not easily captured with rules. Significant infrastructure resources, such as large annotated corpora, would allow the development of machine-learning-based QG models, e.g. for learning question formation rules thus eliminating the need to manually create such rules which is both time-consuming and error-prone.

Heilman and Smith proposed one of the five initial tasks for the first QG-STEC, Ranking Automatically Generated Questions.

North Carolina State University. Kristy Boyer and James Lester focus on automatic QG in task-oriented dialogue. Their JAVATUTOR project (Boyer, Vouk, & Lester, 2007; Boyer et al., 2009) aims to create adaptive task-oriented dialogue systems by using data-driven approaches that learn dialogue policies from corpora. Research has shown that the effectiveness of human tutorial dialogue is facilitated by rich tutor-student interactions (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001), students' self-explanations (Chi, Leeuw, Chiu, & LaVancher, 1994), and tutors' targeted, adaptive feedback (Forbes-Riley & Litman, 2009; Fossati et al., 2010). To best support these activities, tutorial dialogue systems must be able to generate questions. For example, to engage in mixed-initiative interaction with a student, a tutoring system must be able to not only answer student questions, but pose questions as well. Similarly, supporting students' self-explanations also involves asking questions that elicit those explanations, especially deep questions (Graesser & Person, 1994).

Most tutorial dialogue systems handle QG with a set of hand-authored questions or question templates, which offer limited flexibility and content coverage. An automated, data-driven approach to QG would address these limitations. Automatic QG therefore holds significant promise for increasing the effectiveness of tutorial dialogue systems, including task-oriented tutorial dialogue systems such as JAVATUTOR (Boyer et al., 2009). QG infrastructure would greatly enable the development of effective tutorial dialogue systems.

To facilitate research on automatic QG within task-oriented tutorial dialogue, additional annotated corpora are needed, ideally in multiple domains. These corpora would consist of successful human-human textual or spoken dialogue annotated with dialogue act tags that capture cognitive and affective aspects of each utterance. Task structure would also be tagged, as would the correctness of students' task actions (Boyer et al., 2010; Fossati et al., 2010; Lane, 2004). With such annotated corpora, it would be possible to explore several important research questions: (1) When should the tutor ask questions? (2) What should the question topics be? (3) What are potential surface realizations of the question? The answers to these questions will contribute to creating the next generation of tutorial dialogue systems that are more flexible, and more effective, than their predecessors.

University of Colorado. Rodney Nielsen and colleagues at The University of Colorado are researching QG and Question Answering in a variety of contexts including intelligent tutoring systems, other educational settings, clinical informatics, and personal dialogue. They have been active contributors to the recent QG research efforts including the QG workshops (Becker et al., 2009; Becker et al., 2010; Nielsen, 2008; Nielsen et al., 2008). Nielsen was part of the steering committee for the first QG-STEC, led the evaluation methods and metrics team at the first QG workshop, and proposed one of the five initial tasks for the QG-STEC, the Concept Identification and Ordering task.

The University of Colorado's research on QG focuses on several basic questions including those related to the first step in the QG processing pipeline, as described next. Target Concept Identification (TCI) is a critical subtask in QG (Nielsen, 2008; Vanderwende, 2008) - if the concept is not important, is the question really worth asking? TCI requires deciding which concepts in general are worth discussing, what their relative merit is, how they depend on one another, and how to elicit them from the learner or help the learner discover and comprehend them given the current dialogue context.

While Target Concept Identification is performed during the dialogue and is a context-sensitive task, it is also important to identify a priori the set of key question-worthy concepts in the knowledge source. Given the application domain, the objective of Key Concept Identification (KCI) is to identify the most important content in the text, that for which questions should be or are likely to be generated. We believe this task is much less sensitive to the application domain and would thus appeal to a broader research community. As part of the planning grant, we will consider the issues discussed above in the context of existing dialogue systems and decide whether it is reasonable or necessary to restrict initial corpus development to KCI versus the context-sensitive task of TCI. Related issues, such as Concept Sequencing (defining a logical sequence of concepts) and Concept Relation Identification and Classification (the detection and labeling of inter-concept relationships) will also be considered.

The University of Pennsylvania. The University of Pennsylvania's (UPenn) research team (Rashmi Prasad, Aravind Joshi) participated in all three QG workshops and also in Task A of the first QG-STEC. Their research focus includes the generation of higher-order questions, similar to the medium/discourse-level questions in QG-STEC's Task A. As opposed to questions generated from the content of a single clause, higher-order questions are derived from "discourse relations" that hold between clausal and sentential arguments.

Discourse relations in a given input should be reliably pre-specified in future QG-STECs so that participants can focus on the higher-order QG challenges rather than on developing the discourse-level resources needed to identify the discourse relations. In keeping with this idea, the first QG-STEC provided the discourse relation annotations using an existing automatic discourse parser (HILDA, duVerle & Prendinger, 2009) based on Rhetorical Structure Theory (RST, Mann and Thompson, 1988). The discourse relations thus obtained were not as reliable as hoped.

As part of planning the QG infrastructure, the goal is to explore high-precision shallow (or low-level) discourse parsing that has been the focus of much recent research in NLP. Mannem et al (2010) have argued that for QG, not all discourse relations are required as the input knowledge and that the complete discourse structure of a text need not be specified. Indeed, much of the recent research on shallow discourse parsing (Dinesh et al., 2005; Wellner et al., 2007, 2009; Pitler et al., 2009; Pitler & Nenkova, 2009; Prasad et al., 2010) exploits the Penn Discourse Treebank (PDTB) corpus (Prasad et al., 2008), which contains low-level annotations of discourse relations.

Rashmi Prasad and Aravind Joshi at UPenn have led the development of the PDTB and have the necessary extensive experience and expertise for the annotation of discourse relations. They will explore a semi-automatic annotation approach towards developing the discourse-level infrastructure for QG.

Research By Other Team Members. Calvo and colleagues explore the role of question generation in supporting writing activities (Liu & Calvo, 2009; Liu, Calvo, & Rus, 2010). Hirashima (2009) automatically generates physics problems and study their impact on students' problem solving skills. Automated Question Generation is being used to scaffold self-questioning strategies automatically to help children in grades 1-3 understand informational text (Chen, Aist, & Mostow, 2009).

FUTURE RESEARCH DIRECTIONS

It is important to note that our emphasis was on automated Question Answering and Generation processes that try to mimic the typical human Question Answering process of asking questions in natural language and receiving a similar answer. Asking questions can be done a myriad of other forms and systems that handle these different ways of asking questions are currently built. Questions can be asked by consulting Frequently Asked Questions list, creating questions using a limited number of primitives available as GUI (Graphical User Interface) elements, Point & Query facilities where the users point to (or clicks on) a hot spot on the display and then a menu of contextually-relevant questions appears (see Graesser et al., 2002 for an example), questions that are asked by a combination of speech, gesture, and gaze, or questions inferred from someone's actions. It is beyond to the scope of this chapter to provide an overview of all these forms of inquiry and answering systems.

An interesting development is the creation of community-based Question Answering (cQA) services offered by major Internet search companies, e.g. Yahoo!Answers, in which users can ask questions that are being answered voluntarily by other users. The asker and other users rate the answers resulting in answer filtering and ranking process that will eventually bring the best answers at the top.

CONCLUSION

Question Answering and Generation are two important challenges for the Google generation. Since the advent of the Internet, the information focus shifted from access to information to information overload. Effective and natural retrieval methods such as Question Answering have been explored and developed. We can now effectively provide answers to factoid questions that seek specific nuggets of information that are well-defined. Providing user-tailored answers to deep questions is still an open problem that waits to be explored at some point in the future. Asking factoid questions presupposes knowing what you are looking for. When you do not know what to ask for, as in learning, the reverse process of asking questions takes central stage. Question Asking or Generation has been explored sporadically until few years ago. The recently created Question Generation community provides a coherent and systematic framework for advancing the field and impact on other fields, such as learning technologies. Eventually, Question Answering and Generation will work in tandem as important features of future information delivery approaches, leading to both livable-documents or textbook that can both give you quick answers are help you ask question while lecturing them.

REFERENCES[i]

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

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

Graesser, A.C., Hu, X., Person, N.K., Jackson, G. T., Toth, J. (2002). *Modules and Information Retrieval Facilities of the Human Use Regulatory Affairs Advisor (HURAA)*. The 7th annual world conference on E-learning in Corporate, Government, Healthcare, & Higher Education. AACE: Montreal, Canada.

Graesser, A., Ozuru, Y., & Sullins, J. (2009). What is a good question? In M. G. McKeown & L. Kucan (Eds.), *Threads of coherence in research on the development of reading ability* (pp. 112-141). NY: Guilford.

Lehnert, W. G. (1977). The process of question answering, Yale University, New Haven, CT, 1977.

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R., & Rus, V. (1999). LASSO: A Tool for Surfing the Answer Net, in Proceedings of the Text Retrieval Conference (TREC-8), November, 1999.

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

Rus, V. (2002). Logic Form For WordNet Glosses and Application to Question Answering, Computer Science Department, School of Engineering, Southern Methodist University, **PhD Thesis**, May 2002, Dallas, Texas.

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, Arlington, VA, April 20-21, 2007.

Rus, V. & Graesser, A.C. (Eds.). (2009a). The Question Generation Shared Task and Evaluation Challenge. ISBN: 978-0-615-27428-7.

Rus, V. & Lester, J. (Eds.). (2009b). Proceedings of the 2nd Workshop on Question Generation. July 6, 2009, Brighton, UK.

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

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

Voorhees, E. M. & Tice, D.M. (2000). The TREC-8 question answering track evaluation. In E.M. Voorhees and D.K. Harman, editors, Proceedings of the Eighth Text REtrieval Conference (TREC-8). Electronic version available at http://trec.nist.gov/pubs.htrul, 2000.

Woods, W. A., Kaplan, R. M. and Webber, B. N., 1972. *The Lunar Sciences Natural Language Information System: Final Report*, BBN Report 2378, Bolt Beranek and Newman, Inc., Cambridge, MA.

Waltz, D. L., 1978. "An English Language Question Answering System for a Large Relational Database," *Communications of the ACM*, 21(7):526-539.

KEY TERMS & DEFINITIONS

Question Answering: The task of providing answers from large collection of documents to naturally-asked questions.

Question Generation: The task of generating questions from various inputs such as raw text, semantic representation, or databases.

[i] References should relate **only** to the material you actually cited within your chapter (this is not a bibliography). References should be in APA style and listed in alphabetical order. Please do not include any abbreviations.