Automatic Question Generation for Educational Applications – The State of Art

Nguyen-Thinh Le[1], Tomoko Kojiri[2], Niels Pinkwart[1]

{nguyen-thinh.le, niels.pinkwart}@hu-berlin.de
[1]Humboldt Universität zu Berlin, Germany

kojiri@kansai-u.ac.jp
[2]Kansai University, Japan

Abstract. Recently, researchers from multiple disciplines have been showing their common interest in automatic question generation for educational purposes. In this paper, we review the state of the art of approaches to developing educational applications of question generation. We conclude that although a great variety of techniques on automatic question generation exists, just a small amount of educational systems exploiting question generation has been developed and deployed in real classroom settings. We also propose research directions for deploying the question technology in computer-supported educational systems.

Keywords: automatic question generation, educational technology

1 Introduction

Recently, the research area of automatic question generation for educational purposes has attracted attention of researchers from different disciplines. Question generation is defined by Rus et al. (2008) as follows: "Question generation is the task of automatically generating questions from various inputs such as raw text, database, or semantic representation". This definition indicates that the type of input for question generation can vary: it can be, for example, a sentence, a paragraph or a semantic map. According to Piwek and Boyer (2012), research on question generation has a long tradition and can be traced back to the application of logic to questions. One of the first works on questions was proposed by Cohen (1929) to represent the content of a question as an open formula with one or more unbound variables. While research on question generation has been being conducted for long time, deploying automatic question generation for educational purposes has raised interests in different research communities in recent years.

Studies have reported that deploying questions in teaching encourages students to self-explain, which has been shown to be highly beneficial for learning (Chi et al., 1994). With novice computer scientists, asking effective questions during the early phases of planning a solution can support the students' comprehension and decompo-

adfa, p. 1, 2011.

sition of the problem at hand (Lane, 2005). Asking targeted, specific questions is useful for revealing knowledge gaps with novices, who are often unable to articulate their questions (Tenenberg & Murphy, 2005). In addition, in the view of improving meta-cognitive abilities, asking students to generate questions by themselves may enable students to construct meaningful knowledge and to employ various meta-cognitive strategies by themselves (Yu et al., 2005).

In this paper, we survey the state of the art of educational systems which deploy questions. The goal of this paper is to answer the following questions: 1) which methodologies can be applied to generate questions? 2) How can questions be deployed in educational settings?

2 Method

In order to answer the research questions introduced in the previous section, we will select peer-reviewed scientific reports on question generation systems for educational purposes. That is, not only complete systems but also work in progress will be taken into account in this paper.

We will group educational applications of question generation into the same class if they follow the same educational purpose and discuss their technical approaches. Since the aim of this review is to find out the current deployment of educational applications of questions, evaluation studies of existing works will be summarized.

3 Question Generation

3.1 Educational Applications of Question Generation

In this subsection, we review the current educational systems deploying question generation. The systems are classified into three classes according to their educational purposes: 1) knowledge/skills acquisition, 2) knowledge assessment, and 3) educational systems that use questions to provide tutorial dialogues.

First class: Knowledge/skills acquisition

The purpose of the first class of educational applications of question generation includes knowledge/skills acquisition. One of the first automatic question generation systems which have been developed to support learning was proposed by Wolfe (1976). The author proposed a system called AUTOQUEST to help novices learn English. Questions are generated from reading sources provided to students. Kunichika et al. (2001) applied a similar approach based on syntactic and semantic information extracted from an original text. Their educational purpose was to assess the grammar and reading comprehension of students. The extracted syntactic features include subject, predicate verb, object, voice, tense, and sub-clause. The semantic information contains three semantic categories (noun, verb and preposition) which are used to determine the interrogative pronoun for the generated question. For example,

in the noun category, several noun entities can be recognized including person, time, location, organization, country, city, and furniture. In the verb category, bodily actions, emotional verbs, thought verbs, and transfer verbs can be identified. The system is also able to extract semantic relations related to time, location, and other semantic categories, when an event occurs. Evaluations showed that 80% of the questions were considered as appropriate for novices learning English by experts and 93% of the questions were semantically correct.

Mostow and Chen (2009) developed an automated reading tutor which deploys automatic question generation to improve the comprehension capabilities of students while reading a text. The authors investigated how to generate self-questioning instruction automatically on the basis of statements about mental states (e.g., belief, intention, supposition, emotion) in narrative texts. The authors proposed to decompose the instruction process into four steps: describing a comprehension strategy, modeling its use, scaffolding its application, and prompting the child (who is the user of the reading tutor) to use it. The step of describing a comprehension strategy aims at explaining the user when to apply self-questioning. Then, the reading tutor poses a question about the sentence of a reading text in order to illustrate the use of self-questioning. During the step of scaffolding the application of self-questioning, the tutor system helps the child construct a question by choosing from four characters in the context of the given reading text from the on-screen menu (e.g., the town mouse, the country mouse, the man of the house, the cat), three question types (Why, What, How), and three question completers on a menu-driven basis. The reading tutor gives positive feedback in case the child constructed a correct question and invites the child to try again in case she/he created counterfactual questions. The step of prompting the child to use the self-questioning strategy encourages her/him to develop a question and to find an appropriate answer from the given text. The reading tutor has been evaluated with respect to the acceptability of menu choices (grammatical, appropriate, and semantically distinct), to the acceptability of generated questions, and to the accuracy of feedback. The authors reported that only 35.6% of generated questions could be rated as acceptable. 84.4% of the character choices and 80.9% of the question completer choices were classified as acceptable. However, the accuracy of detection of counterfactual questions was 90.0% which is high for generating plausible feedback. Applying a similar approach, Chen and colleagues (Chen et al., 2009) developed a reading tutor for informational texts for which another set of question templates need to be defined.

Also in the same class of educational applications, Liu and colleagues (Liu et al., 2012) introduced a system (G-Asks) for improving students' writing skills (e.g., citing sources to support arguments, presenting the evidence in a persuasive manner). The approach to generating questions deployed in this system is template-based. It takes individual sentences as input and generates questions for the following citation categories: opinion, result, aim of study, system, method, and application. The process of generating questions consists of three stages. First, citations in an essay written by the student are extracted, parsed and simplified. Then, in the second stage, the citation category is identified for each citation candidate. In the final stage, an appropriate question is generated using pre-defined question templates. For example, for the cita-

tion category “opinion”, the following question templates are available: “Why +subject_auxiliary_inversion()?”, “What evidence is provided by +subject+ to prove the opinion?”, “Do any other scholars agree or disagree with +subject+?”. In order to instantiate these question templates, the “+subject_auxiliary_inversion()” operation places the auxiliary preceding a subject, and the “+subject+” operation replaces the place holder with a correct value. The evaluation has been conducted with 33 PhD students-writers and 24 supervisors. Each student has been asked to write a research proposal. Each proposal was read by a peer and a supervisor, who both were asked to give feedback in form of questions. In total, questions were produced from four sources for each student’s proposal: questions generated from a supervisor, from a peer, from the G-Asks system, and from a set of five generic questions: 1) Did your literature review cover the most important relevant works in your research field? 2) Did you clearly identify the contributions of the literature reviewed? 3) Did you identify the research methods used in the literature reviewed? 4) Did you connect the literature with the research topic by identifying its relevance? 5) What were the author’s credentials? Were the author’s arguments supported by evidence? Each question producer generated a maximum of five questions. Students evaluated 20 questions at most for each student’s proposal based on five quality measures: 1) grammatical correctness, 2) clearness, 3) appropriateness to the context, 4) helpfulness for reflecting what the author has written, 5) usefulness. Evaluation studies have reported that the system could generate questions as useful as human supervisors and significantly outperformed human peers and generic questions in most quality measures after filtering out questions with grammatical and semantic errors (Liu et al., 2012).

In the contrast to approaches to generating questions using text as an input, Jouault and Seta (2013) proposed to generate semantics-based questions by querying semantic information from Wikipedia database to facilitate learners’ self-directed learning. Using this system, students in self-directed learning are asked to build a timeline of events of a history period with causal relationships between these events given an initial document (that can be considered a problem statement). The student develops a concept map containing a chronology by selecting concepts and relationships between concepts from the given initial Wikipedia document to deepen their understandings. While the student creates a concept map, the system also integrates the concept to its map and generates its own concept map by referring to semantic information of Wikipedia. The system’s concept map is updated with every modification of the student’s one. In addition, the system extracts semantic information from DBpedia (Bizer et al., 2009) and Freebase (Bollacker, 2008) which contains semantic representation of Wikipedia in order to select and add related concepts into the existing map. Thus, the system’s concept map always contains more concepts than the student’s map. Using these related concepts and their relationships the system generates questions for the student to lead to a deeper understanding without forcing to follow a fixed path of learning.

Second class: Knowledge Assessment.

The second class of educational applications of question generation aims at assessing knowledge of students. Heilman and Smith (2010) developed an approach to generat-

ing questions for assessing students' acquisition of factual knowledge from reading materials. The authors developed general-purpose rules to transform declarative sentences into questions. The approach includes an algorithm to extract simplified statements from appositives, subordinate clauses, and other constructions in complex sentences of reading materials. Evaluation studies have been conducted to assess the quality and precision of automatic generated questions using Wikipedia and news articles. The authors evaluated the question generation approach with 15 English speaking university students who were asked to rate the system generated questions with respect to a list of deficiencies (ungrammatical, no sense, vagueness, obvious answer, missing answer, wrong "WH"-word, formatting errors, e.g., punctuation, and others). The participants of the evaluation study were asked to read a text of an article and to rate approximately 100 questions generated from the text. The authors reported that their system achieved 43.3% precision-at-10[1] and 6.8 acceptable questions could be generated from a source text of 250 words (Heilman & Smith, 2010). However, no evaluation studies with respect to the contribution of generated questions for learning can be found in literature.

For the purpose of assessing vocabulary of students, there are several attempts to automatically generate multiple-choice closed questions. In general, the process of generating questions for vocabulary assessment is determining which words to remove from the source sentence. The purpose of this step is to emphasize which vocabulary students should learn. If multiple-choice questions are supposed to be generated, wrong alternative answers (also referred to as distracters) for a specific multiple-choice question are required. Mitkov and colleagues (Mitkov et al., 2006) developed a computer-aided environment for generating multiple-choice test items. The authors deployed various natural language processing techniques (shallow parsing, automatic term extraction, sentence transformation, and computing of semantic distance). In addition, the authors exploited WordNet, which provides language resources for generating distracters for multiple-choice questions. The question generation process of this system consists of three steps. First, key terms, which are nouns or noun phrases with a frequency over a certain threshold, are extracted using a parser. The second step is responsible for generating questions. For this purpose, a clause filtering module was implemented to identify those clauses to be transformed into questions. The clauses are selected if they contain at least a key term, are finite, and are of a Subject-Verb-Object structure or a Subject-Verb structure. In addition, transformation rules have been developed to transform a source clause to a question item. The third step is deploying hypernyms and coordinates (which are concepts with the same hypernym) in WordNet to retrieve concepts semantically close to the correct answer. If WordNet provides too many related concepts, only the ones which occur most frequently in the textbook (which is used for generating multiple-choice questions) are selected. The authors demonstrated that the time required for generating questions including manual correction was less than for manually creating questions alone (Mitkov et al., 2006):

[1] "We calculate the percentage of acceptable questions in the top N questions, or precision-at-N. We employ this metric because a typical user would likely consider only a limited number of questions." (Heilman & Smith, 2010).

For 1000 question items, the development cost would require 30 hours of human work using the system, while 115 hours would be required without using the system. In addition, the quality of test items which have been generated and post-edited by humans was scored better than those produced manually without the automatic support of the system.

For the educational purpose of assessing vocabulary, Brown and colleagues (Brown et al, 2005) developed the system REAP which is intended to provide students with texts to read according to their individual reading levels. The system chooses text documents which include 95% of words that are known to the student while the remaining 5% of words are new to the student and need to be learned. After reading the text, the student's understanding is assessed. The student's responses are used to update the student model in order to provide appropriate texts in the next lesson. The authors suggested six types of questions: definition, synonym, antonym, hypernym, hyponym, and cloze questions. In order to generate questions of these types, the system REAP uses data from WordNet. When a word is input in WordNet, it may appear in a number of synonym sets (or synsets): nouns, verbs, adjectives, and adverbs and a synset can be linked to other synsets with various relations (including synonym, antonym, hypernym, hyponym, and other syntactic and semantic relations). While the definition, synonym, antonym, hypernym, and hyponym question types can be created directly using appropriate synsets' relations, the cloze questions are created using example sentences or phrases retrieved from the gloss for a specific word sense in WordNet. Once a question phrase with the target word is selected, the present word is replaced by a blank in the cloze question phrase. In order to validate the quality of system-generated questions, the authors asked three researchers to develop a set of question types that could be used to assess different levels of word knowledge. Experimental results have reported that with automatically generated questions, students achieved a measure of vocabulary skill that correlates well with performance on independently developed human-generated questions.

Third class: Tutorial Dialogues

The third class of educational applications of question generation includes providing tutorial dialogues in a Socratic manner. Olney and colleagues (Olney et al., 2012) presented a method for generating questions for tutorial dialogue. This involves automatically extracting concept maps from textbooks in the domain of Biology. This approach does not deal with the input text on a sentence-by-sentence basis only. Rather, various global measures (based on frequency measures and comparison with an external ontology) are applied to extract an optimal concept map from the textbook. The template-based generation of questions from the concept maps allows for questions at different levels of specificity to enable various tutorial strategies, from asking more specific questions to the use of less specific questions to stimulate extended discussion. Five question categories have been deployed: hint, prompt, forced choice question, contextual verification question, and causal chain questions. Studies have been conducted to evaluate generated questions based on a rating scale between 1 (most) to 4 (least). All questions have been rated based on five criteria: 1) Is the question of the target type? 2) Is the question relevant to the source sentence? 3) Is the

question syntactically fluent? 4) Is the question ambiguous? 5) Is the question pedagogic? Results have been reported that the prompt questions (M=1.55) were significantly less to be of the appropriate type than the hint questions (M=1.2), and less to be of the appropriate type than the forced choice questions (M=1.27). Regarding the fluency, hint questions (M=1.56) were significantly more fluent than prompts (M=2.92), forced choice questions (M=2.64), contextual verification questions (M=2.25), and causal chain questions (M=2.4). With respect to pedagogy, hint questions were significantly more pedagogic than prompts (M=3.09), forced choice questions (M=3.21), contextual verification questions (M=3.3), and causal chain questions (M=3.18). With regard to the relevance of generated questions, there was no significant difference between the five question categories and the relevance lies between 2.13 and 2.88. Ambiguity scores (between 2.85 and 3.13) across the five question categories shows a tendency that questions were slightly ambiguous. Note that no results were available with regard to learning effectiveness through using generated questions.

Also with the intention of supporting students using conversational dialogues, Person and Graesser (2002) developed an intelligent tutoring system that improves students' knowledge in the areas of computer literacy and Newtonian physics using an animated agent that is able to ask a series of deep reasoning questions[2] according to the question taxonomy proposed by Graesser & Person (Graesser & Person, 1994). In each of these subjects a set of topics has been identified. Each topic contains a focal question, a set of good answers, and a set of anticipated bad answers (misconceptions). The system initiates a session by asking a focal question about a topic and the student is expected to write an answer containing 5-10 sentences. Initially, the system used a set of predefined hints or prompts to elicit the correct and complete answer. Graesser and colleagues (Graesser et al., 2008) reported that with respect to learning effectiveness, the system had a positive impact on learning with effect sizes of 0.8 standard deviation units compared with other appropriate conditions in the areas of computer literacy (Grasser et al., 2004) and Newtonian physics (VanLehn, Graesser et al., 2007). Regarding the quality of tutoring dialogues, the authors reported that conversations between students and the agent were smooth enough that no participating students left the tutoring session with frustration, irritation, or disgust. However, with respect to students' perception, the system earned averaged ratings, with a slightly positive tendency.

Lane & VanLehn (2005) developed PROPL, a tutor which helps students build a natural-language style pseudo-code solution to a given problem. The system initiates four types of questions: 1) identifying a programming goal, 2) describing a schema for attaining this goal, 3) suggesting pseudo-code steps that achieve the goal, and 4)

[2] Categories of deep reasoning questions:
Causal antecedent: What state or even causally led to an event or state?
Causal consequence: What are the consequences of an event or a state?
Goal-orientation: What are the goals or motives behind an agent's action?
Instrumental/procedural: What instrument or goal allows an agent to accomplish a goal?
Enablement: What object or resource allows an agent to perform an action?
Expectational: Why did some expected event not occur?

placing the steps within the pseudo-code. Through conversations, the system tries to remediate student's errors and misconceptions. If the student's answer is not ideal (i.e., it cannot be understood or interpreted as correct by the system), sub-dialogues are initiated with the goal of soliciting a better answer. The sub-dialogues will, for example, try to refine vague answers, ask students to complete incomplete answers, or redirect to concepts of greater relevance. For this purpose, PROPL has a knowledge source which is a library of Knowledge Construction Dialogues (KDCs) representing directed lines of tutorial reasoning. They consist of a sequence of tutorial goals, each realized as a question, and sets of expected answers to those questions. The KCD author is responsible for creating both the content of questions and the forms of utterances in the expected answer lists. Each answer is either associated with another KCD that performs remediation or is classified as a correct response. KCDs therefore have a hierarchical structure and follow a recursive, finite-state based approach to dialogue management. PROPL has been evaluated with the programming languages Java and C and it has been reported that students who used this system were frequently better at creating algorithms for programming problems and demonstrated fewer errors in their implementation (Lane & VanLehn, 2005).

Table 1. Existing question generation systems for educational purposes

Educational purpose	**System**	**Support type**	**Evaluation**
Developing knowledge/ skills	Wolfe (1976)	learning English	-
	Kunichika et al. (2001)	grammar and reading comprehension	Quality of questions
	Mostow & Chen (2009)	Reading tutor	Quality of questions
	Liu et al. (2012)	Academic Writing Support	Quality of questions
	Jouault & Seta (2013)	Self-directed learning support	-
Knowledge assessment	Heilman & Smith (2010)	Assessing factual knowledge	Quality of questions
	Mitkov et al. (2006)	Assessing vocabulary	Time effectiveness for generating questions
	Brown et al. (2005)	Assessing vocabulary	Quality of questions
Socratic dialogues	Olney et al. (2012)	Providing feedback in form of questions	Quality of questions
	Graesser et al. (2008)	Tutor for Computer literacy and Newtonian physics	Quality of questions, learning effectiveness, and students' perception
	Lane & VanLehn (2005)	Tutor for programming	Learning effectiveness

In summary, existing educational applications of question generation can be classified into three classes based on their educational purposes (Table 1): 1) question generation for knowledge and skills acquisition, 2) question generation for knowledge assessment, and 3) question generation for development of tutorial dialogues. From this table, we can notice that most educational applications of question generation fall into the first class. In addition, most evaluation studies focused rather on the quality of question generation than on the learning effectiveness contributed by the question generation component. In the next section, we compare different approaches to generating questions.

3.2 Approaches to Automatic Question Generation

Rus et al. (2008) regarded question generation as a discourse task involving the following four steps: 1) when to ask the question, 2) what the question is about, i.e., content selection, 3) question type identification, and 4) question construction.

The first issue involves strategies to pose questions. The second and the third issues are usually solved by most question generation systems in a similar manner using different techniques from the field of natural language processing such as parsing, simplifying sentences (Knight and Marcu, 2000), anaphor resolution (Kalady et al., 2010), semantic role labeling (Mannem, et al., 2010), or named entity recognizing (Ratinov and Roth, 2009).

While most question generation systems share common techniques on the second and third step of the process of question generation, their main difference can be identified when handling the fourth issue, namely constructing questions in grammatically correct natural language expression. Many question generation systems applied transformation-based approaches to generate well-formulated questions (Kalady et al., 2010; Heilman and Smith, 2009; Ali et al., 2010; Pal et al., 2010; Varga and Le, 2010). In principle, transformation-based question generation systems work through several steps: 1) Delete the identified target concept, 2) place a determined question key word on the first position of the question, and 3) convert the verb into a grammatically correct form considering auxiliary and model verbs. For example, the question generation system of (Varga and Le, 2010) uses a set of transformation rules for question formation. For subject-verb-object clauses whose subject has been identified as a target concept, a "Which Verb Object" template is selected and matched against the clause. For key concepts that are in the object position of a subject-verb-object, the verb phrase is adjusted (i.e., auxiliary verb is used).

The second approach for question formation, which is also employed widely in several question generation systems, is template-based (Wyse and Piwek, 2009; Chen et al., 2009; Sneiders, 2002). The template-based approach relies on the idea that a question template can capture a class of questions, which are context specific. For example, Chen et al. (2009) developed the following templates: "what would happen if <X>?" for conditional text, "when would <X>?" and "what happens <temporal-expression>?" for temporal context, and "why <auxiliary-verb> <X>?" for linguistic modality, where the place-holder <X> is mapped to semantic roles annotated by a semantic role labeler. Note that these templates have been devised to generate ques-

tions from an informational text. For narrative texts, Mostow and Chen developed another set of question templates (2009). Wyse and Piwek (2009) developed a similar approach which consists of rules and templates. Rules are represented in form of regular expressions and are used to extract key concepts of a sentence. Pre-defined questions templates are used to generate questions for those concepts. No evaluation results have been documented for this system.

Table 2 shows the evaluation results of different existing question generation systems. From the table we can notice that the template-based systems (Chen et al., 2009; Mostow and Chen, 2009) achieved considerable results, whereas there seems to be room for improvement of the transformation-based systems (Kalady et al., 2010; Pal et al., 2010; Varga and Le, 2010).

Table 2. Evaluation results of existing question generation systems

System	**Question type**	**Evaluation Results**
Kalady et al. (2010)	Yes-No, Who, Whom, Which, Where, What, How	Recall=0.68; Precision=0.46
Ali et al. (2010)	Yes-No, Who, Which, Where, What, When, How, Why	Recall=0.32; Precision=0.49
Varga & Le (2010)	Who, Whose, Whom, Which, What, When, Where, Why, How many	Relevance[3]=2.45(2.85); Syntactic Correctness & Fluency=2.85(3.1)
Mannem et al. (2010)	Who, When, What, Where, why, How	Low acceptance. No statistic data available.
Pal et al. (2010)	Yes-No, Who, When, Which, What, Why, How many, How much	Satisfactory results. No statistic data available
Chen et al. (2009)	Question templates for informational text	79.9% plausible questions[4]
Mostow & Chen (2009)	Question templates for narrative text	71.3 % plausible questions

In addition to texts as input for question generation, structured database can also be used. Sneiders (2002) developed question templates whose answers can be queried from a structured database. For example, the template "When does <performer> perform in <place>?" has two entity slots, which represent the relationship (Performer-perform-place) in the conceptual model of the database. Thus, this question template can only be used for this specific entity relationship. For other kinds of entity relationships, new templates must be defined. Hence, this template-based question generation approach is mostly suitable for applications with a special purpose. However, to develop high-quality templates, a lot of human involvement is expected. Another type of

[3] The evaluation criteria Relevance and Syntactic correctness and fluency are rated by from 1 to 4, with 1 being the best score. Values outside and inside in the brackets indicate ratings of the 1st and 2nd human.

[4] The evaluation results are calculated as the average of the plausibility percentage of three different question types: 86.7% (condition), 65.9% (temporal information), 87% (modality).

structured database as input for question generation is using semantic representation. Jouault and Seta (2013) deployed semantic representation of Wikipedia for question generation. They use ontological engineering and linked open data (LOD) techniques (Heath & Bizer, 2011) in order to generate semantics-based adaptive questions and to recommend documents according to Wikipedia to help students create concept maps for the domain of history. One of the great advantages of adopting semantic information rather than natural language resources is that the system can give adequate advice based on the machine understandable domain models without worrying about ambiguity of natural language.

4 Directions for Question Generation in Educational Systems

In Section 3.1 we have reviewed educational applications of question generation. We have identified eleven systems for knowledge and skill acquisition purposes, and only two of them have been evaluated with respect to learning effectiveness. Furthermore, successful deployment of these educational systems in educational settings was not documented.

Although automatic question generation can be achieved using a variety of natural language processing techniques which have gained wide acceptance, there is a lack of strategies for deploying question generation into educational systems. A similar finding has also been identified by Mostow and Chen (2009) especially for the purpose of training reading comprehension: most existing work in this field rather randomly chooses sentences in a text to generate questions than posing questions in an educational strategic manner.

Research on question generation with focus on educational settings needs to develop further. In this paper, we propose three directions for deploying question generation in educational settings. First, in the area of Intelligent Tutoring Systems, several research questions can be investigated, e.g., if the intent of the questions is to facilitate learning, which question taxonomy should be deployed? Given a student model in an Intelligent Tutoring System, which question type is appropriate to pose the next questions to the student? The second research direction is deploying semantic information available on the Internet (e.g., Wikipedia, WordNet) to generate questions. The goal of generating semantics-based questions might include stimulating students' brainstorming, reminding students to additional information, and supporting students solving problems. The third research direction focuses on developing meta-cognitive skills of students. Using questions in teaching is known to be beneficial. Asking students to generate questions helps students recall knowledge and deepen learned content. In addition, it might also develop thinking skills. Deploying automatic question generation in educational systems may use model questions to help students improve the skill of creating questions and thus, meta-cognitive skills of students.

5 Conclusions

In this paper, we have reviewed numerous educational applications of question generation and technical approaches to generating questions. From the technical point of view, many technical approaches for generating questions are successful. Although question generation has a long history, the number of prototypes of question generation for educational purposes is still small.

For the research area in deploying question generation for educational purposes, we propose three research directions. First, question generation should be deployed in Intelligent Tutoring Systems in order to support students in problem solving. The second research direction is deploying semantic information available on the Internet (e.g., Wikipedia, WordNet) to generate semantics-based questions in self-directed or constructivist learning environments. The third research direction promotes applying automatic question generation in order to develop meta-cognitive skills of students, especially the skill of generating questions.

6 References

1. Ali, H., Chali, Y. & Hasan, S. A. (2010). Automation of question generation from sentences. In Boyer, K. E. and Piwek, P. (eds.), Proceedings of the 3rd Workshop on Question Generation, held at ITS 2010, pp.58-67.
2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S.: DBpedia - A crystallization point for the Web of Data (2009). *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), pp. 154-165.
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the SIGMOD International Conference on Management of Data*, pp. 1247-1250, ACM.
4. Brown, J., Frishkoff, G., & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.
5. Chen, W., Aist, G., & Mostow, J. (2009). Generating questions automatically from Informational text. In: Craig, S. D. & Dicheva, D. (eds.), *Proceedings of the 2nd Workshop on Question Generation*, held at AIED 2009, pp.17-24.
6. Chi, M. T. H., Lee, N., Chiu, M. H., & LaVancher, C. (1994): Eliciting Self-Explanations Improves Understanding. *Cognitive Science*, 18(3), pp. 439-477
7. Cohen, F.S. (1929). What is a Question? *The Monist* 39: 350-364
8. Graesser, A. C. & Person, N. K. (1994). Question Asking during Tutoring. *American Educational Research Journal*, 31(1), pp. 104-137.
9. Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: a tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36(2):180-92.
10. Graesser, A. C., Rus, V., D'Mello, S. K., & Jackson, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning*, pp. 95-125, Information Age Publishing.
11. Heath, T. & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers

12. Heilman, M. & Smith, N. A. (2009). Question generation via over-generating transformations and ranking. Report CMU-LTI-09-013, Language Technologies Institute, School of Computer Science, Carnegie Mellon University
13. Jouault, C., & Seta, K. (2013). Building a Semantic Open Learning Space with Adaptive Question Generation Support. In: *Proceedings of the 21st International Conference on Computers in Education*.
14. Kalady, S., Elikkottil, A., & Das, R. (2010). Natural language question generation using syntax and keywords. In Boyer, K. E. and Piwek, P. (eds.), *Proceedings of the 3rd Workshop on Question Generation*, held at ITS 2010, 1-10.
15. Knight, K. & Marcu, D. (2000). Statistics-based summarization – step one: Sentence compression. *Proceedings of the 17th National Conference of the American Association for AI.*
16. Kunichika, H., Katayama, T., Hirashima, T. & Takeuchi, A. (2001): Automated Question Generation Methods for Intelligent English Learning Systems and its Evaluation. *Proceedings of the International Conference on Computers in Education*, pp. 1117-1124.
17. Lane, H. C. & Vanlehn, K. (2005): Teaching the tacit knowledge of programming to novices with natural language tutoring. *Journal Computer Science Education*, 15, pp. 183–201.
18. Liu, M., Calvo, R.A., & Rus, V. (2012). G-Asks: An Intelligent Automatic Question Generation System for Academic Writing Support. In *Dialogue and Discourse* 3 (2), 101-124.
19. Mannem, P., Prasady, R., & Joshi, A. (2010). Question generation from paragraphs at UPenn: QGSTEC system description. In Boyer, K. E. and Piwek, P. (eds.), *Proceedings of the 3rd Workshop on Question Generation*, held at ITS 2010, p.84-91.
20. Mitkov, R., Ha, L. A., & Karamanis, N. (2006) A computer-aided environment for generating multiple-choice test items. Journal Natural Language Engineering 12 (2): 177–194. Cambridge University Press.
21. Mostow, J. & Chen, W. (2009). Generating instruction automatically for the reading strategy of self-questioning. *Proceeding of the Conference on Artificial Intelligence in Education*, pp.465-472.
22. Olney, A.M., Graesser, A., & Person, N.K. (2012) Question Generation from Concept Maps. In *Dialogue and Discourse* 3 (2), 75–99.
23. Pal, S., Mondal, T., Pakray, P, Das, D., & Bandyopadhyay, S. (2010). QGSTEC system description - JUQGG: A rule-based approach. In Boyer, K. E. and Piwek, P. (eds.), *Proceedings of the 3rd Workshop on Question Generation*, held at ITS 2010, pp. 76-79.
24. Person, N. K., & Graesser, A. C. (2002). Human or Computer? AutoTutor in a Bystander Turing Test. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, Stefano A. Cerri, Guy Gouardères, and Fábio Paraguaçu (Eds.). Springer-Verlag, pp. 821-830.
25. Piwek, P. & Boyer, K. E. (2012). Varieties of Question Generation: introduction to this special issue. *Dialogue & Discourse*, 3(2) pp. 1–9.
26. Ratinov, L. & Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. *Proceedings of the 13th Conference on Computational Natural Language Learning.*
27. Rus, V., Cai, Z. & Graesser, A. (2008). Question Generation: Example of A Multi-year Evaluation Campaign. In: Rus, V. and A. Graesser (eds.), *Online Proceedings of 1st Question Generation Workshop*, NSF, Arlington, VA.

28. Sneiders, E. (2002). Automated question answering using question templates that cover the conceptual model of the database. *Proceedings of the 6th Int. Conference on Applications of Natural Language to IS*, pp. 235-239.
29. Tenenberg, J. & Murphy, L. (2005): Knowing What I Know: An Investigation of Undergraduate Knowledge and Self-Knowledge of Data Structures. *Journal Computer Science Education*, 15(4), pp. 297-315.
30. Varga, A. & Le, A. H. (2010). A question generation system for the QGSTEC 2010 Task B. In Boyer, K. E. and Piwek, P. (eds.), *Proceedings of the 3rd Workshop on Question Generation*, held at ITS 2010, pp. 80-83.
31. Wyse, B. & Piwek, P. (2009). Generating questions from OpenLearn study units. In Craig, S. D. & Dicheva, D. (eds.), *Proceedings of the 2nd Workshop on Question Generation*, held at AIED 2009, pp. 66-73.
32. Yu, F.-Y., Liu, Y.-H. & Chan, T.-W. (2005). A Web-Based Learning System for Question-Posing and Peer Assessment. *Innovations in Education and Teaching International*, 42(4) pp. 337-348.
33. Vanlehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C.P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1), pp. 3-62