Tapping into the Power of Automatic Question Generation

Ibrahim Eldesoky Fattoh
Faculty of Information Technology. Misr University for Science &Technology, Egypt

Amal Elsayed Aboutabl
Computer Science Dept., Faculty of Computers& Information, Helwan University, Egypt

Mohamed Hassan Haggag
Computer Science Dept., Faculty of Computers & Information, Helwan University, Egypt

ABSTRACT

Question Generation (QG) is an important element of learning environments, information seeking systems, help systems, and other applications. There are a number of distinct research subfields which are concerned with the Automatic Question Generation (AQG) Systems. This research tries to have a wide look on existing automatic question generation systems, and some trials of overcoming its difficulties from different points of views.

General Terms

Natural Language Processing, Natural Language Generation

Keywords

Automatic Question Generation, Questions Taxonomy, Multiple Choice Questions, Entity Based Questions.

1. INTRODUCTION

Questions have been studied as part of the task of Question Answering in the field of Natural Language Processing (NLP). At the beginning, question answering research focused on answering questions from databases and knowledge representations [1], but in the past two decades has refocused on retrieving answers from text – e.g., in 1999 the evaluation of question-answering systems became part of the Text Retrieval Conference (TREC) series. Simultaneously, there has been a strand of research on advisory dialogue systems [2]. All the previous systems were primarily aimed at responding to the user's questions. Recently, there has been a broader transformation in the field of Natural Language Processing researches in Question Generation task. Since 2008, researchers from different communities, such as, Discourse Analysis, Dialogue Modeling, Formal Semantics, Intelligent Tutoring Systems, Natural Language Generation, Natural Language Understanding, and Psycholinguistics, have met annually at the Question Generation workshop. AQG system would be useful for building an automated trainer for learners to ask better questions, and for building a better hint and question asking facilities in intelligent tutoring systems [3]. Another benefit of QG is that it can be a good tool to help in improving the quality of the Question Answering (QA) systems. Available studies revealed that humans were not very skilled in asking good questions. Therefore, they would benefit from automated QG systems to assist them in meeting their inquiry needs [4]. In this research, a survey about automatic question generation and different variety of work that has been done is discussed. The rest of the paper is organized as follows: section 2 discusses the basic steps for automatic question generation systems; section 3 introduces two distinct taxonomies of questions, section 4 the state of the art in which a description of the previous work, finally section 5 introduces a conclusion with some remarks.

2. STEPS FOR AUTOMATIC QUESTION GENERATION

The automatic question generation problem is derived from Natural Language Processing tasks. Its two major sub tasks are natural language understanding and natural language generation. Question generation is the task of generating reasonable questions from an input; the input can be structured like database or unstructured like text [5]. QG can be divided into deep QG and shallow QG [6]. Deep QG generates deep questions that involve more logical thinking (such as why, why not, what-if, what-if-not and how questions) whereas shallow QG generates shallow questions that focus more on facts (such as who, what, when, where, which, how many/much and yes/no questions).

The two main tasks that have been involved in the process of text to QG are, content selection (the informative sentence selected for question generation) and question formation (transformations on the informative sentence to interrogative one to get the question). Question formation further has the subtasks of:

- Finding suitable question type (what- where- when, etc.)
- Auxiliary and main verb transformations and
- Rearranging the phrases to get the final question as shown in figure 1.

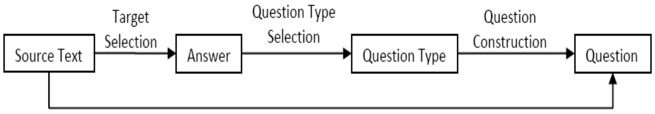

Fig.1. Steps for text to question generation

3. QUESTIONS TAXONOMY

Both [7] and [8] proposed two different taxonomies for question types in their analysis of tutoring transcripts. In [7] the questions were categorized into three categories; simple, intermediate, and complex. In [8], a primary and secondary taxonomy has been done. In the primary taxonomy, the questions were categorized into five categories; description, method, explanation, comparison, and preference questions. After finishing the primary taxonomy, several secondary taxonomic dimensions are supplemented like Collins' question type, and Bloom's taxonomy. Table 1 shows the category types by [7] in the first column and primary category types by [8] in the second column.

Table 1. Categories of Questions Types

Categories by [7]	**Primary Categories By [8]**
Simple/Shallow Questions • Verification: invites a yes or no answer. • Disjunctive: Is X, Y, or Z the case? • Concept completion: Who? What? When? Where? • Example: What is an example of X?	Description Questions • Concept Completion: Who, what, when, where? • Definition: What does X mean? • Feature Specification: What features does X have? • Composition: What is the composition of X? • Example: What is an example of X?
Intermediate Questions • Feature specification: What are the properties of X? • Quantification: How much? How many? • Definition: What does X mean? • Comparison: How is X similar to Y?	Method Questions • Calculation: Compute or calculate X. • Procedural: How do you perform X?
Complex/ Deep Questions • Interpretation: What is the significance of X? • Causal antecedent: Why/how did X occur? • Causal consequence: What next? What if? • Goal orientation: Why did an agent do X? • Instrumental/procedural: How did an agent do X? • Enablement: What enabled X to occur? • Expectation: Why didn't X occur? • Judgmental: What do you think of X?	Explanation Questions • Causal Antecedent: What caused X? • Causal Consequence: What will X cause? • Enablement: What enables the achievement of X? • Rationale Questions • Goal Orientation: What is the goal of X? • Justification: Why is X the case?
	Comparison Questions • Concept Comparison: Compare X to Y? • Judgment: What do you think of X? • Improvement: How could you improve upon X?
	Preference Questions • Free Creation: requires a subjective creation. • Free Option: select from a set of valid options.

The taxonomy by [8] deviates from [7] in the addition of secondary dimensions and the hierarchical structure of the primary taxonomy. They also added question classes and moved some classes to secondary dimensions. These taxonomies could be useful in a number of ways in the question generation task. First, it could be used in the main task to specify the type of question that systems should generate from a text snippet [9]. Second, if the overall question generation task is conceived of as consisting of Concept Selection, Question Type Determination, and Question Construction, then the output of the Type Determination task could be one or more of the labels from the primary taxonomy.

4. STATE OF THE ART

In this section a description of the previous work that have been done in automatic question generation is explored, here a classification of the work is proposed according to the questions formats, multiple choices questions which asks about a word in a given sentence, the word may be an adjective, adverb, vocabulary, etc., the second format is the entity questions systems or Text to Text QG (like factual questions) which asks about a word or phrase corresponding to a particular entity in a given sentence, the question types are like what, who, why etc.

4.1 Multiple Choice Questions Systems (Gap Fill Questions)

The research by [10] presented an automatic question generation system that can generate gap-fill questions for content in a document. Gap-fill questions are fill-in-the-blank questions with multiple choices provided. The system finds the informative sentences from the document and generates gap-fill questions from them by first blanking keys from the sentences and then determining the distractors for these keys. Syntactic and lexical features are used in this process. the authors had introduced some features as a basis for sentence selection like is it a first sentence, common tokens, is it has an abbreviation and superlatives, length and others. In the key selection, the system first generates a list of keys using part of speech tagging (POS) then select the best key from this list depending on three parameters which are; number of occurrences of the key in the document, does it is a word in the title, and height of the key in the syntactic tree. The distractor selection also depends on some features like measuring the contextual similarity between the distractor and the keys in which they present, Dice coefficient score between gap fill sentence and the sentence containing the distractor, and the difference in term frequencies of distractor and keys in the chapter being tested. The system has been tested using two chapters of the biology book and has been evaluated manually by two biology students. The sentence selection module take 0.7 inter evaluator agreement, the key selection take 0.75 inter evaluator agreement, and 0.60 are useful gap fill question which have at least one good distractor.

The authors of [11] introduced a prototype for an automatic quiz generation system for English text to test learner comprehension of text content and English skills. They used the semantic network (SemNet for short), to represent the relationship between a vocabulary and its context. The semantic network constitutes of players, actions, attributes, and relationships between them. They did two generators for two types of questions. The first generator is for sense comprehension of adjectives; the generator will extract attributes (adjectives) of players from the SemNet of a given text as questionnaire vocabularies and form multiple-choice cloze questions by scooping these vocabularies out of the corresponding sentences. The options of a question include the right answer and the distractors. The right answer is substituted by the synonym or a similar adjective of the applied sense of the questionnaire adjective. The candidates of the target substitute are acquired from WordNet. And then, most of them will be filtered out by two filtering steps. In the first step, they check whether a candidate can be an attribute of the corresponding player in the knowledge base. In the second step, they check whether the attributive relationship of the remaining candidate is raised frequently through similar texts. The second generator is for anaphor comprehension, they told that to understand the meaning of a text; a learner must integrate these subnets by connecting each anaphor with its antecedents. Thus, they designed a question generator to examine whether a learner understand the association between an anaphor and its antecedent. The generator identifies the antecedent of an anaphor and form a multiple-choice cloze question by scooping the anaphor out of its sentence. The options comprise its antecedent and the distractors.

The next research proposed an automatic question generation module based on clustering and classification [12]. Through the process of development, the words of source text are clustered into partitions based on the algorithm parameters, to gain all the information for selection of the best suitable sentences for question generation. The word to be distracted from the sentence for a question generation is determined by the classification method using neural network. The system consists of many sub modules like text preprocessing (annotation and stemming), domain preprocessing (topic based clustering of words and generation of term hierarchy), selection of key words, selection of candidate words using the word clusters, grammatical alignment of the candidate words, generation of the test, and evaluation module. A semi-automated selection method is used for determining the base sentences of the questions. A CPN (Counter Propagation Network) variant was invoked in the selection of candidate words for the question. To determine the word class and stem of the words, a simplified free version of the Szószablya framework is applied. Their question generation system is capable to generate three types of sentences (concept, definition, declarative sentence) from which each of the three types appeared in the prepared test paper in the form of multi-choice questions. The system was tested and evaluated in the second term of academic year 2011-2012. 45 people took part in testing from who 40 were students and 5 were experts. The system was compared with manually generated tests. The performed analysis showed that the automatically generated questions are as good as the manually constructed tests and could be used in future e-learning frameworks.

Another research that generates multiple choice questions for understanding evaluation of adjectives in a text is found in [13]. Based on the sense association among adjectives, an adjective being examined can be usually substituted by some other adjectives. The system was able to generate three types of questions: questions for collocations, questions for antonyms, and questions for synonyms or similar words. The basic idea of the system is as follow; for a given sentence, extract an adjective-noun pairs that exist, then for each adjective-noun pair, if it is a collocation, generate a question for it. If the original sentence has words which have negative meanings, generate a question for antonyms. Also generate question for synonyms or similar words. The candidates of a substitute are gathered from WordNet and filtered by web corpus searching. To evaluate the generated questions, they choose Far East senior high school English textbook Book One, which contains 12 articles, as the experiment material. Experiment results have shown that the proposed answer determination approaches and question filtering strategies are effective in precision.

The authors of [14] proposed **ArikIturri** system which is an Automatic Question Generator for Basque language test questions, which is independent from the test assessment application that uses it. The information source for this question generator consists of linguistically analyzed real corpora, represented in XML markup language. The system is an NLP based which is able to generate four different types of questions: Fill in the blank (FBQ), word formation, multiple choice, and error correction. Two kinds of language resources were used: NLP tools and specific linguistic information for question generation. The life cycle of the system as follows, the sentence retriever module selects candidate sentences from the source tagged corpus. In a first step, it selects the sentences where the specified linguistic phenomena appear. Then, the candidates selector studies the percentages of the candidates in order to make random selections of sentences. Once the sentences are selected, the answer focuses identification marks out some of the chunked phrases as answer focuses depending on the morpho syntactic information of the phrases. Then, the item generator creates the questions depending on the specified exercise type. It is probable that some questions are ill-formed. Because of that, the system included the ill-formed questions rejecter in the architecture. The results of the evaluation of ArikIturri with multiple choice and error correction question types were presented. These results demonstrate that the automatic generator is good. In fact, the well-formed questions are more than %80.The experiments carried out during the implementation of the system have proved that the source corpus and the NLP techniques used in the process of question generation determine the quality of the obtained questions.

Authors in [15] proposed an approach for AQG for vocabulary assessment; they generated 6 types of questions: definition, synonym, antonym, hypernym, hyponym, and cloze questions. They retrieve the data from WordNet after choosing the correct sense for it. Each of the 6 types of questions can be generated in several forms, the primary ones being word bank and multiple-choice. In word bank, the testee sees a list of answer choices, followed by a set of questions or statements. Concerning the distractor choice, the question generation system chooses distractors of the same part of speech and similar frequency to the correct answer, as recommended by Coniam (1997). In the assessment of the questions they focused on the automatically generated multiple-choice questions, with distracters based on frequency and POS. Four of the six computer-generated question types were assessed: the definition, synonym, antonym, and cloze questions. The percentage of questions generated for the four types were above 60% for 156 word list.

4.2 Entity Based Questions Systems

The authors of [16] presented a system that automatically generates questions from natural language text using discourse connectives. They explored the usefulness of discourse connectives for QG by analyzing the senses of the connectives that help in QG and proposed a system that uses this analysis to generate questions of the type why, when, give an example and yes/no. They analyzed four subordinating conjunctions, since, when, because and although, and three adverbials, for example, for instance and as a result. The system goes through the entire document and identifies the sentences containing at least one of the seven discourse connectives. Then the system finds the question type on the basis of discourse relation shown by discourse connective.

Table 2. Discourse connectives and their question type used in [16]

Discourse Connectives	Sense	Question type
Because	Causal	Why
Since	Temporal causal	When- why
When	Causal + temporal conditional	When
Although	Contrast concession	Yes / no
As a result	Result	Why
For example	Instantiation	Give an example – where
For instance	Instantiation	Give an instance – where

Table 2 shows the 7 discourse connectives and their senses and the question types for each one as they suggested. The task of content selection involves finding the target argument of the discourse connective. The target argument is the part to be asked for. It may be a sentence or a clause before the discourse connective or after it. Identification of target argument is done in two steps. The system first locates the syntactic head or head verb of the target argument and then extracts it from the dependency tree of the sentence. After that set of transformations are applied on the content to get the final question. The system was evaluated manually. The evaluation was performed by two graduate students with good English proficiency. Evaluators were asked to rate the questions on syntactic and semantic correctness. The system was tested on two datasets, QGSTEC-2010 development dataset and the overall system is rated 6.3 out of 8 on this dataset, the total number of questions generated for this dataset is 61. The second data set is five Wikipedia articles (football, cricket, basketball, badminton and tennis) and the overall rating of the system is 5.8 out of 8 and the total number of questions is 150 for this dataset. By carrying out error analysis on the system there were 4 error types found that affect the overall rating of the system. The errors were co-reference resolution, parsing, inter-sentential connectives, and non-handling of predicative adjuncts.

An automatic Sentence-to-Question generation system presented in [17], where given a sentence, the system generates a set of questions for which the sentence contains, implies, or needs answers. A syntactic parser was used to build elementary sentences from the input complex sentences. A named entity recognizer and a part of speech tagger are applied on each of these sentences to encode necessary information. The classification of sentences based on their subject, verb, object and preposition for determining the possible type of questions to be generated. In their research they considered the factoid questions only like what, where, when , who, and so on. They divided the system into 3 basic modules, data preprocessing, elementary sentence construction, and sentence classification and question generation. In the data preprocessing the Oak system used to tokenize the sentences, also the named entity tagger to extract entities from the sentences like a person name, dates, places etc., and a part of speech tagger used to provide information about verbs and their tenses. In the second module the Charniak parser used to construct the syntactic tree for the statements to get the elementary sentence. The subject, object, preposition and verb for each elementary sentence used to classify the sentences in the third module. For example, if a sentence has the structure: “Human Verb Human”, it will be classified as “whom and who” question types. That system was tested on TREC dataset, the Recall and Precision measures used to evaluate the results. For the “When”, “Where” and “Who” type questions, the Recall was similar. The type “What”, got the lower Recall. The precision was high for the types “Who and Where“, the type “When “was still above 0.5, the other types were hovering around 0.4.

The authors of [18] proposed a framework for question generation composes general purpose rules to transform declarative sentences into questions using NLP tools and a statistical component for scoring questions. The input sentence passed through three stages in that framework, transforming the source sentence, question transducer, and question ranker. In the first stage, the selected sentence or a set of sentences from the text is transformed into one declarative sentence by optionally altering or transforming lexical items, syntactic structure, and semantics. Many existing NLP transformations exploited in this stage, including extractive summarization, sentence compression, sentence splitting, paraphrasing, textual entailment, lexical semantics for word substitution. In the second stage, the question transducer takes as input a declarative sentence and produces as output a set of possible questions. It identifies the answer phrases which are the targets of WH-movement and converts them into question phrases. In the system, answer phrases can be noun phrases or prepositional phrases, which enables who, what, where, when and how much questions. Tregex, a tree query language used in order to implement the rules for transforming source sentences into questions, and Tsurgeon, a tree manipulation language built on top of Tregex. A set of Tregex expressions marks the phrases in an input tree which cannot be answer phrases due to constraints on WH-movement. After marking unmovable phrases, the transducer iterates over the possible answer phrases. For each one, it copies the input tree, then removes the answer phrase and generates possible question phrases from it. The declarative sentence is turned into a question by executing a set of defined syntactic transformations (WH-movement, subject-auxiliary inversion, etc.). In the third stage, the questions are scored and ranked using discriminative ranker based on the logistic regression model. The questions are ranked according to features of the source sentences, input sentences, the question, and the transformations used in generation. The training and data sets used are from different corpora like WIKI_ENG, WIKI_SIMP, and Section 23 of the Wall Street Journal data in the Penn Treebank. The manual evaluation shows that the system achieves 43.3% precision-at-10, generating approximately 6.8 acceptable questions per 250 words of source text.

5. CONCLUSION REMARKS

This research discussed a set of researches that generated AQG systems, the systems classified according to the question formats into two types. The first type (multiple choice) accepts a declarative input in Natural Language or a Knowledge Representation formalism (concept maps), and output consisting of interrogative sentences. The algorithms for solving these problems vary, in a number of ways, the basic differences are found in the methods used to find the key to be asked and finding distrctors for that key, finding the key depends on the objective of the question (ask for vocabulary, adjective etc.,), the methods for finding the detractors are different, some uses a set of words form WordNet tool, some uses other adjectives from bank of words their system has been built. Others used the POS and word senses to get distractors. According to evaluation of those systems, all of them are manually evaluated by students or experts. The percentage of evaluation ranges from 60% to 80% for these systems. . The second type (Entity based) is a Text-to-Text QG systems, the input is a declarative sentence, a preprocessing step is involved, which divides the complex sentence up into sentences of a size that is appropriate for generating questions. Most systems carry out syntactic processing and some semantic processing to arrive at an intermediate representation. The amount of semantic processing exists in recognition of named entities only. The approaches share the use of transformation rules. Also in this type the objective of the questions specified the algorithms used, one used the senses of the words to be asked to specify the question word, the other depended on the Named entity tagger to specify the question word. For more improvements in those systems, the inclusion of more semantic information can be added and concentration on a classification modules of the sentences.

6. REFERENCES

[1] Bronnenberg, W.J., H.C. Bunt, J. Landsbergen, P. Medema, R. Scha, W.J Schoenmakers And E. Van Utteren (1979). The Question-Answering System Phliqa 1. In: L. Bolc (Ed.), Natural Communication With Computers. Macmillan.

[2] Winograd, T. (1972) Understanding Natural Language. Academic Press, New York.

[3] Graesser A. C., Vanlehn K., Rose C. P., Jordan P. W. & Harter D. (2001). Intelligent Tutoring Systems With Conversational Dialogue. Ai Magazine, 22(4), 39–52.

[4] Rus V. & Graesser A. C. (2009). The Question Generation Shared Task And Evaluation Challenge. In Workshop On The Question Generation Shared Task And Evaluation Challenge, Final Report, The University Of Memphis : National Science Foundation.

[5] Rus, V., Cai, Z., And 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, September 25-26, 2008, Nsf, Arlington, Va.

[6] Graesser, A. C., Ozuru, Y., & Sullins, J. (2009). What Is A Good Question. In M. Mckeown (Eds), Festscrift For Isabel Beck. Mahwah, Nj: Erlbaum.

[7] Graesser, A. C., & Person, N. K. (1994). Question Asking During Tutoring. American Educational Research Journal, 31 , 104-137.

[8] Nielsen ,R., Buckingham ,J., Knoll ,G., Marsh , B., Palen ,L. (2008). A Taxonomy Of Questions For Question Generation. In Workshop On The Question Generation Shared Task And Evaluation Challenge.

[9] Nielsen, R. (2008). Question Generation: Proposed Challenge Tasks And Their Evaluation. In Workshop On The Question Generation Shared Task And Evaluation Challenge, Arlington, Va.

[10] Agarwal , M. And Mannem ,P. (2011). Automatic Gap-Fill Question Generation From Text Books. In Proceedings Of The 6th Workshop On Innovative Use Of Nlp For Building Educational Applications. Portland, Or, Usa. Pages 56-64.

[11] Sung, L., Lin, Y., And Chern, M.(2007). An Automatic Quiz Generation System For English Text. Seventh Ieee International Conference On Advanced Learning Technologies (Icalt 2007).

[12] Bednarik, L., And Kovács, L. (2012). Implementation And Assessment Of The Automatic Question Generation Module. Coginfocom 2012 • 3rd Ieee International Conference On Cognitive Info-communications • December 2-5, 2012, Kosice, Slovakia.

[13] Lin, Y., Sung, L., And Chern, M (2007). An Automatic Multiple-Choice Question Generation Scheme For English Adjective Understanding. Workshop On Modeling, Management And Generation Of Problems/Questions In Elearning, The 15th International Conference On Computers In Education (Icce 2007), Pages 137-142, Hiroshima, Japan.

[14] Aldabe, I., Lopez De Lacalle, M., Maritxalar, M., Martinez, E., And Uria, L. (2006). Arikiturri: An Automatic Question Generator Based On Corpora And Nlp Techniques, Ser. Lecture Notes In Computer Science, Vol. 4053, Pp. 584–594. Springer, Heidelberg.

[15] Brown, J., Firshkoff, G. And Eskenazi, M. (2005) Automatic Question Generation For Vocabulary Assessment. Proceedings Of Hlt/Emnlp-2005, 819–826. Vancuver, Canada.

[16] Agarwal, M., Shah, R., And Mannem, P. (2011).Automatic Questions Generation Using Discourse Cues. In The Proceedings Of The 6th Workshop On Innovative Use Of Nlp For Building Educational Applications, Aclhlt

[17] Ali, H., Chali, Y., And Hasan, S. (2010). Automatic Question Generation From Sentences. In Proceedings Of Qg2010: The Third Workshop On Question Generation, 2010.

[18] Heilman M., and Smith N.A. "Question generation via overgenerating transformations and ranking". Technical Report, DTIC Document, 2009.

7. AUTHOR’S PROFILE

Ibrahim Eldesoky Fattoh is currently working as a teacher assistant, Computer science department, Faculty of Information Technology, Misr University for Science and Technology (MUST), Cairo, Egypt. He is a Ph.D student, Faculty of Computer and Information, Helwan University, Egypt in the area of Automatic Question Generation under supervision of Prof. Mohammad Hassan Haggag and Assoc. Prof. Amal Elsayed Aboutabl. He finished his B. Sc. and Master degrees at Faculty of Computers and Information, Helwan University, Egypt. His master thesis was entitled "Supervised Immune System for Information Filtering". His research interests include Artificial Intelligence, Natural Language Processing, Data Mining, Text Mining, and Computational Intelligence.

Amal Elsayed Aboutabl is currently an Associate Professor at the Computer Science Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt. She received her B.Sc. in Computer Science from the American University in Cairo and both of her M.Sc. and Ph.D. in Computer Science from Cairo University. She worked for IBM and ICL in Egypt for seven years. She was also a Fulbright Scholar at the Department of Computer Science, University of Virginia, USA in 2009. Her current research interests include parallel computing, performance evaluation and text processing.