# Automatic question generation approaches and evaluation techniques

**Manisha Divate* and Ambuja Salgaonkar**

Department of Computer Science, University of Mumbai, Santacruz, Mumbai 400 098, India

**The objective of this article is to review several automatic question generation systems and find why automatic question generation is still an attraction for researchers. The focus is mainly on the task of question generation, analysis of the approaches and evaluation of various methods of automatic question generation. Pointers for further research are included.**

**Keywords:** Automatic question generation, evaluation techniques, quality enhancers, ranking, sentence simplification.

A study reveals that an average student asks over 26 questions per hour in one-on-one human tutoring sessions; in contrast, the student poses 120 questions per hour in a learning environment that forces her to ask questions in order to access any information[1]. Conversely, students learn more deeply if prompted by questions[2].

Conventionally, questions are constructed and assessed by tutors. It has been a trend for several decades that automatic question generation (AQG) system generates questions from the corpora using natural language processing.

AQG systems were first developed in the 1976 (ref. 3). They have been created for English language and vocabulary, medicine, education and using multimedia. The sequence of developments is as follows: learning words[4–6], English[7], grammar testing[8], medicine[9], academic writing[10], literature review[11], education[12] (henceforth, Heilman and Smith AQG is abbreviated as HSAQG), multimedia[13], and finally a recent major development, on-line learning[14]. This article presents a review of more than 50 contributions in the domain of AQG.

## Types of questions

In classroom practice, a tutor evaluates the comprehension of a learner by asking gap-fill type questions (GFQs), multiple choice questions (MCQs), factoid-based questions (FBQs) and deep learning-type questions (DLQs).

### Gap-fill questions

A stem is a good question or problem to be solved[15]. To identify a stem and generate a GFQ, an informative sentence is selected from a given document. The selection of information involves identification of semantic features in the entire document.

Next, a key phrase or answer phrase (assume it is a noun phrase) is selected; term frequency plays an important role. A distractor (not expected to occur in the question) is a choice given to a learner. A good distractor could be a synonym of the key phrase or an important term in the domain of the key phrase. Distractors in Revup, an AQG, are selected from word2vec, a vector of words[16].

Text summarization features like length of a sentence, number of common tokens, number of noun and pronouns, and position of a sentence are generally considered[17].

### Multiple choice questions

MCQ is a wh-type question that comes with a set of multiple distractors and a single correct answer[18]. Framing an MCQ is a three-step process:

S1. Term extraction,

S2. Selection of distractors,

S3. Question generation.

In term extraction, noun and noun phrases are selected using a shallow parser tool. A key phrase is a term whose frequency is above a threshold.

A question is generated by following a question template. Sentences in subject–verb–object or subject–verb form are good candidates for an MCQ.

Base sentence: 'The verb is the most central element in a clause'.

Question template: 'Which HVO', where H is a hypernym of the underlined term in the sentence, and phrase 'part of speech' is a hypernym of a term 'The verb'.

The heuristic while framing a question is, replace the term 'The verb' by 'part of speech'. Thus:
'Which part of speech is the most central element in a clause?'[19].

### Factoid-based questions

Questions beginning with a wh-phrase, i.e. what, who, which, why, when, where, etc. are factoid questions. How is a wh-phrase by default. (These words are capitalized

*For correspondence. (e-mail: divate.manisha.79@gmail.com) here when they are used in a technical, rather than generic, sense.) Answering FBQs helps a learner to understand the subject in depth. The formulation of an FBQ primarily needs syntactic information[20–22]. Three types of FBQs can be generated from informational text[23]:

(1) Questions about conditional context are the ones that have stems containing phrases like if, then, so, etc.

Example 1:

Sentence: If it rains, the picnic will be cancelled.

Question: What would happen if it rains?

(2) Questions about temporal information are the ones whose stems contain date and time expressions.

Example 2:

Sentence: Tilak obtained Bachelor of Arts in first class in Mathematics from Deccan College of Pune in 1877.

Question: When did Tilak obtain Bachelor of Arts in first class in Mathematics from Deccan College of Pune?

(3) Stems of questions on possibility and necessity contain in their expressions the words would, will, should, could, must, may.

Example 3:

Sentence: You should read daily two hours for your exam.

Question: Why should you read daily?

It is observed that factoid questions starting with who and whom are easy to generate because (1) they are based on a subject, the entity person, and (2) while framing a question, sentence sequence remains the same. Questions with Where-phrase are easy to predict and generate because they based on the entity location[24].

### *Deep learning questions*

Deeper learning is the process by which a learner becomes capable of applying her learning to new situations. The contribution of DLQs is in developing critical thinking within a learner that, independent of the teacher's efforts, enhances the learner's understanding of the subject[25]. Unlike FBQs, a DLQ often starts with a question stem such as why, what caused, how did it occur, what if, how does it compare, or what is the evidence[26].

A tool for pedagogically generating DLQs that aid students in essay writing has been discussed earlier[27]. An AQG for generating questions to drive group discussions has been provided earlier[28]. A base sentence is selected using text categorization to generate FBQs which is further information retrieval and text summarization form a basis for generating subjective-type DLQs.

Example 4:

FBQ: What was inscribed on the side of the barn?

Prefix: Discuss in detail.

DLQ: Discuss in detail what was inscribed on the side of the barn.

Example 5:

FBQ: Does psychological manipulation unite the animals against a supposed enemy?

Prefix: Why?

DLQ: Why does psychological manipulation unite the animals against a supposed enemy?

The choice of prefix determines the nature of deep learning.

After taking note of student citations, the G-Asks[11] an AQG system, triggers questions that ask for evidence regarding opinion, result, system, application, method and aim.

Example 6:

Answer based on student's opinion of the source sentence:

Source sentence: 'Cannon (1927) challenged this view, mentioning that physiological changes were not sufficient to discriminate emotions.'

DLQ*:* Why did Cannon challenge this view mentioning that physiological changes were not sufficient to discriminate emotions?

## The AQG process

The process of manual question generation has the following three steps: (i) reading the text, (ii) finding an important idea or answer phrase and (iii) transforming the idea into a question. In comparison, an AQG executes these four modules: (i) sentence simplification, (ii) answer phrase selection, (iii) sentence transformation, and (iv) question ranking and evaluation[12,14,21].

### *Sentence simplification*

Sentence simplification includes splitting of sentences with independent clauses, appositive phrases, prepositional phrases, discourse marker, and relative clauses. Table 1 shows the various clauses considered for sentence simplification. While simplification makes some aspects of question generation easier, it also introduces new problems that must be handled.

Example 7: 'The boy went to school on Monday, and he came home on Thursday.'

Case 1: Independent clauses are separated.

The boy went to school on Monday.

He came home on Thursday.

Case 2: Resolve anaphora-pronoun reference in particular.

The boy came home on Thursday.

(Replace the pronoun 'he' by the noun phrase 'the boy'.)

Case 3: Identify insignificant prepositional phrases (PPs) and remove them.

Example 8: Because of this, the sal tree is revered by many Buddhist people around the world.

**Table 1.** Clauses considered while simplifying a sentence

| Clauses | Definition | Example | Phrases | Sentence simplification |
|---|---|---|---|---|
| Independent clause | An independent clause is one that can stand by itself as a simple sentence. | I love Mexican food but my stomach dislikes it.<br>Juliet loves her little brother; he is generous and kind. | Nil | I love Mexican food. My stomach dislikes it. |
| Appositive phrase | Appositive is a noun or pronoun placed beside another noun or pronoun to describe or identify it. | Most people have never played polo, a very expensive game. | Nil | Most people have never played polo. |
| Leading prepositional phrase | A phrase with preposition and its object. | On the far side of the camping ground, they saw the lion slowly walking away from them into the woods. | Above, against, at around, before, behind, below, besides, between, by, for, in, of, off, on, over, through, to, under, with | They saw the lion slowly walking away from them into the woods. |
| Relative clauses | A clause which gives extra information about the noun. | I bought a new house which is very big.<br>Children who hate chocolate are uncommon. | When, where, who, which, that | I bought a new house. |
| Discourse marker | Discourse markers are essentially linking words. They show how one piece of conversation is connected to another. | She did not win the contest; however, she managed to deliver a satisfactory performance. | However, nevertheless, so, well, anyways | She did not win the contest. She managed to deliver a satisfactory performance. |
| Noun participle | A participle phrase begins with a present or past participle. | While waiting for take-off, the flight attendants passed out magazines. | Present participle phrase ends with -ing and past participle phrase end with -ed | The flight attendants passed out magazines. |

After removing PPs: The sal tree is revered by many Buddhist people around the world[29].

Note that not all the PPs are insignificant.

Example 9: During EI Niño, warm water moves eastward instead.

Removal of PPs results in the loss of important temporal information, and the result is: warm water moves eastward[14].

Sometimes, sentence simplification depends upon two linguistic phenomena: (1) semantic entailment and (2) presupposition[29].

(Def: *A* semantically entails *B* if and only if whenever *A* is true, *B* is also true.)

Consider the following two cases:

Case 1: Removal of discourse markers and adjunct modifiers in *A* leads to semantically entailed *B*.

Def: A discourse marker is a word or phrase that does not change the truth – conditional meaning of the sentence.

Def: An adjunct modifier is an optional part of a sentence, clause or phrase that, if removed or discarded, will not otherwise affect the remainder of the sentence.

Example 10:

*A*: However, Jefferson did not believe the Embargo Act, which restricted trade with Europe, would hurt the American economy.

Discourse marker: However.

Adjunct modifier: Which restricted trade with Europe.

*B*: Jefferson did not believe the Embargo Act would hurt the American economy.

Case 2: Remove the conjunctions to separate clauses and verb phrases.

Example 11:

*A*: John studied on Monday but went to the park on Tuesday.

Conjunction: But.

*B*: John studied on Monday. John went to the park on Tuesday.

Sentences *A* and *B* are semantically entailed.

Def: Presupposition is an implicit assumption about the world or background belief relating to an utterance whose truth is taken for granted in discourse (Wikipedia).

*A* presupposes *B* if and only if *B* is true independent of *A*.

Example 12:

*A*: Jane no longer writes fiction.

*B*: Jane once wrote fiction.

Simplification is by forgetting *A* and continuing only with *B*.

## *Selection of an answer phrase*

Given a simplified sentence, an AQG that generates FBQs or MCQs identifies a noun and PP as an answer phrase[30]. Another AQG employs semantic role information and name entity recognition to identify an answer phrase[31].

**Table 2.** Sample templates for question generation

| Base sentence | Question | Template type | Question template |
|---|---|---|---|
| After the foliage period is completed, bulbs may be dug up for replanting elsewhere. | When would bulbs be dug up for replanting elsewhere? | Temporal | When- would-<*X*>? |
| If all medicines in the world were thrown into the sea, it would be all the better for mankind and all the worse for the fishes. | What would happen if all medicines in the world were thrown into the sea? | Conditional | What would happen if <*X*>? |
| Any surface moisture should be dried, then the bulbs may be stored up to about 4 months for a fall planting. | Why should any surface moisture dried? | Modality | Why <Aux_verb> <*X*>? |
| Physicians began to think of the pill as an excellent means of birth control for young women. | Why did physicians begin to think of the pill as an excellent means of birth control for young women?<br>What evidence is provided by physicians to prove the opinion?<br>Do any other scholars agree or disagree with physicians? | Opinion | Why + subject_auxiliary_inversion?<br>What evidence is provided by +subject+ to prove the opinion?<br>Do any other scholars agree or disagree with +subject+? |

Example 13:

Sentence: Tilak was born in 1879 in Ratnagiri.

Answer phrases: Tilak (NP), in 1879 (PP), in Ratnagiri (PP).

Sentence: A wind coming from the south is given as 180 degrees.

Answer phrase marked using SRL (semantic role labeler): As 180 degrees [AM-MNR].

### *Sentence transformation/question generation*

Here the AQG takes a simple declarative sentence and an answer phrase as input and produces a set of possible wh-questions as output. Wh-questions are generated using the following three approaches: (1) Template-based; (2) syntax-based and (3) semantic-based.

*Template-based approach:* Good quality templates for question generation can be generated through human intervention, and based on an answer phrase appropriate templates are selected. Since it is difficult to build templates for generic topics, this approach is recommended for special-purpose applications.

Templates for generating questions based on a conditional context in a given text involve phrases like what-would-happen-if, when-would-*X*-happen, what-would-happen-when and why-*X*. Here *X* is semantic roles tagged by SRL tool. Templates involving the word ‘when’ are useful for generating temporal questions and those involving ‘why’ are useful to generate linguistic modality questions[23]. An AQG uses an appropriate question template for an individual after classifying their citations into categories like opinion, result, system, aim, method and application[11].

Table 2 provides examples of template-based questions. Temporal, conditional and modality phrases are underlined in the base sentences in the table.

*Syntax-based approach:* In this approach questions are generated by manipulating the syntax tree of a sentence into an interrogative. Question generation consists of pipelined operations like marking unmovable phrases, question phrase insertion, decomposition of main verb, subject auxiliary inversion and question generation[3,12].

Yes–no type questions are formulated by simply performing subject-auxiliary verb inversion (i.e. placing auxiliary verb in front of subject of a sentence), and FBQs are formulated on subject noun phrases (NPs), object NPs, and likewise, on appositive, participle and adverbial phrases[32].

Example 14:

Sentence: Dhoni plays cricket.

Answer phrases: Dhoni, cricket.

Mark unmovable phrase: Dhoni plays cricket.

Decompose main verb: Plays → does + play.

Dhoni does play cricket.

Subject – auxiliary inversion: Does Dhoni play cricket.

If Yes–no-type question: Does Dhoni play cricket?

If answer phrase is non-subject, insert wh-word: What does Dhoni play?

FBQ: What does Dhoni play?

*Semantic approach:* The semantic approach is used to define more logical and deep learning questions. The semantic role label indentifies semantic arguments associated with the verb of a sentence and their specific roles. The semantic role labeller identifies mandatory arguments A0 (subject), A1 (object), A2 (indirect object) and optional arguments like AM-TMP, AM-LOC and AM_MNR, which are useful in the construction of questions.

Table 3 shows the output generated by SRL tool developed by Illinois University[33].

Question constructed from Table 3:

Sentence: Yesterday teacher taught us English in class.

S1: Replace semantic roles [A0] – teacher by who and [A2] – us by students.

Question 1: Who taught students English in class yesterday?

S2: Replace AM-LOC by ‘Where’ clause.

Question 2: Where did teacher teach student English yesterday?

Lindberg’s[2] AQG generates the questions based on the rules which are developed by extracting semantic role patterns from the base sentence. The Name Entity Tagger tool and ASSERT, a semantic role labeller (SRL) tool, have been employed for extracting the predicate argument structure of the sentence[31]. The JUQGG system uses the Swirl SRL tool to identify the semantic roles in an input sentence and its dependency relations to formulate questions[32].

The NLPWin tool used in Microsoft applications like spell checking, grammar checking, search and machine translation (MT) contains a component named ‘logical form’ which identifies the semantic relationships of the arguments within a sentence and generates wh-questions[34].

## Ranking

Every AQG system has different criteria for evaluating the quality of questions. In turn, this is dependent on the learner model in a tutoring system. Checking the correctness of grammar, implication of negation, etc. within a question is a cumbersome task. A ranker takes unranked questions and metadata (or features like verb tense, subject–auxiliary verb inversion, pronoun) that describe a method of generating questions from an input sentence.

Given a sufficient number of instances of a feature set, a ranking model predicts the rank of a question according to the acceptability viewpoint. The AQG of Liu *et al.*[35] predicts a question rank using 11 features, while the AQG of Heilman and Smith[29] considers 187 features. Logistic regression-based ranking models seem popular.

Table 4 lists the generally used features.

In a complex and long sentence, questions are generated from the main as well as subordinate clauses. In a two-step ranking process, first the AQG ranks a question based on the depth of a predicate. A question generated from the main clause will get higher rank than the one generated from the subordinate clause. In the second step, questions with more pronouns are given lower rank. This ranking approach ignores grammatical and information content aspects of a question[36].

**Table 3.** Semantic roles assigned by Illinois SRL

| Sentence | Semantic roles |
|---|---|
| Yesterday | teacher [A0] |
| teacher | |
| taught | V: teach.01 |
| us | student(s) [A2] |
| English | |
| in | location [AM-LOC] |
| class | |
| – | |

Depending upon the length of an answer phrase the AQG generates three types of questions: medium (one-phrase answer), specific (one-word answer) and general (one-paragraph answer). Medium questions are generated from a sentence containing semantic roles (assigned by ASSERT SRL, like ARGM-CAU, ARGM-PNC, ARGM-DIS) and the scope of the answer is beyond a single word. For the generation of specific questions, sentences with semantic roles ARGM-TMP, ARGM-LOC are considered, and the span of argument is the answer scope. General questions are generated on the first sentence of a paragraph and their answer scope is the entire paragraph.

To rank such questions topic scoring is considered as the first element in question ranking. A sentence having a good topic score possesses good information content and hence is a good candidate for question generation. The second element in ranking is language model probability. Simple bigram models with Laplace smoothing are used to generate sentence probability[37].

AQG uses topic relevance and syntactic correctness for ranking a question. Subtopics in a given text are identified using the latent Dirichlet allocation (LDA) method. After identifying subtopics, the extended string subsequence kernel (ESSK) method is used to calculate their similarity with generated questions. For syntactic correctness, the tree kernel function is used. The sentence and questions are parsed into the syntactic tree using the Charniak parser and then similarity between two syntactic trees is found. Duh[38] observed that ranking using Regression SVM and Rank SVM models gave similar results under the same feature set. However, Rank SVM gave significant improvements when intra-set features were incorporated.

A 16-dimentional feature vector has been defined to represent a question and a multiple linear regression model has been evolved for ranking the questions on a five-point scale[39]. Fifty-five ranked questions have been classified using J48 classifier. Table 5 shows the accuracy of this model.

## Evaluation

The evaluation of any AQG is carried out on the basis of multiple criteria, namely, user satisfiability, linguistic well-foundedness, maintainability, cost efficiency, output quality and variability[40].

Table 6 shows the evaluation techniques employed in different AQG models.

**Table 4.** Features set

| Feature | Attributes |
|---|---|
| Length | Length of the answer phrase, source sentence and question |
| Grammatical correctness | Proper nouns, pronouns, adjectives, adverbs, conjunctions, noun phrases, prepositional phrases, subordinate clauses, tense of the main verb |
| Transformable features | Removal of appositives and parentheses, subject is the answer phrase |
| Semantic feature | Noun, verb and preposition |
| Negation | Not, nor, never |
| Vagueness | Vague noun phrase in the base sentence, question and answer phrase |

**Table 5.** Precision, recall and *F*-score values on a five-point scale

| Class | Precision | Recall | *F*-score |
|---|---|---|---|
| *A* (4.1 to 5) | 0.926 | 0.892 | 0.908 |
| *B* (3.1 to 4) | 0.65 | 0.710 | 0.675 |
| *C* (2.1 to 3) | 0.706 | 0.780 | 0.730 |
| *D* (1.1 to 2) | 0 | 0 | 0 |
| *E* (0 to 1) | 0 | 0 | 0 |

**Table 6.** Automatic question generation evaluation techniques

| Evaluation techniques | Role |
|---|---|
| Intrinsic evaluation[42] | The functionality of a system is evaluated against the gold standard result. |
| Extrinsic evaluation[42] | System output is assessed with respect to its impact on a task external to the system itself, e.g. this evaluation technique measures user's learning gain, efficiency in terms of time or effectiveness. |
| Black-box evaluation | Performance of an AQG is measured against a known sample dataset, with respect to parameters such as speed, reliability, resource consumption and accuracy in annotation. |
| Glass-box evaluation | The design of the system (i.e. the algorithms and linguistic resources used) is tested. |
| Automatic evaluation | This technique mimics the behaviour of human assessors while evaluating an AQG by comparing its output with the gold standard. |
| Manual evaluation | Human judges are employed to evaluate system performance. It is simple and available relatively easily. Subjectivity, slow speed and expensive human resources are some of the issues. |
| Formative evaluation[43] | The primary purpose is to inform the designer as to whether progress is being made towards the intended goals. |
| Summative evaluation[43] | Intended to assess whether the defined goals have been achieved by the final version of the AQG. |
| The Bystander Turing test (BTT)[14] | Intended to check if a human can differentiate the questions generated by an AQG from those generated by human. Likert scale is provided for ranking of the questions |

**Table 7.** AQG evaluation based on precision, recall and *F*-score

| AQG | Precision | | Recall | | *F*-score | |
|---|---|---|---|---|---|---|
| Liu *et al.*[11] | 0.73 | | 0.71 | | 0.7 | |
| Ali *et al.*[21] | 0.587 | | 0.276 | | 0.38 | |
| Le and Pinkwart[45] | 0.796 | | 0.276 | | 0.41 | |
| Lindberg[2] | Background knowledge | | Background knowledge | | Background knowledge | |
| | Yes | No | Yes | No | Yes | No |
| | 0.47 | 0.79 | 0.22 | 0.92 | 0.3 | 0.85 |
| Liu *et al.*[46] | Not available | | Not available | | 0.79 | |

**Table 8.** AQG evaluation using Cohen's kappa measure

| AQG | Cohen's kappa |
|---|---|
| Agarwal and Mannem[17] | 0.7 |
| Zhao *et al.*[47] | 0.78 |
| Liu *et al.*[11] | 0.57 |
| Liu *et al.*[35] | 0.65 |
| Le and Pinkwart[45] | 0.086 |

## Results

It is important to define well-founded evaluation metrics for AQG tasks. Classification performance is measured using balanced F-score, precision and recall (Table 7).

Let us consider the following:

$Q_{\text{aqg}}$: The number of questions generated by a AQG.

$Q_{\text{manually}}$: The number of questions generated manually.

$$\text{Precision} = \frac{Q_{\text{aqg}} \cap Q_{\text{manually}}}{Q_{\text{aqg}}},$$

**Table 9.** Syntactic deficiency in an AQG model

| | AQG models | | | |
|---|---|---|---|---|
| Syntactic deficiency | HSAQG[41] | Divate and Salgaonkar[39] | McConnell[37] | Lindberg[2] |
| Ungrammatical | 14% | 5.17% | 35.5% | 85% |
| Does not make sense | 20.6% | 4.31% | 33.6% | 63% |
| Vague | 19.6% | 11.21% | 23.4% | 78% |
| Wrong Wh word | 4.9% | 30.17% | 20.6% | Not available |
| Formatting errors | 8.9% | 6.90% | 0.03% | Not available |

**Table 10.** Percentage of questions generated by an AQG model (from the given 90 sentences 360 output questions to be generated)

| AQG | Percentage coverage of output questions |
|---|---|
| MrsQG[48] | 98.3 |
| WLV[49] | 45.8 |
| JUQGG[32] | 58.1 |
| Lethbridge[21] | 46.7 |

**Table 11.** Questions classification using J48 classifier

| Rank | Number of questions |
|---|---|
| A | 27 |
| B | 19 |
| C | 8 |
| D | 1 |
| E | 0 |

$$\mathrm{Recall} = \frac{Q_{\mathrm{aqg}} \cap Q_{\mathrm{manually}}}{Q_{\mathrm{manually}}},$$

$$F\text{-score} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{(\mathrm{Precision} * \mathrm{Recall})}.$$

Precision-at-*N* is the percentage of acceptable questions in the top *N* questions. According to Heilman and Smith[41] precision-at-20 for the linear regression ranking model is 45%, i.e. 9 out of 20 questions are acceptable.

In the later models logistic regression, linear regression and RankSVM have been employed for rank computation[14,29,46]. Generally, ranked questions are evaluated by human annotators. When more than one human annotator evaluates the performance of an AQG, the agreement between them (Cohen's kappa) is computed as follows

$$\text{Cohen's kappa} = 1 - \frac{1 - P_{\mathrm{o}}}{(1 - P_{\mathrm{e}})},$$

where $P_{\mathrm{o}}$ and $P_{\mathrm{e}}$ show the observed and estimated values of agreement among the annotators. Table 8 compiles Cohen's kappa values for various AQG models.

Usually the questions generated by an AQG are classified on the basis of the following types of syntactic deficiencies: ungrammatical, does not make sense, vague, obvious answer, missing answer, wrong wh-word, formatting, other. Table 9 gives the syntactic deficiencies identified by the AQG models.

Table 10 shows the volume of output generated by various AQGs models. Clearly, MrsQG outperforms the other three (Yao *et al.*, unpublished). Also, sufficient structural variants are covered due to the reproduction rules of this tool. Further, note that the accuracy of the NER tool results in good coverage on required questions.

## Future trends: automatic question quality enhancer

In one of our experiments with HSAQG, it was observed that more than 50% of the questions generated by an AQG was acceptable to humans[39]. Results are compiled in Table 11, where rank *A* indicates that the question is a well-formed one and rank *E* indicates that it is not acceptable. Ranks *B*–*D* fall in between.

Clearly, further research is necessary for improving the acceptability of the AQG-generated questions. We call such systems as automatic question quality enhancers (AQQEs). The challenges are in the removal of formatting errors, bringing extra precision of questions to enhance clarity and, in selecting suitable answer phrases.

In our observations, the questions generated by employing HSAQG are free from the six types of infirmities, namely correct verb tense, subject–auxiliary verb inversion, leading conjunction phrase, appositive phrase, and question form is negative. This could be an interesting input while designing an AQQE.

## Conclusion

AQG is a thrust area for researchers in natural language processing (NLP). A summary of the recent literature is given here regarding the types of questions, a generic process for automatic generation of questions and representative approaches to developing an AQG. Notable contributions, including a seminal work[3] have been compiled, and evolutionary trends in this area highlighted.

The template-based approach is simpler and more effective compared to the semantic and syntactic approaches, because once a human annotator provides a high-quality question template, generating questions is a mechanical task. In this process, the cost of employing a human annotator is hidden.

It has been observed that despite a significant increase in the number of tools and resources in NLP over the years, the challenges in producing a satisfactorily performing AQG have not significantly changed: the design of templates, identification of semantic roles, processing of complex sentences for identifying answer keys, to name a few.

We have also discussed methods of evaluating AQG. Defining an objective evaluation measure is a critical task; it has remained a difficult research problem for decades. Measures for testing and evaluation of the functionality of an AQG have been widely researched. Generally the performance of an AQG has been computed on the basis of precision, recall and *F*-score. To facilitate novice researchers, we have explained these three parameters in detail.

There is a potential for formulating a mechanism for measuring the impact of an AQG on external parameters, namely students' learning gain, time efficiency and computing the effectiveness of the system from that perspective.

A novel concept, AQQE, has been discussed. Findings of our experiments in the same context have been discussed.

In summary, this article highlights the research in AQG since the seminal work that dates back to 1976 till the recent contributions in 2015. The pointers for futuristic trend should motivate the inquisitive readers to take research in this domain ahead.

1. Rus, V. and Graesser, A. C., The question generation shared task and evaluation challenge. National Science Foundation, The Univesity of Memphis, 2009, pp. 1–48; doi:10.1016/0004-3702(73)90013-1.
2. Lindberg, D., Popowich, F., Nesbit, J. and Winne, P., Generating Natural Language Questions to Support Learning On-Line. In 14th European Workshop on Natural Language Generation, Sofia, Bulgaria, 2013, pp. 105–114.
3. Wolfe, J. H., Automatic question generation from text – an aid to independent study. In *ACM SIGCUE Outlook*, 1976; doi: 10.1145/953026.803459.
4. Aist, G., Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *Int. J. Artif. Intell. Educ.*, 2001, **12**, 212–231.
5. Mostow, J. *et al.*, Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technol. Instr. Cogn. Learn.*, 2004, **2**, 103–140.
6. Brown, J. C., Frishkoff, G. A. and Eskenazi, M., Automatic question generation for vocabulary assessment. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, Vancouver, p. 826; doi:http://dx.doi.org/10.3115/1220575.1220678.
7. Kunichika, H., Katayama, T., Hirashima, T. and Takeuchi, A., Automated question generation methods for intelligent English learning systems and its evaluation. In International Conference on Computers in Educationa, Melbourne, Australia, 2004.
8. Chen, C. Y., Liou, H. C. and Chang, J. S., Fast: an automatic generation system for grammar tests. In Proceedings of the COLING/ACL English Instrustion and Assessment, Sydney, 2006, pp. 1–4; doi:10.3115/1225403.1225404.
9. Wang, X.-J., Tu, X., Feng, D. and Zhang, L., Ranking community answers by modeling question-answer relationships via analogical reasoning. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '09, ACM Press, 2009, pp. 179–186; doi:10.1145/1571941.1571974.
10. Liu, M., Calvo, R. A. and Rus, V., Automatic question generation for literature review writing support. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Berlin, Heidelbrg, 2010.
11. Liu, M., Calvo, R. and Rus, V., G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue Discourse*, 2012, **3**, 101–124.
12. Heilman, M. and Smith, N. A., Question generation via overgenerating transformations and ranking. *Framework*, No. CMU/LTI-09-013, Carnegie-Mellon Univ Pittsburgh pa Language Technologies Inst, 2009.
13. Skalban, Y., Ha, L. A., Specia, L. and Mitkov, R., Automatic question generation in multimedia-based learning. *COLING (Posters)*, 2012, **1**, 1151–1160.
14. Lindberg, D., Automatic Question Generatin from Text for Self-Directed Learning, 2013.
15. Aldabe, I. and Maritxalar, M., Automatic distractor generation for domain specific texts. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6233 LNAI, 2010, pp. 27–38.
16. Kumar, G., Banchs, R. E. and Haro, L. F. D., RevUP: Automatic Gap-Fill Question Generation from Educational Texts, 2015, pp. 154–161.
17. Agarwal, M. and Mannem, P., Automatic gap-fill question generation from text books. In Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Portland, Oregon, 2011, pp. 56–64.
18. Narendra, A., Agarwal, M. and Shah, R., Automatic cloze-questions generation. In Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, 2013, pp. 511–515.
19. Mitkov, R. and Ha, L. A., Computer-aided generation of multiple-choice tests. In Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications using Natural Language Processing, Stroudsburg, PA, USA, 2003; doi:10.3115/1118894.1118897.
20. Aldabe, I., De Lacalle, M. L., Maritxalar, M., Martinez, E. and Uria, L., ArikIturri: An automatic question generator based on corpora and NLP techniques. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Berlin, 4053 LNCS, 2006, pp. 584–594.
21. Ali, H., Chali, Y. and Hasan, S. A., Automatic question generation from sentences. In Proceedings of the Conference on Traitement Automatique de la Langue Naturelle, Montreal, Canada, 2010, pp. 19–23.
22. Chali, Y. and Hasan, S., Towards automatic topical question generation. In COLING, Mumbai, 2012, pp. 475–492.
23. Chen, W., Aist, G. and Mostow, J., Generating questions automatically from informational text. In Workshop Proceedings of the AIED 2009 14th International Conference on Artificial Intelligence in Education, Brighton, UK, 2009, pp. 17–24.

24. Ali, H., Automatic question generation: a syntactical approach to the sentence-to-question generation case. Ph D diss., Department of Mathematics and Computer Science, University fo Lethbridge, Lethbridge, Alta, 2012.
25. Corley, M. A. and Rauscher, W. C., Deeper learning through questioning, TEAL Fact Sheet No. 12, Teaching Excellence in Adult Literacy, US, 2013, pp. 1–5.
26. Yao, X., Bouma, G. and Zhang, Y., Semantics-based question generation and implementation. *Dialogue Discourse*, 2012, **3**, 11–42.
27. Liu, M. and Calvo, R. A., An automatic question generation tool for supporting sourcing and integration in students' essays. In Proceedings of the Fourteenth Australasian Document Computing Symposium 2009, University of Sydney, 2009, pp. 90–97.
28. Adamson, D. *et al.*, Automatically generating discussion questions. In Artificial Intelligence Education: 16th International Conference AIED 2013, Memphis, TN, Springer, USA, 2013, vol. 7926, pp. 81–90.
29. Heilman, M. and Smith, N., Extracting simplified statements for factual question generation. In Proceedings of the 3rd Workshop on Question Generation, Pittsburg, PA, USA, 2010, pp. 11–20.
30. Heilman, M. and Smith, N. A., Ranking automatically generated questions as a shared task. In Second Workshop on Question Generation, 2009, pp. 30–37.
31. Chali, Y. and Hasan, S. A., Towards topic-to-question generation. *Comput. Linguist.*, 2014, **70**, 1–20.
32. Boyer, K. E. and Piwek, P., Proceedings of QG2010: The Third Workshop on Question Generation. In At The Tenth International Conference on Intelligent Tutoring Systems (ITS2010), Pittsburg, 2010, 91.
33. Punyakanok, V., Roth, D. and Yih, W., The importance of syntactic parsing and inference in semantic role labeling. *Comput. Linguist*., 2008, **34**(2), 257–287.
34. Schwartz, L., Aikawa, T. and Pahud, M., Dynamic language learning tools. In Proceedings of InSTIL/ICALL Symposium 2004, Microsoft Research, 2004.
35. Liu, M., Calvo, R. A. and Rus, V., Automatic generation and ranking of questions for critical review. *J. Edu. Technol. Soc.*, 2014, **17**(2), 333–346; https://rafaelacalvo.files.wordpress.com/2013/11/14-jets_ming.pdf (accessed on 2 September 2015).
36. Mannem, P., Prasad, R. and Joshi, A., Question generation from paragraphs at UPenn: QGSTEC system description. In Proceedings of QG 2010: The Third Workshop on Question Generation, Pittsburg, PA, USA, 2010.
37. McConnell, C., Mannem, P., Prasad, R. and Joshi, A., A new approach to ranking over-generated questions. In AAAI Fall Symposium, Palo, Alto, CA, 2011, pp. 45–48.
38. Duh, K., Ranking vs. regression in machine translation evaluation. *ACM Trans. Comput–Hum. Interact.*, 2008, **17**(1), 191–194.
39. Divate, M. and Salgaonkar, A., Ranking model with a reduced feature set for an automated question generation system. In International Conference on Natural Langauge Processessing (ICON), Thiruvananthapuram, India, 2015, pp. 221–230.
40. Rus, V., Cai, Z. and Graesser, A. C., Evaluation in Natural Language Generation: The Question Generation Task. In Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, Hilton Arlington, 2007.
41. Heilman, M. and Smith, N. A., Good question! Statistical ranking for question generation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational, Linguistics, Los Angeles, California, 2010, pp. 609–617.
42. Lin, C. Y., Automatic question generation from quesries. In Workshop on Question Generation Shared Task, Pittsburg, 2008, pp. 156–164.
43. Resnik, P. and Lin, J., Evaluation of NLP systems. In *Handbook Computational Linguistic and Natural Language Processing*, 2010, pp. 271–295; doi:10.3115/993268.993346.
44. Person, N. and Graesser, A. C., Human or computer? AutoTutor in a Bystander turing test. In Proceedings of E-Learn 2002 – World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, 2002, pp. 778–784.
45. Le, N. and Pinkwart, N., Evaluation of a question generation approach using semantic web for supporting argumentation. *Res. Pract. Technol. Enhanc. Learn.*, 2015, **10**, 3.
46. Liu, M., Calvo, R. A., Aditomo, A. and Pizzato, L. A., Using Wikipedia and conceptual graph structures to generate questions for academic writing support. *IEEE Trans. Learn. Technol.*, 2012, **5**, 251–263.
47. Zhao, S., Wang, H., Li, C., Liu, T. and Guan, Y., Automatically generating questions from queries for community-based question answering. In Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang, Mai, Thailand, 2011, pp. 929–937.
48. Yao, X., Question generation with minimal recursion semantics. In *Language and Communication Technologies*, University of Groningen and Saarland University, 2010, pp. 1–92.
49. Varga, A. and Ha, L., Wlv: a question generation system for the qgstec 2010 task b. In Proceedings of the Third Workshp on Question Generation, Pittsburg, 2010, pp. 80–83.

Received 28 June 2016; accepted 29 May 2017

doi: 10.18520/cs/v113/i09/1683-1691