

# Automatic question generation in multimedia-based learning

*Yvonne SKALBAN<sup>1</sup>, Le An HA<sup>1</sup>, Lucia SPECIA<sup>2</sup>, Ruslan MITKOV<sup>1</sup>*

(1) UNIVERSITY OF WOLVERHAMPTON, Stafford Street, Wolverhampton, WV1 1LY, UK

(2) UNIVERSITY OF SHEFFIELD, 211 Portobello, Sheffield, S1 4DP, UK

Yvonne.Skalban@wlv.ac.uk

Ha.L.A@wlv.ac.uk

l.specia@dcs.shef.ac.uk

R.Mitkov@wlv.ac.uk

## ABSTRACT

We investigate whether questions generated automatically by two Natural Language Processing (NLP) based systems (one developed by the authors, the other a state-of-the-art system) can successfully be used to assist multimedia-based learning. We examine the feasibility of using a Question Generation (QG) system's output as pre-questions; with different types of pre-questions used: text-based and with images. We also compare the psychometric parameters of the automatically generated questions by the two systems and of those generated manually. Specifically, we analyse the effect such pre-questions have on test-takers' performance on a comprehension test about a scientific video documentary. We also compare the discrimination power of the questions generated automatically against that of questions generated manually. The results indicate that the presence of pre-questions (preferably with images) improves the performance of test-takers. They indicate that the psychometric parameters of the questions generated by our system are comparable if not better than those of the state-of-the-art system.

---

**KEYWORDS:** Automatic question generation, question evaluation, psychometric parameters

---

## 1 Introduction

Questions are an integral part of teachers' instructional activities. Teachers spend between 35% to 50% of their instructional time conducting questioning sessions (Cotton, 2001). Research in education (Hamilton, 1985; Klauer, 1984; Rothkopf, 1982; Hamaker, 1986; Anderson & Biddle, 1975) has shown that pre-questions, i.e. questions which are supplied to test-takers before receiving learning material, can have beneficial effects on student learning in reading activities. Pre-questions can help focus learners' attention on the learning material targeted by the questions and they also increase the learning effect through repetition (Thalheimer, 2003). The manual creation of questions is time-consuming and requires the knowledge of domain experts. Research in Natural Language Processing (NLP), indicates that systems for Question Generation (QG) can assist teachers in this laborious task, thus saving time and resources. Semi-automatic QG systems can produce test questions up to 4 times faster than a human expert, without compromising quality (Mitkov et. al, 2006). In this experiment, we examine whether the questions produced by our system can be successfully used as pre-questions and thus support creators of assessment materials. Two different types of pre-questions are investigated: text-based and with supporting image. This experiment also serves to test whether pre-questions have a beneficial effect in combination with audio-visual learning material as opposed to reading material; we analyse the effect pre-questions have on test-takers' performance on a comprehension test about a scientific video documentary. We also examine whether or not questions generated automatically (by two systems) have the same psychometric parameters as those generated manually. The psychometric parameters of questions, such as their discrimination power, are among the most important measures of the quality of the questions.

## 2 Related Work

QG has frequently been employed in educational contexts. Applications include systems which automatically create learning resources such as multiple-choice question (MCQ) tests (Mitkov, 2003, Mitkov et. al., 2006), vocabulary exercises (Brown et. al., 2005, Hoshino and Nakagawa, 2007), as well as solutions which promote reading comprehension (Feeney and Heilman, 2008; Gates, 2008). QG systems help promote student learning by providing learning content and forms of assessment which allow for convenient and fast evaluation of student performance. Several systems have been developed to automatically generate questions from texts using NLP techniques, with a system developed by Heilman (2011) showcasing the state-of-the-art. The system generates questions from reading material for educational practice and assessment using existing tools such as the Stanford parser (Klein and Manning, 2003), Tregex expressions for T-Surgeon (Levy and Andrew, 2006), and BBN Identifier (Bikel, et. al., 1998). The QG process follows several stages. Firstly, sentences are simplified by removing certain discourse markers and adjunct modifiers and by breaking sentences down into clauses. Next, pronoun resolution is performed using the ARKref coreference system (Heilman, 2011). A complex set of transformational rules implemented in Tregex is then used to form *who*, *what*, *where*, *when* and *how much* questions from declarative statements. Since one sentence in the source text can give rise to a number of questions, the questions are statistically ranked in terms of quality before being displayed to the user.

### 3 Methodology

This section describes the author's QG system and the experimental setup and execution of the in-class experiment.

#### 3.1 A QG system for documentary videos

The QG system we developed employs existing NLP tools (GATE, Cunningham, et. al., 2002) for pre-processing and a rule-based approach to generate factual questions from documentary videos, utilizing the subtitles accompanying a documentary. Several of GATE's processing resources (PRs) are employed to pre-process the subtitles; steps include tokenization, sentence splitting, POS tagging, dependency parsing, NE recognition, gazetteer look-up, morphological analysis and co-reference resolution. The PRs enrich the text with linguistic information in the form of annotations, which is exploited in the subsequent steps. Pronoun resolution is performed, based on information provided by GATE's pronominal co-referencer. First-mention pronouns are replaced with the longest co-referent in the co-reference chain. In independent clauses in compound sentences, not only first-mention pronouns, but all subject personal pronouns are replaced with their co-referents. Then the compound sentences are split into several sentences with initial conjunctions deleted. Next, several transformational rules, written in a GATE-specific format (JAPE), are applied. These rules consist of a left hand side (LHS), which is used to match a pattern in a GATE corpus (in our case subtitles) and a right hand side (RHS) which is used to perform actions and to manipulate the text and parse trees. We distinguish between question rules and helper rules. Question rules are used to identify question candidates in the source text. By using the linguistic information made available in the pre-processing steps and the application of syntactic transformations (such as WH-movement and subject-auxiliary inversion) declarative sentences are transformed into questions. Currently, six types of 'wh-questions' can be generated: questions about persons (*who*, *whom*), temporal questions (*when*), questions about possessives (*whose*), location questions (*where*) and questions about inanimate entities (*what*). It has been designed to work with video subtitles, and as a result, is able to explore their unique attribute: each utterance has a time-stamp. These time-stamps can be used to link the texts with their relevant video section. In this experiment, we use this feature to extract relevant screenshots for the questions.

#### 3.2 Definitions

*Pre-questions* are supplied to test-takers before receiving learning material (here: the documentary video). Pre-questions are non-scoring and do not require an answer. Pre-questions can be text-only or can be accompanied by a relevant image. In this experiment, images are screenshots extracted from the video.

*Post-questions* are presented to the test-takers after receiving learning material (here: after watching a documentary). Post-questions are generated either manually by a human expert or automatically. The post-questions employed in this experiment are short answer style questions.

*System A* is the QG system designed by the authors, as described in section .

*System B* is the QG system developed by Heilman (2011). Its methodology is explained in section 2.

### 3.3 Research questions

The aim of the experiment is to answer the following research questions:

1. a) Whether the presence of text-based pre-questions helps test-takers to answer post-questions more accurately (i.e. more questions are answered correctly).  
b) Whether the presence of pre-questions *with screenshots extracted from the video* helps the test-takers to answer post-questions more accurately.
2. a) Whether the presence of text-based pre-questions affects the time taken to answer post-questions.  
b) Whether the presence of pre-questions *with screenshots extracted from the video* affects the time taken to answer post-questions.
3. What are the psychometric parameters of questions generated by system A when compared to system B and manually generated questions?

### 3.4 Selection of system-generated post-questions

Due to the nature of their QG approach, both QG systems produced more questions (A: 139, B: 567) than required for the experiment. Only 9 questions were needed from each method for the participants to complete the experiment in approximately one hour. As system B uses certain heuristics to output questions ranked in terms of quality, the top 3 questions corresponding to the respective parts of the video were selected for use in the experiment. From system A's pool of questions, 3 questions per part were selected by a human expert.

### 3.5 Generation and selection of human-generated questions

The manually generated questions were obtained from a high school teacher of English and Media. The teacher was given access to the documentary video and a transcript and was asked to produce comprehension questions that they would also use in their classroom were they to utilize this video in one of their teaching sessions. The teacher was also instructed to generate the questions in such a way that they could be answered solely with information from the video and did not require any additional knowledge. The human expert generated 22 questions in about 80 minutes, 9 of which were selected for the experiment at random.

### 3.6 Selection of pre-questions

For the first two hypotheses, the focus is on whether or not pre-questions help the performance of test-takers, rather than the generation method of pre-questions. As a result, pre-questions were selected manually from system A's pool of generated questions. Pre-questions were selected based on two premises. Firstly, a question was deemed a suitable pre-question if it revolved around an important concept in the documentary. Secondly, a question was selected as a pre-question if the same or a similar question was also generated by one or more of the other systems. For example, the question "What is nuclear fusion?" was selected as a pre-question because it revolves around a central concept in the documentary. In addition, the same question was generated by the human expert. An example for similar questions generated by all three methods can be seen in Table 1. The development of automatic selection methods for pre-questions and their evaluation will be left to future research.

System A	What did some scientists suspect that Rusi Taleyarkhan’s fusion neutrons could in fact be coming from?	From his own neutron generator
System B	What did Mike Saltmarsh think that any fusion finding could be explained by?	From the pulse neutron generator
Manual	What did the other scientists criticise about Taleyarkhan’s first experiment?	Other scientists criticised that the neutrons detected in the experiment might be background neutrons from the neutron generator.

TABLE 1 Questions with similar content generated by all three QG methods

### 3.7 Selection of images

The screenshots are extracted using the following process. After questions have been generated, the source sentence of a question (i.e. the sentence which gave rise to a question) is mapped to the time stamp contained in the subtitles. Then a screenshot is taken from the video at the respective time a source sentence occurs in the video. For example, the sentence “It was Mike Saltmarsh’s task to work out whether the neutrons detected could indeed be from fusion or were simply background neutrons from the neutron generator” which occurred 29 minutes and 15 seconds into the video gave rise to the first question and screenshot in Table 2.


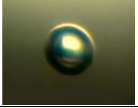

Whose task was to work out whether the neutrons detected could indeed be from fusion or were simply background neutrons from the neutron generator?	Mike Saltmarsh	
What should be produced at exactly the same billionth of a second if fusion was happening?	Fusion neutrons	
What is nuclear fusion?	A nuclear reaction in which atoms are forced together until they fuse, giving off massive amounts of heat, light and energy.	

TABLE 2 Pre-questions with screenshots extracted from the video

### 3.8 Participants and interface

29 students took part in the experiment. All participants were final year undergraduate students at a university Spain reading translation with a major in English. The participants had access to the experiment via an online interface<sup>1</sup>. Instructions for the experiment (e.g. note-taking was allowed, but participants should watch the video only once) were displayed in the interface. The interface provided access to the video and tracked each participant’s answers and time spent to answer each question.

<sup>1</sup> The experiment can be accessed at: <http://www.bootlace.eu/quiz/randq/>

### 3.9 Procedure

The video used was a documentary on ‘nuclear fusion’ (Horizon, 2005). The experiment consisted of three parts, each corresponding to a 10-minute section of the documentary. The participants were divided into three groups. Before each part of the video was shown, participants were given either three pre-questions containing a screenshot extracted from the video, three text-only pre-questions or no pre-questions, depending on their group (cf. Table 3). After each part of the video, the students were asked to answer nine comprehension questions (post-questions) about what they had just seen in the video. Three of those questions had been generated by system A, three by system B and three by a human expert. The post-questions were identical for all participants. This group scenario was used in order to eliminate the problem of cross group performance comparison and cross-question performance comparison.

	Group 1	Group 2	Group 3
Part 1	Pre-questions + screenshots	Pre-questions no screenshots	No pre-questions
Part 2	Pre-questions no screenshots	No pre-questions	Pre-questions + screenshots
Part 3	No pre-questions	Pre-questions + screenshots	Pre-questions no screenshots

TABLE 3 Pre-question scenarios

## 4 Results

### 4.1 Answering research question 1: accuracy

Firstly, a  $\chi^2$  test of independence was used to determine whether the performance across the groups differed significantly; there was no evidence to suggest so. Table 4 shows the breakdown of correctly and incorrectly answered post-questions for each pre-question type ( $Q_{np}$ =no pre-questions,  $Q_{tp}$ =text-based pre-questions,  $Q_{sp}$ =pre-questions with screenshots). Due to time constraints, not all test-takers answered all questions, which is the reason for the total number of questions answered varying for each pre-question type. Proportionally, the highest number of correctly answered questions is observed where test-takers were given pre-questions with screenshots, followed by text-based pre-questions. Test-takers who did not receive any pre-questions at all produced the smallest proportion of correct answers.

Pre-question type	Correct	Incorrect	Total	% correct
$Q_{np}$	75	113	188	39.83
$Q_{tp}$	86	85	171	50.29
$Q_{sp}$	84	60	144	58.33
$(Q_{tp}+Q_{sp})$	(170)	(145)	(315)	(53.97)

TABLE 4 Breakdown of correct and incorrect answers per pre-question type

A  $\chi^2$  test was performed to determine whether these results are statistically significant. When comparing the performance of students who did not receive pre-questions ( $Q_{np}$ ) to the performance of students who received only text-based pre-questions ( $Q_{tp}$ ), the result is

statistically significant ( $p= 0.047$ ). The same applies when the performance of students who did not receive pre-questions is compared with that of students who that received pre-questions with screenshots ( $Q_{sp}$ ); we observed a better statistically significant difference ( $p=0.00085$ ). When text-based pre-questions and pre-questions with screenshots are grouped together ( $Q_{tp}+Q_{sp}$ ) and compared to no pre-questions ( $Q_{np}$ ), the result is also statistically significant ( $p=0.00225$ ). However, when comparing the performance of students who received text-based pre-questions with that of those who received pre-questions with screenshots, we found no statistically significant difference ( $p=0.1537$ ). We can thus conclude that test-takers who receive pre-questions (with or without image) tend to perform better on a comprehension test than those who receive no pre-questions at all. Supplying a screenshot alongside a pre-question results in a more significant difference of correctly answered questions when compared to text-based pre-questions.

#### 4.2 Answering research question 2: time taken to answer post-questions

For each test taker, the time to answer a question was measured. We hypothesized that the presence of pre-questions would affect the time taken to answer post-questions. We observed that the highest mean value (cf. Table 5) occurred in the pre-questions with screenshots condition ( $Q_{sp}$ ), followed by text-based pre-questions ( $Q_{tp}$ ). The lowest average time required to answer a question was observed in the no pre-questions condition ( $Q_{np}$ ). However, there appears to be no significant difference between the means of the different conditions, which is confirmed by a single-factor analysis of variance. We can thus conclude that the presence of pre-questions, with or without screenshot, does not affect the time taken to answer post-questions.

	Min t in s	Max t in s	Mean	SD
$Q_{np}$	2	237	53.26	44.38
$Q_{tp}$	3	403	54.84	55.07
$Q_{sp}$	5	306	58.57	46.49

TABLE 5 Seconds taken to answer post-questions depending on pre-question type

#### 4.3 Answering research questions 3: psychometric parameters

Classical test theory can provide information about the effectiveness of a question (also referred to as 'item'). One measure is item discriminating power (DP) (Gronlund, 1982). DP describes the relationship between student performance on a particular item and their total exam score. DP ranges from -1.0 to 1.0; the higher the value, the more discriminating the item. A high DP means that test takers with overall high scores answered the item correctly, whereas test takers who performed poorly overall did not answer the item correctly. On the converse, a low DP indicates that poorly performing test takers answered an item correctly whereas test takers with overall high scores did not answer an item correctly; this means that the item may be confusing for better scoring test takers. Items with near zero or negative DP should not be used for assessment. To calculate DP, test results need to be ranked from highest to lowest score. Two equal-sized groups are formed, the 'upper group' containing the tests with the highest scores, and the 'lower group' containing those with the lowest scores. DP is calculated as follows:

$$DP = \frac{R_U - R_L}{\frac{1}{2}P}$$

Where DP is the discriminating power,  $R_U$  is the number of right answers from the upper group,  $R_L$  is the number of right answers from the lower group,  $P$  is the number of total participants. The results for the discriminating power for the three QG methods can be seen in Table 6.

	Min	Max	Mean DP
System A	-0.15	0.44	0.16
System B	-0.22	0.22	0.07
Manual	0.15	0.59	0.37

TABLE 6 Discriminating powers for all three QG methods

The manually created questions exhibit the highest average DP, followed by system A and lastly system B. The application of Student's t-test shows that there is a statistically significant difference between system A's mean DP and the manual questions' mean DP ( $p=0.0434$ ). The same applies when comparing system B's mean DP to that of the manual questions. However, no statistically significant difference could be observed between system A's and system B's mean DPs ( $p=0.356988$ ). While this means that neither automatic system's questions are as good as questions generated by human experts at distinguishing between well and poorly performing students, it also means that system A's questions are as good as, if not better than, those generated by the state-of-the-art system.

## Conclusion and directions for future research

Our findings show that both text-based pre-questions and pre-questions with images lead to a larger number of correctly answered post-questions (as opposed to using no pre-questions). Supplying a screenshot alongside a pre-question will result in a statistically more significant difference of correctly answered questions when comparing to no pre-questions. The ability to supply a screenshot alongside a question is unique to our system. The average time taken to answer a question is not statistically significantly different between the pre-question settings. We analysed whether questions generated by our system have a discriminating power (DP), comparable to that of questions generated by human experts and a state-of-the-art system. We found that manually created questions exhibit the highest DP and there is no statistically significant difference between our system and the state-of-the-art system, implying that questions generated by our system are as good as, if not better than, questions generated by the state-of-the-art system. A number of issues need to be addressed in future research. The feasibility of automatically or semi-automatically choosing pre-questions needs to be explored. Furthermore, we aim to investigate whether other images taken from other sources (e.g. Google Image search) can also be used in pre-questions. A large-scale experiment investigating the productivity of generating questions (time taken to post-edit questions vs. time taken to generate questions from scratch) is planned.



## References

- Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.) *The psychology of learning and motivation: Advances in research and theory* (Vol. 9). New York: Academic Press.
- Bikel, D., Schwartz, R., Weischedel, R. (1999). An Algorithm that Learns what's in a Name. *Machine Learning- Special Issue on NL Learning*, 34, 1–3.
- Brown, J.C., Frishkoff, G.A, Eskenazi, M. (2005). *Automatic question generation for vocabulary assessment*. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Canada, 2005.
- Cotton, K. (2001). *Classroom questioning*. School Improvement Research Series.
- Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.
- Dantonio, M. (2001). Developing effective teacher questioning practices. *Learning to Question, Questioning to Learn*, 1, 6-10.
- Feeney, C. and Heilman, M. (2008). *Automatically generating and validating reading-check questions*. In Proc. of the Young Researcher's Track. Ninth International Conference on Intelligent Tutoring Systems.
- Gates, D. M. (2008). *Generating reading comprehension look-back strategy questions from expository texts*. Master's thesis, Carnegie Mellon University.
- Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, 56, 212-242.
- Hamilton, R. J. (1985). A framework for the evaluation of the effectiveness of adjunct questions and objectives. *Review of Educational Research*, 55, 47-85.
- Heilman, M. (2011). *Automatic Factual Question Generation from Text*. Ph.D. Dissertation, Carnegie Mellon University. B-LTI-11-004.
- Horizon (2005). [online]. Accessed 30/05/2012 < <http://www.bbc.co.uk/iplayer>>.
- Klein, D., Manning, C. (2003). *Fast Exact Inference with a Factored Model for Natural Language Parsing*. In Advances in Neural Information Processing Systems 15 (NIPS 2002), 3-10, Cambridge, MA.
- Levy, R., Galen, A. (2006). *Tregex and Tsurgeon: tools for querying and manipulating tree data structures*. Proceedings of LREC 2006.
- Mitkov, R. and Ha, L. A. (2003). *Computer-Aided Generation of Multiple-Choice Tests*. In Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing. Edmonton, Canada, 2003.
- Mitkov, R., Ha, L. A., Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2), pp.177-94.

Hoshino, A., Nakagawa, H. (2007). *Assisting cloze test making with a web application*. In Proceedings of Society for Information Technology and Teacher Education International Conference. Chesapeake, VA, 2007. AACE.

Rothkopf, E. Z. (1982). *Adjunct aids and the control of mathemagenic activities during purposeful reading*. In W. Otto & S. White (Eds.) Reading expository material. New York: Academic Press.

Thalheimer, W. (2003). *The learning benefits of questions*. [online]. Accessed 30/08/2012. <<http://www.learningadvantage.co.za/pdfs/questionmark/LearningBenefitsOfQuestions.pdf>>.