TEDQuiz: Automatic Quiz Generation for TED Talks Video Clips to Assess Listening Comprehension

Yi-Ting Huang, Ya-Min Tseng, Yeali S. Sun
Department of Information Management
National Taiwan University
Taipei, Taiwan
{d97725008, r01725009, sunny}@ntu.edu.tw

Meng Chang Chen
Institute of Information Science
Academic Sinica
Taipei, Taiwan
mcc@iis.sinica.edu.tw

***Abstract*—In the last few years, researchers in the field of e-learning and Natural Language Processing (NLP) have shown an increased interest in automatic question generation. However, little research has discussed the automatic evaluation of listening comprehension in multimedia learning. In this work, we present an automatic quiz generation for TED Talks video clips, called TEDQuiz. TEDQuiz generates multiple–choice questions in two question types, gist-content questions and detail questions. We use a graph-based algorithm, LexRank, to identify the most important part of a talk, as the main concept of a gist-content question. We also proposed an approach to distractor selection for detail question generation that generates grammatically correct but semantically wrong sentences as distractors. The experimental results demonstrated that the measured results from automatically generated questions are comparable with that from manually generated questions because their scores were significantly correlated. Moreover, most subjects agreed that the generated listening comprehension questions were of quality and usefulness.**

Keywords- question generation; computer assisted language learning; multimedia learning

I. INTRODUCTION

In recent years, there has been increasing attention to automatic question generation in the field of e-learning and Natural Language Processing (NLP). Automatic question generation is the task of automatically generating questions from some form of input. It is useful in multiple subareas and has been used in generating instructions in tutoring system, assessing domain knowledge, evaluating language proficiency, assisting academic writing and closed domain question answering.

In order to make learning environment more effective and efficient, many researchers have been exploring the possibility of automatic question generation in various contexts. For example, a wide variety of applications, such as Linguistics and Biology, identified the important concepts in textbooks and generated multiple-choice questions and gap-fill questions [1][10]. In the domain of language learning, a growing number of studies are now available to not only drills and exercises [9][11][16], including vocabulary, grammar, reading questions, but also formal exams, including SAT (Scholastic Aptitude Test) analogy questions and TOEFL (Test of English as a Foreign Language) synonym task [15].

Though substantial studies have been developed on various fields, we observed that these generated questions become insufficient when the goal turns to assessing learners' listening comprehension of a given content. With the development of computers and the Internet, learners have more choices and resources that enable learning. Learning is possible through online courses or other online materials, such as videos on YouTube[1] or TED Talks[2]. There is thus an increasing demand for automatic assisting tools that help learners evaluate their understanding and comprehension in multimedia learning [14]. However, the tests generated from the previous work only evaluate examinees' reading skills and little research has work on tests that measure examinees' listening understanding.

This work presents an automatic quiz generation for assessing listening comprehension of video clips in English. In order to test students with a top-down approach, the system generates two question types, gist-content questions and detail questions. The gist-content questions test the main idea of a listening passage, while the detail questions are related to the details or facts from the passage. Through this approach, English language learners can evaluate their understanding through exercises after watching online listening materials. The experimental results demonstrated that the measured results from the automatically generated questions were significantly correlated with those from the manually generated questions. Moreover, in the questionnaire results, most subjects agreed that the generated listening comprehension questions were of quality.

II. RELATED WORK

Automatic question generation (also called computer-aided question generation) is the task of automatically generating questions, which consists of producing a stem, a correct answer and several distractors, when given a text. The use of computer-aided question generation for educational purpose was motivated as research of reading comprehension consistently found that assessment is helpful

[1] http://www.youtube.com/
[2] http://www.ted.com/

978-1-4799-4038-7/14 $31.00 © 2014 IEEE
DOI 10.1109/ICALT.2014.105

to learning and enhances learners' retention of material. MARCT system [16] investigated reading comprehension and designed three question types, including true-false question, numerical information question and not-in-the-list questions. Huang et al. [9] captured the relations between sentences to generate two kinds of meaningful reading questions based on noun phrase coreference resolution. Mostow and Jang [11] designed reading questions to diagnose different types of comprehension failure. However, in this previous work, these computer-aided question generations were only designed for reading comprehension.

Question generation has been primarily concerned by the natural language processing community through the question generation workshop and the shared task in 2010 (QGSTEC 2010) [12]. The aim of the task is to generate a series of questions based on the raw text from sentences or paragraphs. The question types includes *why, who, when, where, when, what, which, how many/long* and *yes/no questions*. Many generation approaches to *wh*-questions have been developed, inclusive of template-based, syntactic-based, semantic-based, and discourse-based approach. So far, the work of Heilman and Smith [6][7] is one of the state-of-the-arts. They analyzed the structures of sentences and proposed general–purpose rules using part-of-speech (POS) tags and category labels. Their question generation system, which derives simplified sentences from complex sentences and transforms declarative sentences into questions, can generate grammatical and readable questions rather than unnatural or senseless questions. However, unlike the computer-aided question generation, which is directly related to the topic of generating questions for educational purpose, the question generation of these related studies only focused on generating questions (stems) based on the given content, they were not involved in the distractor selection.

The most related work is Liu et al. [5], which produces a listening cloze item by extracting a sentence, mining a gap in the sentence, and selecting words with similar phonemes as distractors. With this approach, learners can click on alternatives to listen to the recorded sounds, and select one alternative as their answer to fill the gap. This method, however, only focuses on the word recognition in listening rather than the comprehension of the listening passage.

III. METHOD

A. System overview

To assist English language learners in listening comprehension, this paper proposed an automatic question generation for multimedia learning. Here, we adopted TED Talks as research materials. TED is an acronym for Technology, Entertainment, and Design. Experts in different fields are invited as speakers to give a short talk in order to spread their positive ideas. The quality of TED Talks is good enough for English language learners to acquire new knowledge and learn language. Moreover, the length of a talk, less than eighteen minutes, is short enough to sustain leaners' attention.

Figure 1 illustrates the system architecture of the proposed computer-aided comprehension question generation process, named TEDQuiz. A learning material, such as streaming media, is crawled by a crawler. Here, the transcripts of the TED Talks are extracted as the given content. In order to reduce the redundancy in spoken language, such as "you know", "hmm", the transcripts are then transformed into simplified sentences in written language by the sentence simplification proposed by Heilman and Smith [8]. When a learning material is given, questions are created from the question generation process, which is composed of gist-content question generation and detail question generation. Questions generated from the gist-content question generation test the understanding of the main idea of the listening passage. These questions that test the understanding of the gist may require the learner to generalize or synthesize the information that he or she has acquired from the learning material. If the learner answers a gist-content question correctly, it probably means that he or she understands the general topic or main idea. Questions from the detail question generation are typically related to the main idea directly. If a learner answers a detail question correctly, it probably means that he or she understands the explicit details or facts listed in the listening passage.

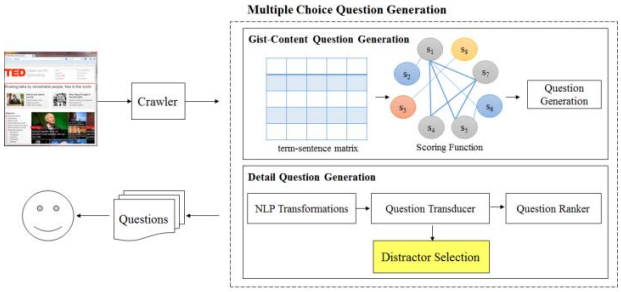

Figure 1. System architecture of TEDQuiz

B. Gist-content question generation

In this section, we discuss how to identify the main idea of a learning material and how to generate gist-content questions. The stems of gist-content questions ask about the overall content of the listening passage, the correct answers describe the closest overall theme of the content, and the distractors refer to only small portions of the content.

To measure the importance of sentences of a learning material, we use LexRank [3] to identify the most important sentences from the given body of text as the main theme of the content. LexRank is a graph-based method for computing the importance of sentences based on the concept of eigenvector centrality. At first, as seen in Figure 1, LexRank builds a graph of all the sentences of a document in a cluster. The nodes represent the sentences and the edges are the cosine similarity between them. After that, LexRank computes the salience of sentences based on their eigenvector centrality in the graph and ranks sentences using a variant of PageRank [2] over the words of sentences. Once the computation of LexRank is done, the most and the least important concepts of the given content can be identified. If a sentence has the largest LexRank score, it probably means that this sentence highly correlates to the other sentences.

The sentence can be treated as the main idea of the given content. On the other hand, when the LexRank score of a sentence is small, it probably means that the sentence could be detailed information or digression.

Once the main idea and off-topic part of the given content are recognized, a gist-content question can be generated. The stem of a gist-content question is generated based on predefined templates, for example, "*What is the main topic of the listening passage?*", "*What is the talk mainly about?*", and "*Which of the following is closest to the main idea of the talk?*". The correct answer of the question is derived from the most important sentence, which is selected by LexRank; On the other hand, the plausible distractors of the question are from the least important sentences. As an example, Figure 2 presents a question generated from the TED Talks "*Russell Foster: Why do we sleep?*[3]". The bolded sentence is selected because it has the highest LexRank score, while the sentences in the distractors are the least salient ones in the given content. All of them are paraphrased in order to distinguish advanced learners, who should understand the main theme of the given learning material and identify the incorrect description within the distractors, from non–advanced learners, who might fail to comprehend the meaning of the given content and choose the incorrect alternatives because of their misunderstanding.

Q1. Which of the following is closest to the main idea of this talk?
(1) We've had three explanations for why we might sleep.
(2) When you're tired, and you lack sleep, you have poor memory, poor creativity, increased impulsiveness, and overall poor judgment.
(3) If you have good sleep, it increases your concentration, attention, decision-making, creativity, social skills, health.
(4) You do not do anything much while you're asleep.

Figure 2. Example of a gist-content question

C. Detail question generation

In this section, we discuss how to generate the distractors of the detail questions. The stems of detail questions are typically factual questions, which refer to *what, when, who, whose*, and *how many*. Here, we extend Heilman and Smith's work [6][7], which composed general-purpose rules to transform declarative sentences into questions. Based on the factual questions generated by Heilman's question generation system, our goal is to generate grammatically correct but semantically wrong sentences as distractors. Figure 3 presents an example of detail question, generated from the same TED Talk, "*Russell Foster: Why do we sleep?*". While the stem of the question is generated by Heilman and Smith's work [6][7], the alternatives are generated by the distractor selection proposed in this paper. The bolded sentence is the answer and the rest of the alternatives are distractors, which have higher rank among the distractor candidates. The bold underline in the bolded sentence marks the head word of the answer phrase. The purpose of the proposed distractor selection is to find suitable replacements (the italic underline in the distractors) and combine them into final distractors.

To produce such plausible distractors, we select words from those that appear in the given content to replace the head word of an answer phrase. The intuition is that when a learner does not fully comprehend a given content, words that also appear in the listening passage become as plausible as the correct answer key, with the premise that the output distractors make sense. The distractor selection analyzes the syntactic, semantic, locational and n-gram information of the content words in the given content, selects three of the content words as replacements, and rewrites the answer sentence by replacing the head word of the answer phrase with the replacements, as distractor candidates.

Q2: What becomes a dominant part of the vasculature?
(1) *Glucose* becomes a dominant part of the vasculature.
(2) *Carbohydrate* becomes a dominant part of the vasculature.
(3) *Caffeine* becomes a dominant part of the vasculature.
(4) *Dew* becomes a dominant part of the vasculature.

Figure 3. Example of a detail question

There are three selection strategies adopted in the distractor selection. The first strategy is based on the semantic information from SuperSense Tagger (SST)[4] and is assigned the highest selection priority. SuperSense Tagger is a tool to assign each content word to one of the 45 WordNet[5] categories. For example, *minute* and *day* are annotated as *B-noun.time* by SST, *toe* and *hand* are tagged as *B-noun.body*. The content words with the same SuperSense tag as the head word of the answer phrase have the highest priorities in the distractor selection. The second strategy is based on syntactic information, part-of-speech (POS) tag, and n-gram information from Google 1T 5-gram corpus[6], with the medium priority. We first select replacements from content words with the same POS tag as the head word of the answer phrase, but we observed that this often leads to nonsensical sentences. Therefore, we further check the corresponding n-gram existence in the corpus to make sure if the context and the replacement can be used in this way. If the n-gram formed by a replacement along with its context is found existing in the Google n-gram corpus, it means that the n-gram appears at least 40 times on the Web. Thus, we have more confidence to take the replacement as a reasonable distractor. Finally, the third strategy is based on the n-gram model from the BNC corpus[7], which is assigned with the lowest priority. With this final strategy, we select content words with the same POS tag as the head word of the answer phrase as replacements. In order to select good replacements, we measure the relevance between distractor candidates and

[3] http://www.ted.com/talks/russell_foster_why_do_we_sleep.html

[4] SuperSense tagger: http://medialab.di.unipi.it/wiki/SuperSense_Tagger
[5] WordNet: http://wordnet.princeton.edu/
[6] Google 1T 5-gram corpus: http://catalog.ldc.upenn.edu/LDC2006T13
[7] BNC corpus: http://www.natcorp.ox.ac.uk/

the corresponding answer sentence to prefer less abrupt choices. Using BNC corpus, we calculate the Bayesian probability of a distractor candidate given all the context words in the answer sentence [11], as the following formula,

$$\Pr(c|\vec{w}) \propto \Pr(c) \prod_{i=1}^{n} \Pr(w_i|c) \quad (1)$$

where c indicates the replacement and w is a vector of context words.

There are two problems to address, regarding the distractor candidates. One is that the distractor candidates could be a reasonable answer, the other is that the multiple-choice question here should contain exactly three distractors. To prevent the former problem, in the beginning of the distractor selection, we filter out the content words that are synonyms or hypernyms of the head word of the answer phrase. We acquire synonym and hyponym information using WordNet. To deal with the second problem, after collecting the distractor candidates, we count the number of distractor candidates. If the number is larger than three, the distance ranker select distractor candidates by locational information. If the number is less than three, the question is discarded.

In Figure 4, the distractor selection is organized as follows,

Step 1. For any content word *c* in the given content, check whether the word *c* is the synonym or the hypernym of the head word of the answer phrase. If it is, the word *c* is filtered out.

Step 2. Select three words with the same SuperSense tag as distractor candidates (as the first strategy).

Step 3. If there are less than three replacements found in Step 2, we turn to select from all words with the same POS tag. And then the words are checked if any of their combinations with the context exists in the Google n-gram Corpus (as the second strategy).

Step 4. If there are less than three replacements found in Step 2 and 3, BNC scores are calculated for all same-POS-tag candidates. Candidates with higher scores are chosen first (as the third strategy).

Step 5. If, at any step, the number of candidates is more than the required number, the closer the original sentence location of a candidate is to the source sentence of the target question, the higher the priority it acquires.

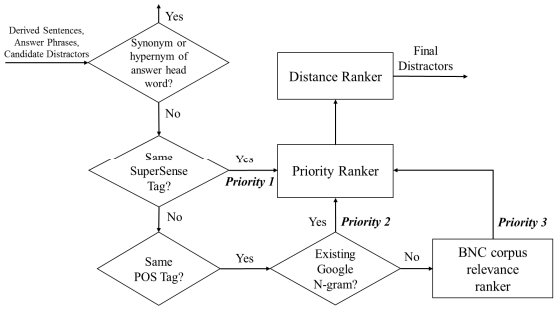

Figure 4. Distractor selection of detail question generation

Given the answer phrase of each question, the head word of the answer phrase is identified by the function provided by Stanford Parser[8]. We replace the head word of the answer phrase with the replacements, and perform noun inflection or verb tense matching accordingly in order to form grammatically correct distractor.

Sometimes, a sentence may lead to several questions. To determine which questions are preferred, we rank the resulting questions by combining scores calculated from predefined features, shown in Table I. First, we consider the relevant portions of a question and an answer in the given content as the relative importance. The importance of a question refers to the LexRank score discussed in the previous section, and that of an answer is generated based on [4]. Then, to represent the quality of distractors, the adopted selection strategy is included as a feature. The higher the priority of the selection strategy, the better the quality. The quality of a stem is represented by the predicted value from the logistic regression model of stem acceptability [6][7]. Finally, questions are ranked by the following formula:

$$\max \sum_{i=0}^{n} \alpha_i f_i \quad (2)$$

Here α is the importance parameter which holds a value in [0,1]. We kept α= 0.25 to give equal importance to each of features.

TABLE I. FEATURES OF DETAIL QUESTION SELECTION

feature	the description of feature
$\mathbf{f_1}$	the relative importance of a question
$\mathbf{f_2}$	the relative importance of an answer
$\mathbf{f_3}$	the selection strategy adopted by the distractor selection
$\mathbf{f_4}$	the linguistic quality of a stem

IV. EXPERIMENT

In this section, we investigated whether learners could be measured by answering the generated questions. We also examined whether the quality of the generated questions could be acceptable for use of computer-based assessment.

A. Experimental setting

Forty senior high school students in Taiwan, who take English as a foreign language (EFL), participated the experiment. During the experiment, the subjects were asked to complete two tasks. Each task consisted of watching a TED Talk, answering five questions generated by TEDQuiz, and then answering five questions created by a human expert. Finally, they answered a questionnaire to express their experience on the quality of generated questions.

The TED talks used in the experiment were "*Gregory Petsko: The coming neurological epidemic*[9]" and "*Andrew Blum: Discover the physical side of the internet*[10]". They

[8] http://nlp.stanford.edu/software/lex-parser.shtml
[9] http://www.ted.com/talks/gregory_petsko_on_the_coming_neurological_epidemic.html
[10] http://www.ted.com/talks/andrew_blum_what_is_the_internet_really.html

were selected as experimental materials because of the difficulty of the content and the length of the talks. Each test was composed of one gist-content question and four detail questions. The questionnaire was extracted from [13], which developed a computer based assessment acceptance model. Since TEDQuiz is in the initial development stage, this study only focuses on one of the constructs in the questionnaire, *Content*. A five–point Likert scale was employed.

B. Measurement validation

To validate the performance of the proposed TEDQuiz, the scores from two different generations were compared. Table II reports the Pearson's correlation coefficient between the scores from the automatically generated test and those from the manually generated test, on the two TED Talks. All of the measures are significantly positively correlated. Especially, the correlation on Gregory Petsko's talk showed a high relevance ($r=0.57$, $p<0.001$) between the two types of generation. This suggests that the measured results from the automatically generated questions can be comparable with the results from the manually generated questions.

TABLE II. THE CORRELATION RESULTS

TED Talks	***Gregory Petsko***	***Andrew Blum***
correlation	0.57 ($p<0.001$)	0.35 *($p<0.05$)*

C. Results from questionnaire

In terms of evaluating the performance of our automatic question generation, four questions in the questionnaire concerning the subjects' perception were investigated. Table III displays the questionnaire results on the detail questions, the percentage and the mean score and standard deviation. Overall, most participants were satisfied with the functionality of the generated questions. Over eighty percent of subjects agreed that the generated questions were understandable. Especially, more than ninety percent of them agreed that the selected questions were highly related to the Talk, and they agreed that the system was useful to them. This data supports the usefulness of the generated questions.

V. CONCLUSION

In this work, we present an automatic quiz generation for TED Talks video clips, called TEDQuiz. It generates multiple–choice questions in two question types, gist-content questions and detail questions. To the best of our knowledge, there has been no prior work on our research topic. In future work, we plan to develop a wide variety of quiz types and implement the generation on a browser extension. Using this idea, English language learners can take tests to evaluate their understanding after watching any video.

ACKNOWLEDGMENT

This work was partially supported by National Science Council, Taiwan, with Grant No. 100-2221-E-001-015-MY3. We are also thankful to Chien-Ming Chen for implementing Web service and technique support; and Hui-Ping Liu for collecting experimental data.

TABLE III. QUESTIONNAIRE RESULTS

Item	**1**	**2**	**3**	**4**	**5**	**M**	**SD**
Questions generated by TEDQuiz were clear and understandable.	2.5	15	32.5	32.5	17.5	3.4	1.0
Questions generated by TEDQuiz were easy to answer.	2.5	20	32.5	32.5	12.5	3.3	1.0
Questions generated by TEDQuiz were relative with the TED Talk.	2.5	2.5	12.5	35	47.5	4.2	0.9
Questions generated by TEDQuiz were useful.	0	7.5	30	37.5	25	3.8	0.9

REFERENCES

[1] M. Agarwal and P. Mannem, "Automatic gap–fill question generation from text books," Proc. Workshop on Innovative Use of NLP for Building Educational Applications, pp. 56–64, 2011.

[2] S. Brin and L. Page. "The anatomy of a large-scale hypertextual Web search engine," Proc. Conference on World Wide Web, 107-117, 1998.

[3] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," Journal of Artificial Intelligence Research, Vol. 22, pp. 457-479, 2004.

[4] C. Y. Lin and E Hovy,. "The Automated Acquisition of Topic Signatures for Text Summarization," Proc. International Conference of Computational Linguistics, 2000.

[5] C. L. Liu, C. H. Wang, Z. M. Gao, and S. M. Huang, "Applications of lexical information for algorithmically composing multiple–choice cloze items," Proc. Workshop on Building Educational Applications Using Natural Language Processing, pp. 1–8, 2008.

[6] M. Heilman and N. A. Smith, "Question generation via overgenerating transformations and ranking," Technical report, Language Technologies Institute, Carnegie Mellon University Technical Report CMU–LTI–09–013, 2009.

[7] M. Heilman and N. A. Smith, "Good question! statistical ranking for question generation," Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 609–617, 2010.

[8] M. Heilman and N. A. Smith. "Extracting Simplified Statements for Factual Question Generation," Proc. Workshop on Question Generation, 2010.

[9] Y. T. Huang, M. C. Chen, and Y. L. Sun, "Personalized Automatic Quiz Generation Based on Proficiency Level Estimation," Proc. International Conference on Computers in Education, pp. 553-560, 2012.

[10] R. Mitkov, L. A. Ha, and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," Natural Language Engineering, Vol. 12, pp. 177-194, 2006

[11] J. Mostow and H. Jang, "Generating diagnostic multiple choice comprehension cloze questions," Proc. Workshop on Innovative Use of NLP for Building Educational Applications, pp. 136–146, 2012.

[12] V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan, "The First Question Generation Shared Task Evaluation Challenge," Proc. International Natural Language Generation Conference, 2010.

[13] V. Terzis and A. A. Economides, "The acceptance and use of computer based assessment," Computers & Education, Vol. 56, pp. 1032–1044, 2011.

[14] R. Trinder, Multimedia in the Business English Classroom: The Learners' Point of View. Computer Assisted Language Learning, Vol. 15, pp. 69-84, 2002

[15] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," Proc. European Conference on Machine Learning, pp. 491-502, 2001.

[16] Y. C. Yang, C. F. Yang, C. M. Chang, and J. S. Chang, "Computered–aid reading comprehension automatic quiz generation," Proc. Computational Linguistics and Speech Processing, 2005