Evaluation of Automatically Generated Pronoun Reference Questions

Arief Yudha Satria and **Takenobu Tokunaga**
School of Computing
Tokyo Institute of Technology
152-8550 Tokyo, Meguro, Ôokayama, 2-12-1, Japan
satria.a.aa@m.titech.ac.jp take@c.titech.ac.jp

Abstract

This study provides a detailed analysis of evaluation of English pronoun reference questions which are created automatically by machine. Pronoun reference questions are multiple choice questions that ask test takers to choose an antecedent of a target pronoun in a reading passage from four options. The evaluation was performed from two perspectives: the perspective of English teachers and that of English learners. Item analysis suggests that machine-generated questions achieve comparable quality with human-made questions. Correlation analysis revealed a strong correlation between the scores of machine-generated questions and that of human-made questions.

1 Introduction

Asking questions has been widely used as a method to assess the effectiveness of teaching and learning activities. By asking questions, teachers can get feedback whether students understand about the teaching materials. In this context, creating questions becomes an important task in teaching and learning activities. Questions are usually made by human experts, which demands manual efforts; thus it is time-consuming and expensive. Automatic question generation is a solution to solve this problem.

Several past studies worked on various kinds of automatic question generation. Heilman and Smith (2009) worked on the automatic question generation for the purpose of reading comprehension assessment and practice. Liu and Calvo (2012) worked on the automatic generation of trigger questions (directive and facilitative) for supporting writing activities. Chali and Hasan (2015) worked on the automatic generation of all possible questions given a topic of interest. Serban et al. (2016) worked on the automatic generation of questions about an image.

Research on automatic question generation has been active, yet there are few studies which elaborate the detailed evaluation process and in-depth analysis of the machine-generated questions. QG-STEC 2010 is the first shared task about question generation that comprises two subtasks: question generation from paragraphs and question generation from sentences (Rus et al., 2010). Human judges were utilised to evaluate question quality by considering five criteria: syntactic correctness and fluency, question type, relevance, ambiguity, and variety.

Liu and Calvo (2012) evaluated their trigger question generation system for academic writing support by comparing machine-generated trigger questions to human-made trigger questions based on five aspects: clarity, correctness, relevance, usefulness for learning concepts, and usefulness to improve the literature review documents. Twenty-three students were instructed to write essays and then to assess the trigger questions if these questions could improve their essays. Because the machine-generated trigger questions were created based on the collected student essays, their analysis showed that they were effective only for the collected student essays while the human-made trigger questions were effective for general essays as well as the collected essays.

Zhang and VanLehn (2016) employed students to rate machine-generated questions and human-made questions based on relevance, fluency, ambiguity, pedagogy and depth. Araki et al. (2016) evaluated their question generation system by judging the questions on three metrics: grammatical correctness, answer existence and inference steps.

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 76–85
Copenhagen, Denmark, September 8, 2017.

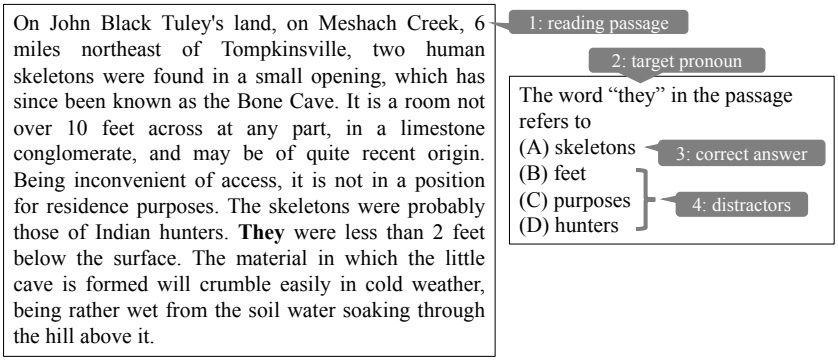

Figure 1: Example of pronoun reference question

Susanti et al. (2017) utilised English teachers and students to evaluate their question generation system. English teachers were asked to distinguish machine-generated questions from human-made questions apart. The English teachers also judged the questions on their usability in a real test and their difficulties using five scale rating. They also received suggestions to improve the questions from the English teachers. Furthermore, students were asked to answer the machine-generated questions and human-made questions; their answers were analysed using item analysis and the analysis based on Neural Test Theory (Shojima, 2007).

To sum up, the evaluation of automatic question generation systems in the past research was performed by utilising human judges and students. In this study, we provide detailed evaluation experiments and analysis of automatically generated pronoun reference questions. Pronoun reference questions consist of four components, i.e. a reading passage, a target pronoun, a correct answer, and three distractors as illustrated in Figure 1. We focus on pronoun reference questions because they measure the test taker's ability to resolve pronoun in reading passages. We argue that resolving pronoun is an important skill for reading comprehension.

The evaluation target of this study is the English pronoun reference questions generated by our system (Satria and Tokunaga, 2017). To the best of our knowledge, there is no other system for generating pronoun reference questions. The system generates questions from human-written texts by performing a sentence splitting technique on nonrestrictive relative clauses. The details of the question generation system are explained in Section 2. We evaluate the questions from two different perspectives following Susanti et al. (2017). The first perspective is from English teachers. We argue that English teachers have the ability to differentiate the good questions from the bad ones because creating questions is one of the teacher's responsibilities in the classroom; thus asking English teachers to judge the quality of machine-generated questions is reasonable. The second perspective is from English learners. Good questions can discriminate high proficiency learners from low proficiency learners. English learners were instructed to answer the questions and their responses were used for analysing the characteristics of the questions.

In what follows, we explain the automatic question generation system to be evaluated (Section 2), followed by the elaboration of the evaluation from the English teacher perspective (Section 3) and the English learner perspective (Section 4). We conclude the evaluation results and point out the possible future research direction (Section 5).

2 Generating pronoun reference questions

Pronoun reference questions such as in Figure 1 ask test takers to identify the antecedent of the target pronoun in the reading passage; thus the correct answer can be obtained by employing an anaphora resolution system to identify the antecedent of the target pronoun. Using this approach, the performance of the anaphora resolution system directly affects the quality of the generated questions. Since the performance of the state-of-the-arts anaphora resolution system is still insufficient to be employed for generating pronoun reference questions, we proposed to utilise nonrestrictive relative clauses to obtain pairs of the correct answer (antecedent) and the target pronoun (Satria and Tokunaga, 2017). The core idea

of our method is transforming a sentence with a nonrestrictive relative clause into two sentences by applying a sentence splitting technique with replacing the relative pronoun with a personal pronoun. An assumption behind our method is that the antecedent identification of relative pronouns is relatively easier than that of personal pronouns because the antecedents of the relative pronouns appear in a restricted region in the sentence.

The system receives human-written texts from Project Gutenberg[1] that span several genres (i.e. science, technology and history) and produces question components based on the texts. The question generation process comprises four steps: correct answer generation, reading passage generation, target pronoun generation, and distractor generation.

The nonrestrictive relative clause is vital in our system because we transform human-written texts by applying the sentence splitting technique regarding nonrestrictive relative clauses to create the correct answer, the reading passage and the target pronoun. Nonrestrictive relative clauses are clauses that do not specify its modifying noun; they only give additional information to it instead. Thus, they can be detached from their main clauses. This property allows the sentence splitting technique to work most of the cases without changing the meaning of the texts.

There are cases, however, where the sentence splitting induces a change of text meaning, mostly due to the introduced pronoun refers to a different antecedent from that referred to by the relative pronoun in the original sentence. For instance, the text (2) is derived from the text (1) by extracting the nonrestrictive relative clause (underlined part) and replacing the relative pronoun "which" with a pronoun "it". The antecedent of "it" in the third sentence looks to be "legend", a subject in the previous sentence. But it should be "knowledge" in the previous sentence when we look at the original sentence where "which", the counterpart of "it" in (2), obviously refers to "knowledge". To exclude such spurious anaphora, we apply the Centering theory (Brennan et al., 1987; Grosz et al., 1995) to see the introduced pronoun refers to the same antecedent as in the original sentence. In this particular example, the Centering theory tells us that "legend" in the second sentence of (2) has a higher status than "knowledge" because the former is a subject and the latter is an element in the prepositional phrase. Thus "legend" is a more probable antecedent of "it", which contradicts the original sentence of (1).

(1) The church of S. Croce has seen another strange death of a Pope, that of Sylvester II. (999-1003), a Frenchman, Gerbert by name. A legend, related first by cardinal Benno in 1099, describes him as deep in necromantic **knowledge**, which he had gathered during a journey through the Hispano-Arabic provinces.

(2) The church of S. Croce has seen another strange death of a Pope, that of Sylvester II. (999-1003), a Frenchman, Gerbert by name. A **legend**, related first by cardinal Benno in 1099, describes him as deep in necromantic **knowledge**. He had gathered **it** during a journey through the Hispano-Arabic provinces.

2.1 Correct answer generation

The identified antecedent of the relative pronoun is used as a correct answer. To identify the antecedent of the relative pronoun, we employed both lexical parser and dependency parser. The lexical parser produces a parse tree of the target sentence, i.e. a sentence that contains a nonrestrictive relative clause. The parse tree is traversed based on hand-made rules (Satria and Tokunaga, 2017) which consider the syntactic attachment and the linguistic feature, i.e. number. The dependency parser produces a set of dependencies which include the acl:relc[2] dependency relation. If only both results from the lexical parser together with hand-made rules and the dependency parser agree on the antecedent of the relative pronoun, the target sentence is further processed in the next steps. The system discards the target sentence which causes discordance on the antecedent of the relative pronoun.

2.2 Reading passage and target pronoun generation

We create a reading passage by splitting a sentence at a nonrestrictive relative clause. Sentence splitting divides the target sentence into two sentences: the main clause and the relative clause.

[1] https://www.gutenberg.org/

[2] http://universaldependencies.org/docs/en/dep/acl-relcl.html

Table 1: Example of the evaluation table filled by the evaluators

question	quality	reading passage	target pronoun	correct answer	distractors	comments
Q1	2	✓			✓	⋮
Q2	1			✓		⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Q60	3					⋮

When splitting the target sentence, the connection between two sentences must be maintained in order to retain the sentence meaning. The connection of those sentences is maintained through the target pronoun. The system creates the target pronoun by replacing the relative pronoun with a personal pronoun with considering linguistic features. Because the target pronoun resides in the reading passage, splitting target sentence and replacing the relative pronoun with the target pronoun complete the reading passage generation. For instance, the text (4) is derived from (3). The underlined non-restrictive relative clause in (3) is taken out into a separate sentence and placed after the main clause in (4). At the same time, the relative pronoun in the relative clause is replaced with the personal pronoun "they". We further confirm that the introduced pronoun "they" surely refers to the subject in the previous sentence regarding the Centering theory.

(3) The **flowers**, which are individually larger than those of the False Acacia, are of a beautiful rosy-pink, and produced in June and July.

(4) The **flowers** are of a beautiful rosy-pink, and produced in June and July. **They** are individually larger than those of the False Acacia.

2.3 Distractor generation

Distractor generation comprises the following three steps.

Candidate generation Since we restrict the antecedent of the pronoun, i.e. the correct answer, to a noun or a noun phrase, distractors must also be nouns or noun phrases. The part-of-speech tagger was employed to extract all nouns and noun phrases in the passage. The incompatible candidates on linguistic features are eliminated from the distractor candidates.

Coreference chain extraction A coreference chain consists of a list of expressions that refer to the same entity in a text. Thus, expressions in the same coreference chain with the correct answer are also a possible correct answer. Therefore, they are eliminated from the distractor candidates.

Candidate ranking Since we need only three distractors, the distractor candidates are ranked on the recency principle. Recently mentioned entities are likely to be maintained in human memory because they are still fresh; thus those entities are likely to be referred to by pronouns. More recently mentioned entities are ranked higher than the less recently mentioned entities. Finally, the three highest ranked candidates are selected as the distractors.

3 Evaluation from English teacher perspective

3.1 Experimental setting

We asked five English teachers[3] to evaluate the quality of 60 machine-generated questions by assigning a score of one, two or three to each question. The meaning of the scores is described below.

1. **problematic**, the question is not usable in a real test. Significant modifications are necessary for real use.

2. **acceptable but can be improved**, the question is usable in a real test as it is, but it can be further improved.

3. **acceptable**, the question has no problem to be used in a real test without any change.

If the question quality is judged to be one or two, the evaluators must further identify the problematic question components by checking

[3]They are non-native English speakers but the TESOL (http://www.tesol.org) certificate holders.

the corresponding columns as shown in Table 1. The evaluators leave the problematic components columns empty for acceptable quality questions. The evaluators may optionally give comments on problematic components or suggestions to improve the question quality.

Table 2: Distribution of pairwise disagreement

evaluator\score	{1,2}	{1,3}	{2,3}
(A, B)	7	4	28
(A, C)	2	7	14
(A, D)	4	6	24
(A, E)	2	8	20
(B, C)	1	1	28
(B, D)	1	0	28
(B, E)	3	4	30
(C, D)	1	0	27
(C, E)	4	3	22
(D, E)	1	5	19
total	26	38	240

Table 3: Distribution of rating

evaluator\score	1	2	3	total
A	10	18	32	60
B	1	35	24	60
C	1	20	39	60
D	0	18	42	60
E	6	16	38	60
total	18	107	175	300

3.2 Result and discussion

First, we investigated the agreement between the evaluators by computing the ordinal Krippendorff's alpha (Krippendorff, 1970); it was 0.05 indicating very low agreement between the evaluators. We further investigated the reason of the low agreement. We calculated the pairwise disagreement frequency between every pair of the evaluators as shown in Table 2. The table indicates that the disagreement between the judgement "acceptable but can be improved" and "acceptable" ($\{2, 3\}$) is dominant (80%). This fact suggests the decision on these two categories is highly subjective. Since they are both acceptable categories, we recalculated the Krippendorff's alpha after merging them into a single category to obtain the value 0.06. The average of the pairwise observation agreement was 0.89 after merging. Table 3 shows the distribution of scores judged by each evaluator. As the table shows, the highly skewed distribution of judgment can be considered as the main reason of a very low alpha despite the fairly high observation agreement.

Table 4: Majority quality scores of 60 questions

majority score	frequency
1	0
2	12
3	39
tie	9
total	60

Table 4 shows the distribution of the quality score calculated by the majority principle. The majority principle means that when at least three evaluators rate a same value, that particular value is defined as the question quality score. Table 5 indicates that there are 39 questions (65%) which the majority of the evaluators rated "acceptable (3)". All nine tie cases get at most two "problematic" rating, i.e. the "problematic" can not be the majority. This means all generated questions were judged "usable in a real test" based on the majority principle.

Table 5: Average quality scores of 60 questions

average score	frequency
1.6	1
1.8	1
2.0	4
2.2	8
2.4	7
2.6	22
2.8	13
3.0	4
total	60

Table 5 summarises the average quality scores of five evaluators with their frequency. Even though the majority quality is the same, the actual rating may be different; thus it yields a different average quality. The question with the score 1.6 gets two ones and three twos. All evaluators agree that this particular question has an error in the correct answer. The question with the score 1.8 gets two ones, two twos and one three. Four evaluators agree that this particular question has an error in the correct answer.

Table 6 summarises the comments from the five evaluators with their frequency. The most common comments are related to the correct answer. This tendency is consistent with the component-wise evaluation of our past research (Satria and

Table 6: Evaluator's comments with frequency

comments	frequency
other option could be the correct answer	71
the reading passage is too long	28
the distractors do not distract	18
the distractors are too distracting	11
the reading passage offers little context	6
there are multiple correct answers	5
the reading passage has many technical word (i.e. too difficult)	4
the correct answer is too obvious	1
the target pronoun is inadequate	1

Tokunaga, 2017). We counted the number of questions with a checked cell in the "correct answer" column of the evaluation table (Table 1) to find 80 such cells in total. This number is roughly the same as that of the comments on correct answers. Among these 80 questions, 12 questions were rated 1 (problematic) and 68 were rated 2 (acceptable but can be improved). These cases suggest that the filtering with the Centering theory should be further improved.

4 Evaluation based on English learner perspective

The evaluation from the English learner perspective was conducted to evaluate the behaviour of machine-generated questions in measuring test taker's proficiency.

4.1 Experimental setting

We prepared three sets of questions each of which contains ten machine-generated questions (MGQs) and ten human-made questions (HMQs), in total 20 questions. These 30 HMQs were randomly selected from TOEFL preparation books while these 30 MGQs were randomly selected from the set of MGQs which were judged acceptable on the majority principle in the evaluation by the English teachers as described in Section 3. The question sets were created so that the difference of the average of question difficulty across the question sets was minimised. The balance of question difficulty among three groups, and between MGQs and HMQs is important because we calculate the student-wise score correlation between scores from MGQs and HMQs as explained later in 4.2.

To balance question difficulty among the question sets, we utilised the reading passage difficulty. A question is considered difficult if its reading passage is difficult and vice versa. The reading passage difficulty is calculated based on the word difficulty in the passages. We employed JACET8000 (Uemura and Ishikawa, 2004), a list of 8,000 English words divided into eight levels of word difficulty based on their word frequency. Level 1 is the most frequent (i.e. the easiest) while level 8 is least frequent (i.e. the most difficult). Words that do not appear in the list are considered even less frequent than level 8; thus they are considered to be level 9. To obtain the reading passage difficulty, we assigned a JACET8000 word difficulty level to every word in the reading passage as illustrated in Figure 2 and calculated the average of the difficulty levels. The average of reading passage difficulty for each question set is presented in Table 7.

Dr.$_1$ M. Aurel$_9$ Stein$_9$, principal$_2$ of$_1$ the$_1$ Oriental$_7$ College$_1$ at$_1$ Lahore$_9$, has$_1$ now$_1$ ready$_1$ for$_1$ publication$_4$ the$_1$ first$_1$ volume$_2$ of$_1$ his$_1$ critical$_3$ edition$_4$ of$_1$ the$_1$ Rajatarangini$_9$, or$_1$ Chronicles$_8$ of$_1$ the$_1$ Kings$_1$ of$_1$ Kashmir$_9$, upon$_1$ which$_1$ he$_1$ has$_1$ been$_1$ engaged$_3$ for$_1$ some$_1$ years$_1$. This$_1$ work$_1$ is$_1$ of$_1$ special$_1$ interest$_1$ as$_1$ being$_1$ almost$_1$ the$_1$ sole$_4$ example$_1$ of$_1$ historical$_2$ literature$_2$ in$_1$ Sanskrit$_9$. It$_1$ was$_1$ written$_2$ by$_1$ the$_1$ poet$_2$ Kalhana$_9$ in$_1$ the$_1$ middle$_1$ of$_1$ the$_1$ twelfth$_1$ century$_1$.

Figure 2: Example of reading passage with word difficulty level (subscripts correspond to the level)

Table 7: Mean of reading passage difficulty

metric	question set	MGQ	HMQ
average JACET8000	Qs1	2.15	2.14
	Qs2	2.13	2.13
	Qs3	2.12	2.12
Flesch-Kincaid grade level	Qs1	9.9	11.2
	Qs2	10.0	9.8
	Qs3	9.6	10.7
Flesch-Kincaid reading ease	Qs1	60.3	46.1
	Qs2	59.9	58.9
	Qs3	65.2	50.5
Dale-Chall readability formula	Qs1	9.0	9.9
	Qs2	9.0	9.1
	Qs3	8.9	9.7

Many metrics to measure text readability have been proposed in the past, such as Flesch-Kincaid grade level (Kincaid et al., 1975), Flesch-Kincaid reading ease (Kincaid et al., 1975) and Dale-Chall readability formula (Dale and Chall, 1948). The first two calculate text difficulty with respect to the number of sentences, words and syllables in the text. The third one takes into account the difficulty of each word as well. Table 7 also shows

Table 8: TOEIC score of each group

student group	question set	TOEIC score mean	TOEIC score SD	number of students
1	Qs1	561	146	31
2	Qs2	559	123	25
3	Qs3	554	122	25

the mean values of these metrics for each question set and generation mode, i.e. machine-generated vs. human-made. Overall, the difficulty of reading passages in every question set is well balanced against every metric.

Eighty-one Japanese university students (57 first year and 24 second year students) were recruited and divided into three groups, 27 students for each group, considering their TOEIC scores; we did our best to minimise the difference of the score distribution and the mean of the scores across these three groups. Each student group was assigned a different question set and instructed to finish the assigned question set within 30 minutes.

4.2 Result and discussion

Although we made three groups of the same number of students (27) and assigned a different question set to each group, four students mistakenly worked on a wrong question set. Therefore the distribution of the number of students in a group was skewed as shown in Table 8. Table 8 also shows the average TOEIC score of each group with a standard deviation (SD).

Table 9: Item difficulty of MGQs and HMQs

	MGQ	HMQ
mean	0.59	0.60
standard deviation	0.24	0.17
minimum	0.20	0.26
maximum	0.96	0.90

The item analysis investigates the test taker's responses to individual question items to evaluate the quality of those items. It often uses two measures: the item difficulty and the item discrimination index. The item difficulty is a proportion of the number of test takers who answered correctly to the number of all test takers (Brown, 2013). The value ranges from 0 to 1 with a larger value representing an easier item. Table 9 shows the descriptive statistics of the item difficulty of the sets of 30 MGQs and 30 HMQs.

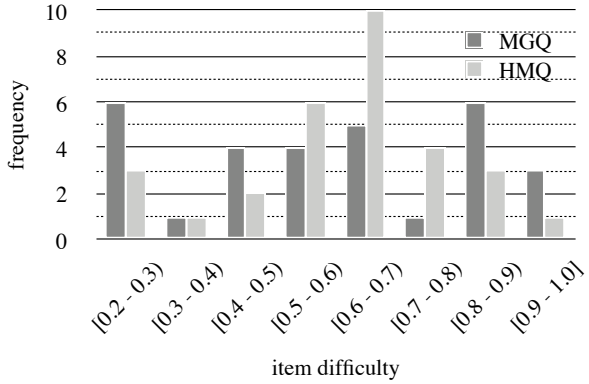

Figure 3: Distribution of item difficulty

Table 9 shows no big difference in mean of the item difficulty between MGQs and HMQs. This result suggests that MGQs have similar difficulty with HMQs. This is consistent with the fact we maintained the balance of question difficulty between MGQs and HMQs as explained in Subsection 4.1. We also provide the distribution of the item difficulty of the MGQs and HMQs in Figure 3. Although the mean is similar between the MGQs and HMQs as shown in Table 9, Figure 3 reveals that the distribution of the item difficulty for HMQs is closer to the normal distribution than that for MGQs. We conducted the Levene's test (Levene et al., 1960) to assess the item difficulty variance homogeneity between MGQs and HMQs to find that their variances are not homogeneous. As we do not care about controlling item difficulty when generating question items, this is a natural consequence.

Mexico, 1818. This species, though not hardy enough for every situation, is yet sufficiently so to stand unharmed as a wall plant. It grows from 10 feet to 12 feet high, with deep-green leaves that are hoary on the under sides. The flowers are bright blue, and produced in June and the following months. **They** are borne in large, axillary panicles. In a light, dry soil and sunny position this shrub does well as a wall plant, for which purpose it is one of the most ornamental. There are several good nursery forms, of which the following are amongst the best: C. azureus Albert Pettitt, C. azureus albidus, C. azureus Arnddii, one of the best, C. azureus Gloire de Versailles, and C. azureus Marie Simon.
(A) leaves
(B) sides
(C) flowers ← correct answer
(D) months

Figure 4: The easiest question

Figure 4 shows the easiest question item while

There are two recesses in the cliff on the opposite side of the little creek formed by the spring. They are 40 to 50 feet above the water, each with an irregular floor of 20 by 30 feet under shelter of the rock. No solid rock is visible in front of them, but a projecting ledge appears on either side about 6 feet below the present average level of the floor; and this is probably the depth of accumulation at the front. **It** seems continuous. It may be less toward the rear. The cavities are in a stratum which is somewhat shelly and crumbles easily.
(A) ledge ← correct answer
(B) depth
(C) accumulation
(D) front

Figure 5: The most difficult question

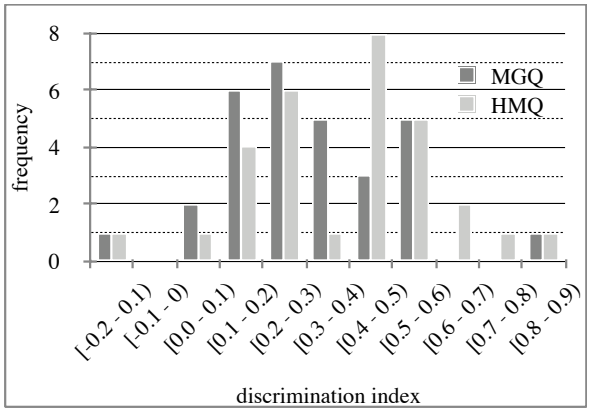

Figure 6: Distribution of item discrimination index

Figure 5 shows the most difficult one in the MGQs in which the target pronoun is in bold and the options are underlined in the reading passage for the readability purpose. Twenty-four out of 25 students answered correctly for the easiest one. This question item is easy because the subject pronoun refers to the subject of the previous sentence. Only five out of 25 students answered correctly for the most difficult question item. Both extremes are not preferable in measuring test taker's proficiency because too easy items lead to very high scores while too difficult items lead to very low scores for the most of test takers.

We calculated the Pearson correlation coefficient between the JACET8000 based reading passage difficulty as we defined in Table 7 and the item difficulty of the MGQs and obtained the value of 0.56. This result suggests that the reading passage difficulty can be one of the important factors for predicting and controlling the item difficulty of question items.

The item discrimination index is a metric to measure the discrimination power of question items (Brown, 2013). The discrimination power is the ability of question items in discriminating high-proficiency test takers from low-proficiency test takers. This metric is vital for language testing because a good test must be able to discriminate test taker's proficiency precisely. The item discrimination index of a question item *i* is computed as follows

$$ID_i = \frac{U_i - L_i}{n},$$

where U_i and L_i represent the number of test takers who correctly answered the question item *i* in the high proficiency group and the low proficiency group respectively, and n denotes the number of test takers in a group. The groups of high and low proficiency are defined as the top 27% of the test takers and bottom 27% of the test takers respectively. The threshold value of 27% is utilised to maximise two characteristics; those two groups must be as different as possible to discriminate clearly, and the number of test takers in each group must be as large as possible to achieve reliability (Popham, 1981; Kelley, 1939).

We computed the item discrimination index for each question item and the average of them. The average is 0.33 for the MGQs and 0.37 for the HMQs. A question item is considered to be acceptable if its discrimination index is greater than or equal to 0.2 (Brown, 1983). According to this criteria, we counted the number of question items of which the discrimination index is greater than or equal to 0.2. Out of 30 question items, the 22 MGQs and 24 HMQs items cleared this condition. Figure 6 shows the distribution of the discrimination index. There seems to be no big difference between the MGQs and HMQs in terms of

The region may be roughly characterized as a vast sandy plain, arid in the extreme; or rather as two such plains, separated by a chain of mountains running northwest and southeast. In the southern part of the reservation this mountain range is known as the Choiskai mountains, and here the top is flat and mesa-like in character, dotted with little lakes and covered with giant pines. **They** in the summer give it a park-like aspect. The general elevation of this plateau is a little less than 9,000 feet above the sea and about 3,000 feet above the valleys or plains east and west of it.
(A) plains
(B) mountains
(C) lakes
(D) pines ← correct answer

Figure 7: MGQ example with a poor discrimination index ($ID = 0.125$)

the average discrimination index (0.33 vs. 0.37) and the number of items clearing the 0.2 criterion (22 vs. 24). Their distribution reveals that the HMQs shows a slightly better distribution than the MGQs. However, the MGQs have comparable discrimination power as the HMQs.

Figure 7 shows an example of MGQ which has a poor discrimination index, i.e. $ID = 0.125$. Three test takers in the high proficiency group and two test takers in the low proficiency group answered correctly. The distractor "mountains" distracted test takers in the high proficiency group very much; thus the number of correctly answered test takers was almost the same between the two groups. The potential reason is that "mountains" appears twice in the text, so it lured the test takers to choose "mountains".

To assess the ability of the MGQs in measuring test taker's proficiency, we calculated the correlation between the test taker's score of the MGQs and other scores including that of the HMQs and TOEIC scores. We argue that the test taker's TOEIC scores provide their true English proficiency. The Pearson correlation coefficient (Pearson, 1896) was calculated, presented in Table 10. The p-value of all the correlation coefficients is less than 0.05.

Table 10 shows that there is no big difference between the MGQs and HMQs in terms of the correlation between the test taker's scores and their TOEIC scores. Furthermore, the correlation with the TOEIC Reading scores is stronger than that with the TOEIC Listening scores. This is a reasonable tendency because the pronoun reference questions are designed for assessing reading comprehension ability.

Table 10: Peason correlation coefficients between test taker's scores

	MGQ	HMQ
TOEIC Listening	0.56	0.57
TOEIC Reading	0.65	0.68
TOEIC Listening & Reading	0.74	0.77
HMQ	0.61	—

5 Conclusion

This paper presented the evaluation of automatically generated pronoun reference questions which ask test takers the antecedent of the specified pronoun in the reading passage. A pronoun reference question was automatically generated by splitting a sentence in a human-written text at a nonrestrictive relative clause and replacing the relative pronoun with a personal pronoun.

The evaluation was performed from two different perspectives: the English teacher perspective and the English learner perspective. Automatically generated 60 question items were evaluated by five English teachers, resulting in that 39 out of 60 (65%) question items were considered acceptable to be used in a real test. We administered 30 MGQs from these acceptable question items together with 30 HMQs from TOEFL preparation books to the 81 university students. The analysis results of the test taker's responses showed that the MGQs achieved comparable quality with the HMQs on their item difficulty and item discrimination index. Furthermore, there was a strong correlation between the MGQ scores and the TOEIC scores of the same test takers.

Possible future work includes controlling item difficulty of the generated questions and generating other types of questions. For instance, our experimental result suggested that the item difficulty of the generated questions had a moderate correlation with the reading passage difficulty. Thus, controlling the passage difficulty might enable us to control the difficulty of the question items. We also need to further explore other factors affecting the item difficulty.

References

Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1125–1136. http://aclweb.org/anthology/C16-1107.

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stanford, California, USA, pages 155–162. https://doi.org/10.3115/981175.981197.

Frederick Gramm Brown. 1983. *Principles of educational and psychological testing*. Holt, Rinehart, and Winston, 3 edition.

James Dean Brown. 2013. Classical test theory. In Glenn Fulcher and Fred Davidson, editors, *The*

Routledge handbook of language testing, Routledge, pages 323–335.

Yllias Chali and Sadid A. Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics* 41(1):1–20. https://doi.org/10.1162/COLI_a_00206.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin* pages 37–54. http://www.jstor.org/stable/1473669.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203–225. http://aclweb.org/anthology/J95-2003.

Michael Heilman and Noah A Smith. 2009. Question generation via overgenerating transformations and ranking. Technical report, DTIC Document.

Truman L Kelley. 1939. The selection of upper and lower groups for the validation of test items. *Journal of educational psychology* 30(1):17–24. http://dx.doi.org/10.1037/h0057123.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement* 30(1):61–70. https://doi.org/10.1177/001316447003000105.

Howard Levene et al. 1960. Robust tests for equality of variances. *Contributions to probability and statistics* 1:278–292.

Ming Liu and Rafael A. Calvo. 2012. Using information extraction to generate trigger questions for academic writing support. In Stefano A. Cerri, William J. Clancey, Giorgos Papadourakis, and Kitty Panourgia, editors, *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 358–367.

Karl Pearson. 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 187:253–318. http://www.jstor.org/stable/90707.

W.J. Popham. 1981. *Modern educational measurement*. Englewood Cliff, NJ: Prentice-Hall.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics, pages 251–257. http://www.aclweb.org/anthology/W10-4234.

Arief Yudha Satria and Takenobu Tokunaga. 2017. Automatic generation of english reference question by utilising nonrestrictive relative clause. In *Proceedings of the 9th International Conference on Computer Supported Education - Volume 1: CSEDU*. INSTICC, ScitePress, pages 379–386. https://doi.org/10.5220/0006320203790386.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 588–598. http://www.aclweb.org/anthology/P16-1056.

K Shojima. 2007. Neural test theory. *DNC Research Note* 7(02):1–12.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2017. Evaluation of automatically generated english vocabulary questions. *Research and Practice in Technology Enhanced Learning* 12(1):11. https://doi.org/10.1186/s41039-017-0051-y.

Toshihiko Uemura and Shinichiro Ishikawa. 2004. Jacet 8000 and asia tefl vocabulary initiative. *Journal of Asia TEFL* 1(1):333–347.

Lishan Zhang and Kurt VanLehn. 2016. How do machine-generated questions compare to human-generated questions? *Research and Practice in Technology Enhanced Learning* 11(1):7. https://doi.org/10.1186/s41039-016-0031-7.