Automatic Generation of English Reference Question by Utilising Nonrestrictive Relative Clause

Arief Yudha Satria and Takenobu Tokunaga
Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan

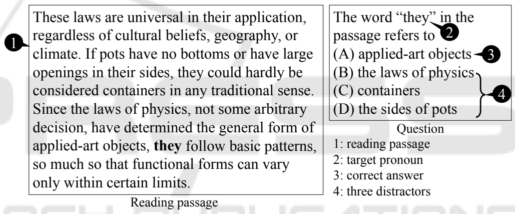
Keywords: Automatic Question Generation, English Reference Test, TOEFL Reference Question.

Abstract: This study presents a novel method of automatic English reference question generation. The English reference question is a multiple choice question which is comprised of a reading passage, a target pronoun, a correct answer (an antecedent of the target pronoun), and three distractors. The reading passage is generated by transforming human made passages using the proposed sentence splitting technique on nonrestrictive relative clauses (NRC). The correct answer is generated by the analysis of parse trees of the passage. The distractors are extracted from the reading passage by using a part-of-speech (POS) tagger and a coreference resolver. Human evaluation showed that 53% of the generated questions were acceptable.

1 INTRODUCTION

Answering questions is one of effective methods to learn a subject. By answering questions, learners could get feedback to what extent they understand about the subject (Davis, 2009). Questions are usually generated by human experts. Creating those questions needs high skill, and is time consuming. The past research showed that human experts in question generation could be replaced with machines. For example, Susanti et al. (2015) attempted to create multiple choice questions for vocabulary assessment while Karamanis et al. (2006) generated questions in the medical domain. Similarly, Skalban et al. (2012) made questions for multimedia-based learning whereas Heilman (2011) constructed WH factual questions. Automatic question generation has been an active research area, particularly in the language learning domain.

From a viewpoint of question consumers, the main resource of question exercises for language learners is usually language learning preparation books, past examination papers and student workbooks and so on. The amount of resources available to accommodate learner needs for practising is still limited. This fact motivates us to leverage abundant number of human made texts for making questions automatically. The present research would contribute to reducing teacher burden on creating those questions and accommodating learners with abundant question exercises as well. Similar to those aforementioned studies, this study works on automatically generating questions, in particular questions for language learning. We focus on reference questions which has been introduced in TOEFL iBT[1].



These laws are universal in their application, regardless of cultural beliefs, geography, or climate. If pots have no bottoms or have large openings in their sides, they could hardly be considered containers in any traditional sense. Since the laws of physics, not some arbitrary decision, have determined the general form of applied-art objects, **they** follow basic patterns, so much so that functional forms can vary only within certain limits.

The word “they” in the passage refers to
(A) applied-art objects
(B) the laws of physics
(C) containers
(D) the sides of pots

1: reading passage
2: target pronoun
3: correct answer
4: three distractors

Figure 1: TOEFL reference question example.

A reference question has four components: (1) a reading passage, (2) a target pronoun, (3) a correct answer, and (4) three distractors as illustrated in Figure 1. The reference question asks test takers to choose the correct antecedent of the target pronoun in the reading passage from the four choices. Although the reference question appears at most twice out of 12 to 14 for each reading passage in TOEFL iBT, we argue that the reference question is crucial to measure test takers’ ability in reading comprehension. When people read a passage and find a pronoun in it, they will naturally resolve the pronoun, i.e. they find its antecedent (Gordon and Scearce, 1995). Therefore, asking the antecedent of a pronoun could evaluate the test takers’ ability in understanding the reading passage.

To the best of our knowledge, there is no research in the past which focuses on generating the reference

[1] https://www.ets.org/toefl/ibt/about

Satria, A. and Tokunaga, T.
Automatic Generation of English Reference Question by Utilising Nonrestrictive Relative Clause.
DOI: 10.5220/0006320203790386
In *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017) - Volume 1*, pages 379-386
ISBN: 978-989-758-239-4

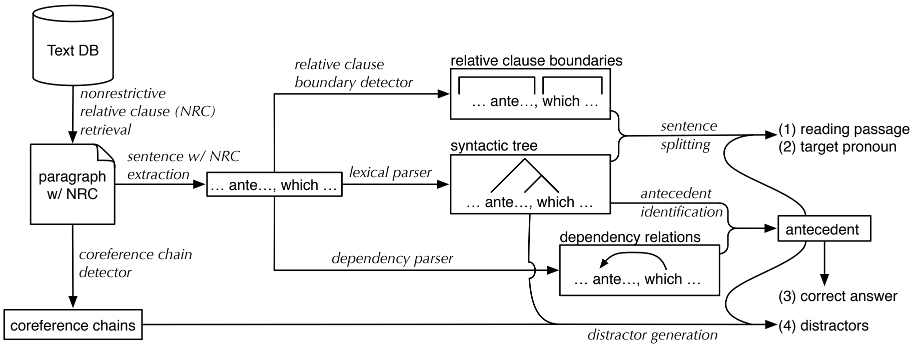


Figure 2: Overview of reference question generation.

question. One possible approach to automatically generate reference questions is by applying a coreference resolver to the reading passage. The coreference resolver could find every pair of a pronoun and its antecedent in the reading passage. The system then will choose a pair of a pronoun and its antecedent as the target pronoun and the correct answer respectively. In this approach, however, the quality of the correct answer will heavily rely on the performance of the coreference resolver. Unfortunately the performance of the currently available coreference resolver remains still unsatisfactory for using our purpose (Lee et al., 2013).

We take a different approach to utilise nonrestrictive relative clauses (NRC) as the resource for pronoun and antecedent pairs. The key idea is splitting a sentence at the relative pronoun and replacing it with an appropriate pronoun to have the pronoun and antecedent pair, assuming that it is easier to identify the antecedent of a relative pronoun within a sentence than that of a pronoun across the sentences. The details of sentence splitting will be explained in section 4.

2 SYSTEM OVERVIEW

Figure 2 illustrates the overview of our system. The system first retrieves human made paragraphs that contain at least one NRC from text databases. The NRCs are retrieved by using simple pattern matching, i.e. by checking if the paragraph includes the string ", which". The sentence with a NRC is extracted from the paragraph to be processed by the lexical parser and the dependency parser to perform *antecedent identification*, and then by the relative clause boundary detector and the lexical parser to perform *sentence splitting*. The antecedent identification must be performed first because the antecedent will be used in sentence splitting to generate the reading passage and the target pronoun, and will be used in *distractor generation*. The antecedent then will be the correct answer. After the reading passage, the target pronoun, and the correct answer are generated, then the coreference chains are identified in the human made paragraph for generating the distractors. Finally, the question could be made by filling the question template with four question components. During the course of the above process, the retrieved paragraph in the first step will be discarded immediately when it is found to generate inappropriate question. Since we have plenty of paragraphs, we take high accuracy at the expense of high recall.

In the overall process, using NRCs for generating reference questions is the key idea of our proposal. A NRC is a relative clause which provides additional information to its modifying noun phrase[2]. It does not define its modifying noun as opposed to the restrictive relative clause. Hence, removing the NRC from the matrix sentence will retain the sentence essential meaning. Consider the following two sentences.

(1) The drill, which is whirled between the palms of the hands, consists of a stalk perhaps a quarter of an inch in diameter.

(2) The drill which is whirled between the palms of the hands consists of a stalk perhaps a quarter of an inch in diameter.

(1) contains a NRC while (2) contains a restrictive relative clause. The difference is located in the existence of a comma before the "which". In (1), the NRC is used to indicate that there is only one drill in the discourse; the relative clause provides additional information about the drill. On the other hand, (2) uses the restrictive relative clause to specify that the speaker is talking about the whirled drill. Removing the relative clause in (1) retains the core meaning of the sentence but this is not the case for (2). Therefore, we can safely transform a sentence with a NRC in a paragraph into two sentences with keeping the meaning of

[2] https://en.oxforddictionaries.com/grammar/relative-clauses

the overall paragraph. At the same time, we obtain the target pronoun by replacing the relative pronoun with an appropriate pronoun in the sentence derived from the relative clause.

3 CORRECT ANSWER GENERATION

In order to generate the correct answer, our system finds the antecedent of the relative pronoun. The antecedent identification must be performed first because the antecedent plays an important role in the other two subprocesses: sentence splitting and distractor generation. We accomplish the antecedent identification by searching the parse tree of the sentence including a NRC, and by employing dependency parser. The following subsections describe the antecedent identification by parse tree search, that by dependency parser, and the aggregation of the both results.

3.1 Parse Tree Search

A parse tree could be utilised to find the boundary of the NRC and to locate the antecedent (noun phrase) which the NRC is attached to. The original sentence is parsed using Stanford Lexical Parser (Klein and Manning, 2003), and the resultant parse tree is traversed to find the antecedent of the NRC by using the following heuristics:

1. Find the nearest noun phrase which is left sister of the NRC in the resultant parse tree.
2. If no such noun phrase is found, find the nearest noun phrase parent of the NRC.

Tregex (Levy and Andrew, 2006), a tool for querying tree data structures, is used in order to locate the antecedent in the resultant parse tree based on the foregoing heuristics.

3.2 Dependency Parser

A dependency parser could also be employed to find the antecedent of NRCs. By using Stanford Dependency Parser (Chen and Manning, 2014), our system will find the attachment of the NRC. Figure 3 shows an example of the dependency parser output. Among many dependencies, the antecedent could be identified by finding a dependency labelled with “`acl:relc`”[3]. This dependency connects the modified noun and the modifying relative clause by a directed arc starting from the antecedent to the relative clause. Note that we omitted the labels of other dependencies for simplicity in the figure.

3.3 Result Aggregation

Both parse tree search and dependency parser provide the antecedent of the NRC. Our system conjoins their results to vote on the correct answer. Only if both results agree on the same antecedent, we adopt the suggested antecedent as the correct answer of the question, then move to the succeeding processes. When the both results do not agree, the analysed sentence is simply discarded and other retrieved paragraph will be tried.

4 READING PASSAGE GENERATION

The reading passage in the reference questions of TOEFL is created by changing college-level textbooks as little as possible to assess how well test takers could understand academic text (Service, 2012). The reading passage in our system is generated from text databases, which are free e-books by Project Gutenberg[4] spanning across several genres: science, history, and technology. The process of reading passage generation comprises three steps: NRC detection, sentence splitting and target pronoun creation.

4.1 NRC Detection

To extract the NRC from the sentence, our system employs the Stanford Lexical Parser (Klein and Manning, 2003) and the relative clause boundary detection algorithm (Siddharthan, 2002). Since detecting NRC boundaries is not always easy, both methods are employed to obtain the reliable relative clause boundaries. As in the correct answer generation in the previous section, only if both agree on the boundary of the NRC, the extracted NRC is processed further, otherwise the paragraph itself is discarded.

4.2 Sentence Splitting

Sentence splitting divides the original sentence into the matrix clause and the relative clause. There are two possibilities in ordering these two clauses when generating the reading passage. Consider the following example.

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

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

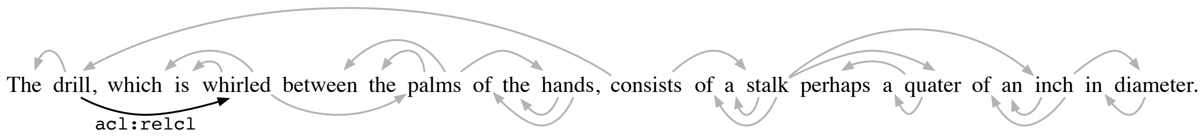

The drill, which is whirled between the palms of the hands, consists of a stalk perhaps a quater of an inch in diameter.
acl:relcl

Figure 3: Example of dependency parser output.

(3) The drill consists of a stalk perhaps a quarter of an inch in diameter. The drill is whirled between the palms of the hands.

(3) is derived from (1) by moving the NRC after the matrix clause and substituting “which” with its antecedent, “the drill”. There is another way to arrange sentences as in (4).

(4) The drill is whirled between the palms of the hands. The drill consists of a stalk perhaps a quarter of an inch in diameter.

The decision in sentence ordering could be made by considering the number of distractor candidates that appear before the pronoun.

4.3 Target Pronoun Creation

The target pronoun of the reference question is derived from the relative pronoun in the original sentence.

(5) The drill consists of a stalk perhaps a quarter of an inch in diameter. It is whirled between the palms of the hands.

(5) is derived from (3) by substituting “the drill” with “it”. This substitution introduces a new pronoun to be the target pronoun of the question. When creating the target pronoun, morpho-syntactic features, e.g. the number, gender, and animacy are considered so that they agree with that of the antecedent. In (5), “the drill” is a singular inanimate noun, thus the pronoun would be “it”.

(6) We know that the temple was built as early as the time of TJsertsen, for in it have been found one or two of his blocks; and no doubt the original **shrine**, which was rebuilt in the time of Philip Arrhidseus, was of the same period, but hitherto no remains of the centuries between his time and that of Hatshepsu had been found.

(7) We know that the **temple** was built as early as the time of TJsertsen, for in it have been found one or two of his blocks; and no doubt the original shrine was of the same period, but hitherto no remains of the centuries between his time and that of Hatshepsu had been found. **It** was rebuilt in the time of Philip Arrhidseus.

There are cases in which the introduced pronoun is prone to be misinterpreted in term of its antecedent. In (7) derived from (6), humans tend to interpret “temple” as the antecedent of the introduced “it” since this interpretation retains the topic across the two sentences by sharing the subject, although it is actually “shrine” in the original sentence (6). Such transformation that distorts the meaning of the original sentence should be avoided. To this end we employ the Centering theory (Brennan et al., 1987; Grosz et al., 1995) to confirm that the introduced pronoun naturally refers to the original antecedent. The Centering theory models natural topic transition in the discourse, thus that could be the fundamentals for predicting the antecedent of pronouns. The Centering theory assigns preference to four types of topic transitions in the order of CONTINUE, RETAIN, SMOOTH-SHIFT and ROUGH-SHIFT. The CONTINUE transition is the most natural and the ROUGH-SHIFT is the least. These topic transitions are determined in terms of the relation between a pronoun and its antecedent. We adopt the pronoun and antecedent pairs that constitute transitions except for ROUGH-SHIFT. The reason why we adopt less natural transitions is to diversify the generated questions.

5 DISTRACTOR GENERATION

Distractors are plausible but incorrect options which distract test takers from selecting the correct answer. We propose the following conditions that must be fulfilled by distractor candidates.

1. The POS of the distractor candidate is either singular noun (NN), proper noun (NNP) or plural noun (NNS) because we deal with only the case where the antecedent of the pronoun is a noun.
2. The distractor candidate has the same number, gender, and animacy features as the correct answer. For instance, a singular distractor would be too obvious for a plural correct answer.

3. The distractor candidate should precede the target pronoun. There are two kinds of reference phenomena: anaphora and cataphora. Anaphora refers back to the preceding word, while cataphora refers forward to the following word. Our system deals with only anaphora because cataphora is rare in normal texts.
4. The distractor candidate should not belong to the same coreference chain as the correct answer. Any word belonging to the same coreference chain as the correct answer would also be the correct answer, thus it must be avoided to be a distractor candidate.

To fulfil the first condition, our system uses Stanford POS Tagger (Toutanova et al., 2003) to extract all nouns for distractor candidates. After all the nouns are extracted, they are checked for their number, gender, and animacy. To avoid creating easy distractors, incompatible noun phrases are discarded to fulfil the second condition. The third condition is achieved by selecting only nouns that appears before the target pronoun. The order of derived sentences in the sentence splitting process is determined so that more distractor candidates are obtained.

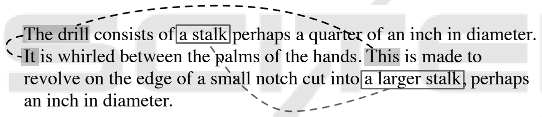

The drill consists of a stalk perhaps a quarter of an inch in diameter. It is whirled between the palms of the hands. This is made to revolve on the edge of a small notch cut into a larger stalk, perhaps an inch in diameter.

Figure 4: Example of coreference chains.

Error sources in distractor generation.
The coreference chain is a set of words which refer to the same entity in the discourse. As the reference question asks the antecedent of the target pronoun, the distractors must not refer to the same entity as the correct answer refers to. In order to fulfil the fourth condition, Stanford Coreference Resolver (Lee et al., 2013) is used to extract all coreference chains in the reading passage. It receives the reading passage as its input and provides all coreference chains as its output. After the coreference chain of the correct answer is determined, the distractor candidates in the same chain are discarded. Figure 4 illustrates an example of coreference chains where the expressions referring to the same entity are linked by dashed lines.

The drill consists of a $\underline{\text{stalk}}_{(3)}$ perhaps a quarter of an $\underline{\text{inch}}_{(2)}$ in $\underline{\text{diameter}}_{(1)}$. **It** is whirled between the palms of the hands. This is made to revolve on the edge of a small notch cut into a larger stalk, perhaps an inch in diameter.

Figure 5: Recency-based distractor candidate ranking.

We need at least three distractors to generate a reference question. When more than three distractor candidates remain after filtering with the four conditions, we need to choose three out of them. It is well known that recently mentioned entity is likely to be referred to by a pronoun (Walker et al., 1998). Keeping this in mind, our system ranks the distractor candidates based on their distance from the target pronoun. The closest distractor candidate is ranked at the highest position as illustrated in Figure 5. Our system then chooses three highest ranked distractor candidates. In the figure, the distractor candidates are underlined with their rank in the parentheses. “Diameter”, “inch” and “stalk” are adopted as distractors in this example.

6 EVALUATION

We evaluate the proposed method in component by component. As a source of texts, we used the texts from the Project Gutenberg. The evaluation are conducted based on the subjective judgement of the author. The following subsections describe the evaluation of the components: correct answer generation, reading passage generation and distractor generation.

6.1 Correct Answer Generation

The correct answer generation was evaluated by measuring the accuracy of antecedent identification. In order to evaluate to what extent the parse tree search and the dependency parser could find the antecedent of the relative pronoun, we retrieved 100 paragraphs from the Gutenberg texts, and applied the parse tree search and the dependency parser to these paragraphs.

Table 1: Number of correctly identified antecedents.

method	correct antecedents
parse tree search	50
dependency parser	51

Table 1 represents the number of correctly identified antecedents in the 100 paragraphs in each method. The numbers are almost the same between the two methods, which agreed on the antecedent in 58 cases, and 41 of them were correct. Therefore aggregating the results from the two methods increases the accuracy from 50% to 70%.
The wrong 17 cases happened when the relative clause came after a noun phrase with a prepositional phrase, as in (8). The both methods happened to attach the relative clause to the object of the preposition, “Abydos” in (8), instead of the correct antecedent, “the oldest temple”.

(8) Petrie has traced out the plan of **the oldest temple** of Osiris at *Abydos*, which may be of the time of Khufu, from scanty evidences which give us but little information.

Another error is subject verb disagreement error; a singular antecedent was chosen for a plural relative pronoun and vice versa. For instance, our system detected “place” as the antecedent of the relative pronoun “which” in (9). Since the predicate of the relative clause is “are”, the correct antecedent must be “junipers and cedars”. There are two cases found in the output of parse tree search.

(9) Going downward they merge into piñons, useful for firewood but valueless as timber, and these in turn give *place* to **junipers and cedars**, which are found everywhere throughout the foothills and on the high mesa lands.

To confirm the effectiveness of the result aggregation, we conducted an additional experiment in which the paragraphs were repeatedly retrieved from the Gutenberg texts and tested if the results of the parse tree search and the dependency parser were agreed on the same antecedent, until the number of the accumulated paragraphs reached 100. Since the paragraph was adopted into the pool only if the both results agreed, it is guaranteed that the both methods agreed on the antecedent of the NRC in the accumulated 100 paragraphs. Among these 100 agreed antecedents, 70 were correct ones and 30 were not. As a result of these experiments, we could say that we can obtain 70% in accuracy for identifying the antecedents of NRCs by applying the result aggregation. As we adopt the antecedent as a correct answer, the performance of the correct answer generation is 70% in accuracy.

6.2 Reading Passage Generation

The reading passage generation was evaluated by measuring the performance of sentence splitting. The main source of errors in the sentence splitting process is the wrongly identified relative clause boundaries. In order to evaluate to what extent the lexical parser and the relative clause boundary detector could find the boundary of relative clause for the purpose of reading passage generation, we retrieved 100 paragraphs randomly from the Gutenberg texts and analysed them by the lexical parser and the relative clause boundary detector.

Table 2 shows the number of correctly identified relative clause boundaries in the 100 paragraphs in each method. The relative clause boundary detector shows around 10% higher accuracy than the lexical parser. In the 100 paragraphs, the both methods agreed on the same 70 boundaries, 67 out of which were correct boundaries. As in the antecedent identification described in the previous subsection, we could gain 8% and 17% in accuracy respectively by aggregating the results from the different methods.

Table 2: Number of correctly identified relative clause boundaries.

method	correct boundaries
lexical parser	79
relative clause boundary detector	88

Similarly to the antecedent identification evaluation, to confirm the aggregation effectiveness, we conducted an additional experiment which accumulated the agreed 100 results by both methods. The result showed that in the 97 out 100 agreed cases the identified boundaries were correct ones. We could say that we can obtain 97% in accuracy for identifying the correct boundary of relative clauses.

Additionally, we evaluated the sentence splitting performance with regard to the extent to which the target pronoun can be interpreted as referring to the correct antecedent. We generated 50 passages with the CONTINUE transition and 50 passages with the SMOOTH-SHIFT transition to ensure the diversity of the generated questions[5].

Table 3: Number of correctly interpretable pronouns.

transition type	pronouns
CONTINUE	48
SMOOTH-SHIFT	41

Table 3 shows the number of correctly interpretable pronouns. The table indicates that the CONTINUE transition gives more appropriate pronouns than SMOOTH-SHIFT because CONTINUE is more natural than SMOOTH-SHIFT. This result suggests that by adopting CONTINUE, our system could generate 48 appropriate pronouns out of 50 questions. However, if we generate all questions by using the CONTINUE transition only, the generated questions will be homogeneous thus the correct answer could be guessed easily.

6.3 Distractor Generation

As we can see from Figure 2, errors in the sentence splitting and the antecedent identification affect the

[5] We did not adopt passages with the RETAIN transition because they tend to be similar passages as those with the CONTINUE transition because of its definition.

performance of the distractor generation. To evaluate the distractor generation independently from these preceding steps, we first collected 100 questions that are free from errors in these preceding steps. The distractor generation was applied to the correctly generated reading passages and pronouns, and the correctly identified antecedents. The generated distractors were judged valid if all three distractors fulfilled the conditions described in section 5.

Table 4: Error sources in distractor generation.

error source	errors
coreference chain detection	6
feature agreement	5

We had 11 cases in which at least one generated distractor was invalid. The main sources of these errors are categorised into two: errors in coreference chain detection, and agreement errors in features of pronouns and their antecedent. The distribution of error sources are summarised in Table 4.

A knowledge of the various Semitic alphabets is necessary for copying inscriptions. Unless the traveler be also acquainted with the languages he had better be cautious about copying Semitic inscriptions; without such knowledge he runs the risk of confusing different Semitic letters. **They** often closely resemble one another. He should, however, be able to make squeezes and photographs.
The word "they" in the passage refers to
(A) letters
(B) inscriptions
(C) languages
(D) alphabets

Figure 6: Example of invalid distractors due to the coreference chain detection error.

Figure 6 shows an example of invalid distractors caused by the coreference chain detection error. The coreference chain detector failed to capture "alphabets" and "letters" in the same coreference chain. Due to this error, (D) alphabets was generated as one of the distractors even though it belongs to the same coreference chain as the correct answer (A) letters.

Some errors occurred in feature agreement. When a person entity was not recognised as a person, it would be chosen as the distractor for the target pronoun "it" as shown in Figure 7. The "Hammurabi" was not recognised as a person, thus it was chosen as one of the distractor. It is obvious that a person could not be a distractor for inanimate correct answer.

We have in total 11 generated questions with at least one inappropriate distractor, suggesting that our system successfully generated distractors with 89% in accuracy.

6.4 Overall Evaluation

Since a question consists of a reading passage, a target pronoun, a correct answer and distractors, the question could be considered acceptable only if all of those components are error-free. We generated 100 questions fully automatically and counted errors in each component. Table 5 shows the number of errors in each question component in the 100 questions.

Table 5: Errors in the components of 100 questions.

question component	#error
reading passage	3
target pronoun	9
correct answer	30
distractors	6

Since a single question might contains errors in multiple components, this categorisation is not disjoint. The most common errors happened in correct answer generation because finding the antecedent of relative clauses is a difficult task. The target pronoun tended to be wrong when it appeared as the object argument in the sentence. In all nine cases of the wrong target pronoun, the pronoun were the object of verbs. From 100 questions, 53 of them are error-free, suggesting that we have 53% in accuracy for automatically generating reference questions.

It is true that Gudea, the Sumerian patesi of Shirpurla, records that Hammurabi rebuilt the temple of the goddess Ninni (Ishtar) at a place called Nina. Now Nina may very probably be identified with Nineveh, but many writers have taken it to be a place in Southern Babylonia and possibly a district of Shirpurla itself. No such uncertainty attaches to Hammurabi's reference to Nineveh. **It** is undoubtedly the Assyrian city of that name.

The word "it" in the passage refers to
(A) Nineveh
(B) reference
(C) Hammurabi
(D) uncertainty

Figure 7: Example of invalid distractors due to agreement error.

7 CONCLUSION

In this paper, we presented a novel method to automatically generate reference questions for evaluating test taker's reading comprehension ability. A reference question consists of four components: a reading passage, a target pronoun, a correct answer and three distractors. It asks test takers the antecedent of the tar

-get pronoun in the reading passage. Questions were automatically generated from existing human made paragraphs so we believe that the present study contributes to reducing teacher burden on creating those questions and accommodating learners with abundant question exercises as well.

In our proposed method, nonrestrictive relative clauses (NRC) were utilised to generate the reading passage using the sentence splitting technique. For generating correct answers, the parse tree search and the dependency parser worked together to enhance the reliability of the generated answer. In creating distractors, the coreference resolver and the part-of-speech tagger were employed.

According to the subjective evaluation of the generated questions, 53% of the questions were acceptable. That means the half of the generated questions could be used for the real test, but the performance still remains far from fully automatic question generation. Our system generated reading passage with 97% in accuracy, correct answer with 70% in accuracy, and distractors with 89% in accuracy; those results show promising potential to generate the reference questions automatically. With the current performance as is, it will be practical to incorporate the proposed components into a kind of authoring system for creating reference questions, so as to reduce the burden of human experts in creating questions.

At the same time, we need to further refine each component generation module to obtain a better performance of the total system. For instance, in order to improve the performance of correct answer generation, checking the agreement between the relative clause and the antecedent could help antecedent identification perform better. In order to remedy the animacy error in distractor generation illustrated in Figure 7, we could incorporate a named entity recogniser in our system to distinguish a person from an inanimate entity. We are also planning to conduct experiments on real English learners to see to what extent the automatically generated questions by our method could discriminate test takers' ability in reading comprehension.

REFERENCES

Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 155–162. Association for Computational Linguistics.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Davis, B. G. (2009). *Tools for teaching*. John Wiley & Sons.

Gordon, P. C. and Scearce, K. A. (1995). Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory & Cognition*, 23(3):313–323.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Heilman, M. (2011). *Automatic factual question generation from text*. PhD thesis, Carnegie Mellon University.

Karamanis, N., Ha, L. A., and Mitkov, R. (2006). Generating multiple-choice test items from medical text: A pilot study. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 111–113. Association for Computational Linguistics.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Levy, R. and Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234. ELRA.

Service, E. T. (2012). *Official Guide to the TOEFL Test, 4th Edition*. McGraw-Hill Education.

Siddharthan, A. (2002). Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL02)*, pages 60–65.

Skalban, Y., Ha, L. A., Specia, L., and Mitkov, R. (2012). Automatic question generation in multimedia-based learning. In *Proceedings of COLING 2012: Posters*, pages 1151–1160, Mumbai, India. The COLING 2012 Organizing Committee.

Susanti, Y., Iida, R., and Tokunaga, T. (2015). Automatic generation of english vocabulary tests. In *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)*, pages 77–78.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Walker, M. A., Joshi, A. K., and Prince, E. F. (1998). *Centering theory in discourse*. Oxford University Press.