Towards Generating Multiple-Choice Tests for Supporting Extensive Reading

Shinjiro Okaku
Department of Informatics
Kyushu University
Fukuoka, Japan
ohkaku@nlp.inf.kyushu-u.ac.jp

Yoichi Tomiura
Department of Informatics
Kyushu University
Fukuoka, Japan
tom@inf.kyushu-u.ac.jp

Emi Ishita
Department of Library Science
Kyushu University
Fukuoka, Japan
ishita.emi.982@m.kyushu-u.ac.jp

Shosaku Tanaka
College of Letters
Ritsumeikan University
Kyoto, Japan
sho@lt.ritsumei.ac.jp

***Abstract*— We propose a method for generating multiple-choice test for an English text selected by a learner and its answer, that are used to make a self-assessment whether the learner comprehends the text after reading it. In our method, the system extracts several important sentences from the text, and replaces one word in each of these sentences with its synonym (if possible). One of these sentences is then selected as a correct optional sentence, while further changes to the polarities or nouns in the remaining sentences are carried out to generate distractor optional sentences for the multiple-choice test. Our method has potential to make extensive reading in English more effective.**

Keywords; Aided Learning; Important Sentence; Paraphrase; Documents on Web; Extensive Reading.

I. INTRODUCTION

Reading extensive English texts is a good training for English learners [1]. They should read texts that are interesting and with an appropriate level to keep up and enhance the learning effect of extensive reading [2]. There are a huge number of English texts available through the Internet that could support such reading.

Learners, whose English abilities are not high, often do not comprehend the content or story of a text after reading it, even if they understood each individual sentence while they were reading it. To acquire a practical English reading ability, they must train their reading comprehension. Learners can do this training effectively when they use a text with accompanying comprehension test as reading material. At present, there is not much reading material like that on the Internet. Therefore, we have developed a method for generating a multiple-choice comprehension test for an English text selected by a learner, using a Natural Language Processing technology. In this paper, we describe how to create a multiple-choice comprehension test for a text and the evaluation for the method.

The rest of this paper is organized as follows. We explain the outline of our system for generating multiple-choice tests and introduce related works in Section 2. The method for extracting important sentences to generate optional sentences for the test is introduced in Section 3, and the method for paraphrasing them is introduced in Section 4. In Section 5 and Section 6, we present the experiments for extracting important sentences and paraphrasing sentences, and the results we obtained. Our final conclusions and future work are discussed in Section 7.

II. OUTLINE OF OUR SYSTEM AND RELATED WORK

The test consists of one sentence (the correct optional sentence) that is consistent with the English text, and several sentences (distractor optional sentences) that are inconsistent with the text. A learner selects a sentence consistent with the text from among optional sentences after reading the text. The system generates these optional sentences from the text selected by the learner as follows. First, it extracts important sentences from the text. We regard "important sentences" as sentences that we have to retain (even temporarily) to understand the content of the text. Second, the system changes elements in these extracted sentences, without changing their meaning, to ensure the test is not dependent on simply memorizing content. Finally, the system selects one sentence among them as the correct optional sentence, and carries out further changes to the polarities, subject or object nouns on the remaining important sentences to generate distractor optional sentences. Thus, the basic techniques for generating optional sentences in the test require a process to extract important sentences, and to change expressions in them without changing their meaning.

In automatic summarization, there are a number of methods to extract important parts or sentences and shorten sentences without changing their meaning [3]. However, state-of-the-art methods for automatic summarization are using machine learning. Therefore we need extensive training data of texts and their summaries to generate test using automatic summarization techniques for an English text selected by a learner among diverse texts available on the Internet. It is difficult and impractical to prepare such training data. There also are many studies on paraphrasing, although there is no free software or resource currently available for public use except the PPDB [4], as far as the

 ISBN: 978-1-61208-385-8

authors know. Therefore, we attempted to generate a multiple-choice test using a method of extracting important sentences without training data and a method of paraphrasing based on a thesaurus.

There are a few studies on a system that generate questions about texts suitable for the proficiency level of English learners, analyze their responses, and give advice to lead them to a correct answer, after reading or listening to a text [5][6]. However, this system is designed to improve grammar and vocabulary for low-level English learners at junior high school in Japan. The purpose of our test is to support extensive reading in English for middle to high-level English learners who want to acquire a practical reading ability.

III. METHOD FOR EXTRACTING IMPORTANT SENTENCES

We extract important sentences in a text using the degree of importance of words based on the *Spreading Activation Model*. Matsumura et al. [7] proposed a method for extracting keywords based on this model. This method extracts the words expressing assertions of a document, taking into account the structure of the document (i.e., the strength of the relationship between words for each segment of the document). It is not dependent upon machine learning. Term frequency-inverse document frequency (TF-IDF) also has been used to extract important words [8] [9], and is not based on machine learning either. However, TF-IDF is not a measure of the importance of words that reflects the structure of a document. Hence, we believe the method proposed in [7] works better than TF-IDF to extract sentences we have to retain to understand the content of a text.

Here, we briefly explain the method for calculating the importance of words proposed in [7]. First, they divide a text into segments, and extract M most frequent words in a segment S_t. Let w_1, w_2, ⋯, and w_L be extracted words from the whole text. The co-occurrence frequency of each pair of extracted words in the segment S_t is used to calculate $R(t)$, the spreading activation matrix of the segment S_t. Thus, the (i, j)-element of $R(t)$ is the strength of the relationship between the word w_i and the word w_j in the segment S_t. The (i, j)-element of $R(t)$ is zero when w_i or w_j does not appear in the segment S_t. Let $\mathbf{a}(t)$ be the L-dimensional column vector, whose i-th element is the activity value of the word w_i in the segment S_t. The activity values of words are calculated according to

$$\mathbf{a}(t) = \big((1-\gamma)I + \alpha R(t)\big)\mathbf{a}(t-1),$$

where all elements in $\mathbf{a}(0)$ are 1. The parameter α is transmission rate, and γ is attenuation rate. Now suppose that the last segment number in a given document is n, then $\mathbf{a}(n)$ expresses the activity values of words after reading the document. Thus, the i-th element in $\mathbf{a}(n)$ expresses the degree of importance of the word w_i in that document. Matsumura et al. proposed the *sharp activity value* as another measure of a word's importance. It is the activity value of a word divided by the activity times of the word.

We conducted a preliminary experiment where only activity value or only sharp activity value was used as measure of a word's importance and importance of a sentence was calculated based on word's importance. However the precision was not so good. In this research, we use the following *mixo-activity value* as a new measure of a word's importance. The mixo-activity value m_w is defined as follows:

$$m_w = \max(a'_w, s'_w),$$

where

$$a'_w = \frac{a_w}{\max_w a_w}, \qquad s'_w = \frac{s_w}{\max_w s_w},$$

a_w is the activity value of the word w, and s_w is the sharp activity value of the word w. The ranges of activity values and sharp activity values are generally different. Then, in the definition of m_w, we use a'_w and s'_w that are transformed so that the maximum values are the same (equal to 1).

We define three measures of importance of a sentence using the degrees of importance of its words as follows (n is the number of words in the sentence) :

(1) The sum of the mixo-activity values of the words in the sentence.

(2) The value (1) divided by n.

(3) The value (1) divided by $\log(n+1)$.

The degree of importance of a sentence according to (1) tends to be high, when the sentence has many words. Therefore, we try to use the measure of (2), the mean mixo-activity value of words in the sentence. However, long sentences are often important. Then we also try to use (3), which will have a value intermediate between (1) and (2).

Thus, we propose three different measures of the importance of sentence using mixo-activity values as the measure of importance of words. These are Wm_i (i = 1, 2, 3), respectively. The number "i" in Wm_i corresponds to the measures listed above. We use all three measures to determine which is the best measure to extract the sentence that we have to retain to understand the content of the text.

IV. METHOD FOR PARAPHRASING SENTENCES

We propose the method of replacing one word in a sentence with its synonym, as a simple method for paraphrasing sentences. However, we cannot simply replace a word with one its synonym because a word generally has many synonyms with different meanings. We have to select an appropriate one among these synonyms, so that the paraphrased sentence has the same meaning as the original sentence.

We focus on a transitive verb and a head noun in its object noun phrase in a sentence. We replace these with their synonyms, if they exist. A transitive verb and its object noun phrase are generally strongly connected to each other. The strength of this connection can be estimated using point-wise mutual information (PMI). Let v be a transitive verb, np be its object noun phrase in a sentence, and n be the head noun of np. In addition, let v' be one of the synonyms of v, n' be

one of the synonyms of *n*, and *np'* is what we get by replacing *n* in *np* with *n'*. For example, one of the synonyms of "provide" is "supply", while one of the synonyms of "example" is "instance"; hence, we get "supplies a perfect example" and "provides a perfect instance" from the original phrase "provides a perfect example". If there is a strong connection between *v'* and *np*, then "*v' np*" is likely a natural expression. Hence, "*v' np*" more likely has the same meaning as "*v np*" because *v'* is one of the synonyms of *v*. Also, if there is a strong connection between *v* and *np'*, then "*v np'* " also will be a natural expression, likely to have the same meaning as "*v np*".

Therefore, we generate the paraphrased sentence of the original sentence *s* as follows:

(1) Extract a transitive verb *v* and its object noun phrase *np* in *s*. Let *n* be a head noun of *np*. If *s* does not have a transitive verb, we do not generate a paraphrased sentence for *s*.

(2) Find the synonyms of *v* and the synonyms of *n* using a thesaurus. We denote the synonyms of *v* as $v'_1, v'_2, \cdots, v'_J$, while the synonyms of *n* are $n'_1, n'_2, \cdots, n'_K$. We get np'_k by replacing *n* in *np* with n'_k. If neither *v* nor *n* has synonyms, we do not generate a paraphrased sentence for *s*.

(3) Calculate the strength of the connection between v'_j and *np* ($j = 1, 2, \cdots, J$), and the connection between *v* and np'_k ($k = 1, 2, \cdots, K$). We denote the strength of the connection between *A* and *B* as *Score*(*A*, *B*).

(4) Find the following set C_V:

$$C_V = \{v'_j \mid Score(v'_j, np) \geq \theta,\ j = 1, 2, \cdots, J\},$$

where θ is a chosen threshold value.
Also, find the following set C_{NP}:

$$C_{NP} = \{np'_k \mid Score(v, np'_k) \geq \theta, k = 1, 2, \cdots, K\}$$

If both C_V and C_{NP} are empty sets, we do not generate a paraphrased sentence for *s*. Otherwise, find the following words *v** and *np**.

$$v^* = \arg\max_{v'_j \in C_V} Score(v'_j, np),$$

$$np^* = \arg\max_{np'_k \in C_{NP}} Score(v, np'_k).$$

If *Score*(*v**, *np*) > *Score*(*v*, *np**), the paraphrased sentence of *s* is what we get by replacing *v* with *v** in *s*. While, if *Score*(*v**, *np*) ≤ *Score*(*v*, *np**), the paraphrased sentence of *s* is what we get by replacing *np* with *np** in *s*.

As described earlier, the strength of the connection between sentence elements can be estimated using the PMI. The PMI between *v* and *np* are defined as follows:

$$PMI(v, np) = \log \frac{P_{V,NP}(v, np)}{P_V(v) P_{NP}(np)}, \quad (1)$$

where

$$P_{V,NP}(v, np) = \frac{f_{V,NP}(v, np)}{\sum_{v,np} f_{V,NP}(v, np)},$$

$$P_V(v) = \sum_{np} P_{V,NP}(v, np),$$

$$P_{NP}(np) = \sum_{v} P_{V,NP}(v, np).$$

$f_{V, NP}(v, np)$ is the frequency of co-occurrence of *v* as a transitive verb and *np* as *v*'s object noun phrase in a corpus.

PMI defined by (1) is not reliable when $f_V(v)$ or $f_{NP}(np)$ is small, because in this case the statistical fluctuation of PMI is large. To consider this, we define H_0 that is threshold value for a word, and we select the appropriate synonym in step (4) that have $f_V(v)$ and $f_{NP}(np)$ higher than H_0.

In this research, we use Wikipedia [10] as a corpus to calculate $f_V(v)$, $f_{NP}(np)$, and $f_{V, NP}(v, np)$. We prepare the body text data of Wikipedia, and parse it to count the frequency of co-occurrence of *v* as a transitive verb and *np* as *v*'s object noun phrase.

V. EXPERIMENT ON EXTRACTING IMPORTANT SENTENCES

We defined the three measures of the importance of sentence, Wm_i (i = 1, 2, 3) in Section III. Next, we need to evaluate which measure works best to extract "the sentence that we have to retain to understand the content of the text" in an experiment.

A. Evaluation Data

First, we chose randomly twenty English texts with about 1,500 words from "The Free Library" [11]. We also chose three test subjects (S, H, P; S is a graduate school student, H is a research student, and P is a professional translator) to have them read these texts and extract five important sentences from each text. We explained to the test subjects that "important sentence" means the sentence that we have to retain (even temporarily) to understand the content of the text. We labeled these extracted sentences as *important* and the remaining sentences as *unimportant* for each subject, and calculated the κ statistic of their inter-subject agreement. Table 1 shows these results.

Generally, there is moderate agreement when $0.4 < \kappa \leq 0.6$, good agreement when $0.6 < \kappa \leq 0.8$, and nearly perfect agreement when $0.8 < \kappa$. Table 1 shows that there is low agreement between subjects, suggesting that it is very difficult to select five important sentences from each text with on average about 57(1130/20) sentences. There also are clear individual differences between subjects. Hence, we assume that any sentence extracted as an important sentence by at least one test subject is important. Using this criterion, we generated our evaluation data, where each sentence was labeled as important or unimportant. We used these data to evaluate the performance of the proposed method below.

 ISBN: 978-1-61208-385-8

TABLE I. κ STATISTIC BETWEEN TEST SUBJECTS

	S	H	P
S		0.273	0.286
H	0.273		0.273
P	0.286	0.273	

TABLE II. PRECISION FOR EXTRACTING IMPORTANT SENTENCES BY EACH MEASURE

Measure	Wm_1	Wm_2	Wm_3
Precision	0.41	0.43	0.47

B. Experiment and Result

We carried out a morphological analysis of each sentence in the text, and removed stop words, using the stop word list in the SMART system [12]. We extracted the basic form of a word with the software tool Tree Tagger [13].

Next, we carried out spreading activation, and extracted five important sentences using the three measures: Wm_i (i = 1, 2, 3) for each text. We set M, the number of words extracted from each segment, to 20% of the number of words by type in the segment S_t and also set a parameter α, transmission rate, to 1.0. We tried some parameter-setting for a parameter γ, attenuation rate. We performed parameter sweep across 0.1, 0.3, 0.5, 0.8 and 1.0. We evaluated the precision of each measure using our evaluation data. Table 2 shows our results when γ = 0.3, that is the best result.

C. Discussion

The precision of extracting important sentences using the three different measures ranges from 41 to 47%. Given that the average number of sentences in these texts is 57, and the average number of important sentences is 10.6, then these measures work fairly well. We calculated the κ statistic between the judgment by the measure Wm_3 and the judgment by each test subject. Table 3 shows the result. Comparing Table 3 with Table 1, we found that these κ statistics are not so low. From this point, we also think that the measure Wm_3 works well.

However, the precision of 47% is not sufficient to use in a practical system to generate a multiple-choice test and its answers, if the evaluation method is appropriate.

As we described in the subsection "Evaluation Data", we had three test subjects extract five important sentences for each text. Then, we assumed that the sentences that we had to retain (even temporarily) to understand the content of the text were the sentences extracted by at least one test subject as important sentences. That is to say, we regarded that the sentence that no test subjects selected as important sentence was not the sentence that we had to retain (even temporarily) to understand the content of the text. However, the sentence that no test subjects selected as important sentence is not always inappropriate for the sentence used in the multiple-choice comprehension test. In the sentences that were selected by the system and were evaluated as unimportant based on the evaluation data, there would be sentences that we had to retain even temporarily to understand the content of the text. The evaluation criterion in this experiment might be too strict, and we have to change the evaluation method, for example, increasing the number of the test subjects.

TABLE III. κ STATISTIC BETWEEN THE JUDGMENT BY THE MEASURE AND THE JUDGMENT BY EACH TEST SUBJECT

	S	H	P
Wm_3	0.197	0.183	0.183

VI. EXPERIMENT ON PARAPHRASING SENTENCES

We proposed a method to paraphrase a sentence by replacing a transitive verb or a head noun in its object noun phrase with its synonym selected according to the strength of their connection, where the strength of the connection between a transitive verb and its object noun phrase is estimated by PMI between them. In this section, we evaluate the performance of the proposed method.

A. Evaluation Data and Tools

We collected articles from "The Free Library", and manually extracted one hundred fifty pairs of transitive verbs and their object noun phrases. When we extracted noun phrases, we removed adverbs, prepositional phrases, and relative clauses to extract noun phrases with the structure "(DETERMINER) (ADJECTIVE) NOUN". We generated candidates of paraphrased expressions for each of these extracted pairs of transitive verb and its object noun phrase using WordNet [14] as a thesaurus to find synonyms for the transitive verbs and nouns. We generated 10 candidates on average for each original expression.

Next, we asked an English editing company to evaluate these candidates and classify them as one of the following four categories:

- Natural and similar meaning
- Unnatural and similar meaning
- Natural and different meaning
- Unnatural and different meaning

Only expressions evaluated as "Natural and similar meaning" are acceptable as paraphrases.

B. Evaluation Method

We define the precision P of the method, as the ratio of the number of the acceptable expressions generated by this method to the total number of the expressions generated by the method. We define the gain G of the method, as the ratio of the number of the expressions generated by the method to the number of all test pairs (150). Parameters in the proposed method are thresholds θ and H_0. When we set θ

(and H_0) to high values, the selected expressions tend to be correct. However, in this case, the method does not generate a paraphrase for many of the original expressions. When we consider our purpose, it is not necessary to paraphrase all of the extracted important sentences. It is sufficient to appropriately paraphrase only one important sentence for a given text and select this as the correct optional sentence, so that the test is not just a simple memorizing test. If we generate a multiple-choice test composed of five optional sentences, then a G value in the range 10 to 40% is sufficient. However, the precision P must be nearly 100%.

In our experiment, we set the goal gain to 10, 20, 30, and 40%, and varied the threshold θ and H_0 for each measure, and find the values of θ and H_0 so that the gain G for the training data is equal to or higher than the goal gain, and the precision P is as high as possible. With these threshold values, we seek the precision and the gain for the test data. We evaluated this procedure using a 5-fold cross validation. In this procedure, the thresholds were varied as follows:

$$\theta = 1, 2, \cdots, 50$$
$$H_0 = 1,\ 10,\ 100,\ 1000,\ 10000$$

C. Results

Table 4 shows the result. We think the precision around 80% at best is fairly good, even though it is simple method, and not based on machine learning. However, the precision must be nearly 100% to generate a multiple-choice test. We discuss further possible improvements to our method in the next section.

D. Discussion

We have to improve our method of paraphrasing sentences to generate a multiple-choice test. At present, most of the unacceptable expressions selected by our proposed method were labeled as "Natural and different meaning" by a proofreader at an English editing company.

By replacing a transitive verb or a head object noun in a verb phrase with its synonym, we expected that we could get a verb phrase with the same meaning as the original verb phrase. In many cases, this is true, if what we get by replacing these words is a natural expression. However, there were many exceptions, as our results show. One of our issues is that we have to select an appropriate expression among expressions estimated as natural using a measure of the strength of the connection. There are studies on disambiguating word sense using the distribution of words around a target word. We think there is potential to select an appropriate expression using this technique. We expect that the candidate expression estimated as natural more likely has the similar meaning to the target verb phrase if it has the similar word-distribution around it to the word-distribution around the target verb phrase.

TABLE IV. GAIN AND PRECISION FOR PARAPHRASING BY 5-FOLD CROSS VALIDATION

θ	H_0	G	P
20	1	40.00%	72.07%
23	1	32.67%	75.31%
28	1	18.67%	77.27%
34	10	10.00%	81.82%

VII. CONCLUSION

We proposed a method for extracting important sentences and paraphrasing them to generate a multiple-choice test for an English text. This test is used to make a self-assessment whether a learner comprehends the text after reading it, and would make extensive reading in English more effective. We evaluated the proposed method with a small-scale experiment and were able to show the potential of our proposed method. Unfortunately, the performance of extracting important sentences was insufficient to form the basis of a practical system to generate a multiple-choice test. The evaluation criterion might be too strict in this evaluation. We have to change the evaluation method. The performance of paraphrasing was insufficient, too. We would carry out further improvements for paraphrasing sentences in our future work. The PPDB [4] that is the large corpus for paraphrasing sentences has been released since 2013. We are going to try an improvement of paraphrasing using the PPDB.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 24652122.

REFERENCES

[1] S. Inagaki and T.Inagaki, "Extensive reading does make a difference: Further evidence from university-level English language education in Japan", language and culture 2010, pp.49-53, 2010 (in Japanese).

[2] S.Abe, "A Trial Extensive Reading Course in English", Artes liberales Vol.42, pp. 99-105, 1988 (in Japanese).

[3] M. Okumura and E. Nanba, "Science of knowledge Auto Summary", Ohmsha, 2005 (in Japanese).

[4] J. Ganitkevitch, B. V. Durme, and C. C. Burch."PPDB: The Paraphrase Database". NAACL-HLT 2013, pp.758-764. 2013.

[5] H. Kunichika, M. Urushima, T. Hirashima, and A. Takauchi, "A Definition of Complexity of Questions for Question and Answer and its Evaluation," Transactions of the Japanese Society for Artificial Intelligence, Vol. 17, pp. 521-529, 2002 (in Japanese).

[6] H. Kunichika, M. Honda, T. Hirashima, and A. Takauchi, "A Method of Judging Answers by Comparing Semantic Information for Questions and Answer," IEICE D-I, Vol. J88-D-I, pp.25-35, 2005 (in Japanese).

[7] N. Matsumura, Y. Ohasawa, and M. Ishizuka, "Automatic Indexing Based on Term Activity," Transactions of the Japanese Society for Artificial Intelligence, Vol. 17, No.4, pp. 398-406, 2002 (in Japanese).

[8] A. Nenkova and K. McKeown, "Automatic Summarization," Foundations and Trends Information Retrieval, Vol.5, No.2-3, pp.103-233, 2011.

[9] R. Iida and T. Tokunaga, "Salient Word Extraction in Discourse and its Application," IPSJ SIG Technical Report 2009-NL-193, No. 9, pp.1234-1241, 2009 (in Japanese).

 ISBN: 978-1-61208-385-8

[10] http://en.wikipedia.org/ [retrieved: May, 2014]

[11] http://www.thefreelibrary.com/ [retrieved: January, 2014]

[12] G.Salton and M. J. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, 1983.

[13] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [retrieved: 1, 2014]

[14] http://wordnet.princeton.edu/ [retrieved: January, 2014]

 ISBN: 978-1-61208-385-8