An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding

Yi-Chien Lin, Li-Chun Sung, Meng Chang Chen
Institute of Information Science, Academia Sinica, Taiwan
{jascha, mico63, mcc}@iis.sinica.edu.tw

ABSTRACT
In this paper, we propose a multiple-choice question generation methodology for understanding evaluation of adjectives in a text. Based on the sense association among adjectives, an adjective being examined can be usually substituted by some other adjectives. In order to discourage learners from answering questions by recalling memorized answers, and to expose learners with more vocabularies, the answer of the questions generated is a substitute of the adjective being examined. The candidates of a substitute are gathered from WordNet and filtered by web corpus searching. Based on the proposed methodology, the choice candidates that are not selected as the answer are still useful. They are more distracting than the dissimilar adjectives and can be used as distractors to improve the distinguishability of a question.

Keywords
Multiple-choice question, automatic question generation

1. INTRODUCTION
Quizzing is a popular approach to measure learners’ English learning effect. One common quizzing form is the multiple-choice question. A multiple-choice question usually consists of a description about the quizzing target and several choices, including one answer and other incorrect choices. In such a question, incorrect choices play an important role in distracting the less-proficient examinees and are therefore called distractors. In the traditional English learning environment, a teacher can manually prepare the corresponding multiple-choice questions for a well-produced learning material in advance. However, the pre-prepared materials may be uninteresting to learners. Thus, to keep learner’s interest, the UWiLL system [1] permits a learner to select any web pages that fit his/her interest as learning materials based on the concept of ubiquitous learning. Unfortunately, in such a system, manual question generation becomes infeasible. To overcome this problem, one solution is to generate questions automatically from given texts.

1.1 Related Work
At present, some methodologies of generating multiple-choice question by computer have been proposed. Generally speaking, the question generation processes can be summarized as follows: (i) identify the quizzing target (terms in a given text), (ii) generate the question description from the sentence containing the quizzing target, and (iii) generate choices, including an answer and some distractors.

For example, Mitkov and Ha proposed a computer-aided question generation environment [2]. For a given text, the quizzing targets are nouns and noun phrases selected by frequency or some pre-defined syntactic pattern. However, the last two steps of the generation process in this environment cannot perform automatically because of two impediments: (i) the question description is generated based on the semantic of the quizzing target which may be unknown, and (ii) a teacher must get involved in distractor filtering and refinement.

After that, some automatic methodologies were proposed [3][4][5][6]. To avoid the first impediment, the question description in these methodologies is just the sentence containing the quizzing target modified by replacing the quizzing target with a blank. And the second impediment is avoided by extracting the terms dissimilar to the quizzing target as distractors based on some thesaurus such as WordNet [7]. In the above works about English vocabulary assessment, some of them found distractors with the help of a thesaurus [5][6]. Some researchers chose some words in the same article as distractors directly [4], and other researchers used a more complicated mathematical approach to determine the distractors among a number of words which have the same part-of-speech and word type [3].

However, for these previous works, the answer is just the quizzing target. Thus, a generated question may be answered by recalling memorized answers, not by the understanding of the quizzing target.

1.2 Goal
To motivate learners to answer questions by understanding instead of by memorizing, we propose an idea to represent the answer by a substitute of the quizzing target. In this paper, because the relationships among adjectives are more definite than other types of vocabularies (e.g. nouns or noun phrases), we apply the idea on adjectives and propose a multiple-choice question generation methodology for understanding evaluation of an adjective in a sentence. Based on the proposed methodology, an examinee will answer such a question correctly if he/she understands the meaning of the adjective being examined. The examinee can also learn the related adjectives in the meantime.

The remainder of this paper is organized as follows. Section 2 describes the process of our proposed methodology. We show the evaluation of the generated questions in section 3. Conclusion and future work are discussed in section 4.
2. THE PROPOSED METHODOLOGY

In general, an adjective may own some senses. The applied sense of an adjective will be determined during adding the adjective in a sentence of a text. Since a sense can be shared by some adjectives, an applied adjective in a sentence may be substituted by other adjectives that own the same sense as the applied sense of the adjective. The final substitution decision is based on whether a substitute candidate can be used to modify the noun or noun phrase modified by the original adjective.

Thus, to check the availability of substitute candidates of an adjective being examined, in the step of quizzing target identification, we will extract each adjective-noun pair and the sentence containing the pair in a given text. The modifying relationship is identified based on the Link Grammar Parser [8] developed by Carnegie Mellon University.

Basically, we will try to generate a multiple-choice question for each adjective-noun pair. According to the characteristic of an adjective and the sentence containing it, the question will be generated based on one of three predefined question types. The question types formulate the question description form and the characteristic of choices. We will introduce them in section 2.1. The process of choice candidate generation and filtering are described through section 2.2 to 2.5.

2.1 Question Types

There are three types of questions our system is able to generate: questions for collocations, questions for antonyms, and questions for synonyms or similar words.

Given an article, we treat it as a number of sentences and make use of Link Grammar Parser to parse each sentence. For a given sentence, one or more adjective-noun pairs are extracted. For instance, in this sentence “Learning English is not an easy job.”, an adjective-noun pair (easy, job) is extracted. For each adjective-noun pair, we then try to generate a question about the original adjective, which is “easy” in the above example. The generated multiple-choice question has a description, which asks the examinee to choose an adjective from following four choices, including one answer and three distractors. According to the characteristic of an adjective-noun pair and the structure of a sentence, questions can be divided into three types by the following procedures shown in Figure 1.

The sentence which generates a question for collocations must contain an adjective-noun pair collocation. In general, words in a collocation cannot be replaced. Therefore, the question asks the examinee to find an adjective from the choices to replace the original adjective, and the answer for a question of this type is like “the adjective cannot be replaced here” or “none of the above”. An example of a question of this type is shown in Figure 2.

| 1. Given a sentence, one or more adjective-noun pairs are extracted. |
| 2. For each adjective-noun pair, if the pair is a collocation, try to generate a question for collocation. |
| 3. If the original sentence contains words that have negative meanings, try to generate a question for antonyms. |
| 4. Try to generate a question for synonyms or similar words. |

Figure 1: The procedure of determining the question type

In this sentence “In high school, I was crazy about English songs.”, the adjective “high” can be replaced with:

1. low
2. advanced
3. broad
4. none of the above

Figure 2: An example question for collocations

For an adjective being examined, if the sentence containing it is negative and can be transformed to an affirmative sentence by eliminating the negative term and replacing the adjective with its antonym, we can generate a question whose answer is the antonym of the adjective. In this paper, we decide whether a negative sentence can be transformed based on the existence of a predefined pattern in the linkage of the sentence obtained by Link Grammar Parser. The pattern is “not – EBm – O – A – adj”. Figure 3 illustrates an example of a question of this type.

In this sentence “Learning English is not an easy job.”, the adjective “high” can be replaced with:

1. uneasy
2. casual
3. effortless
4. difficult

Figure 3: An example question for antonyms

For other adjective-noun pairs, our system generates questions for synonyms or similar words. In a question of this type, the description asks the examinee to find an adjective which has a similar meaning as the original adjective. Figure 4 is an example for questions of this type. The flow of question generation is illustrated in Figure 5, and the detailed process of generating questions for synonyms or similar words is illustrated in Figure 6.
In this sentence “Learning English is not an easy job.”, the meaning of the adjective “easy” is similar to:

1. available
2. comfortable
3. light
4. simple

Figure 4: An example question for synonyms or similar words

2.2 Generating Choice Candidates

By consulting WordNet database, choice candidates are obtained. Besides, we also collect some adjectives that can be randomly selected as choice candidates if necessary. Choice candidates are divided into two groups, simply answer candidates and distractor candidates.

As mentioned above, words in a collocation are usually irreplaceable. Therefore, all we have to do for a question for collocations is just generate some distractors and make examinees confused. To do this, the best source of a distractor is synonyms of the original adjective. If synonyms are not as many as needed, similar words or antonyms will be another source of distractors.

For questions for antonyms, the answer is an antonym of the original adjective. To generate answer candidates, we consult WordNet and obtain all antonyms of the original adjective. Synonyms and similar words are collected as distractor candidates.

On the other hand, the answer candidates of a question for synonyms or similar words come from synonyms and similar words, and the distractor candidates from antonyms. It is noted that answer candidates may move to distractor candidates in some situations, which will be described later.

2.3 Filtering out Unsuitable Choice Candidates

After generating a number of choice candidates, we have to filter out some of them which are clearly incorrect and cannot distract examinees. Since Google search engine is well-known and covers enormous web pages, we utilize Google to check whether an adjective can be one of modifiers of a noun.
In the three types that questions for adjectives are divided into, questions for collocations do not try to find a word that can replace the original adjective because we do not think the original adjective is replaceable. On the other hand, for questions for antonyms and questions for synonyms or similar words, we look for an adjective that not only has an opposite or the similar meaning as the original adjective but also can modify the noun. Moreover, adjectives are used to describe properties of nouns in two forms: attributive and predicative uses [9]. An example for attributive uses is “an easy job”, and “a job is easy” for predicative uses. Therefore, we develop two search patterns corresponding to these two forms of adjective uses.

For an adjective adj and a noun n, we consult Google search engine and apply two patterns, which are “adj n” and “n is adj”, as search strings. A threshold th is set and a choice candidate will be filtered out if the sum of these two search result counts returned by searching these two strings is smaller than the threshold th. Otherwise, we remain the adjective in choice candidates.

2.4 Determining the Answer
Differing from the previous studies on multiple-choice question generation [2][3][4][5][6], our proposed approach does not use the original adjective as the answer. To determine the answer from the answer candidates, we expand the query string by adding some words that may represent the main idea of the whole text and again consult Google to obtain the search result count of each choice candidate. Since the added words represent the main idea of the whole text, the search result count after adding it in the search string may indicate the string “adj n” or “n is adj” has some relationship with the main idea of the document. Here we propose two methods to generate these related words. Intuitively, the first sentence of an article tells about the main idea of the whole story. We remove stop words from this sentence and use the remaining as the related words. In addition, we also develop a simple model to rank all non-stop words in an article and then collect the top ones. Let \( t_1, t_2, ..., t_n \) be the non-stop words in an article, and \( \text{tf}(t_i) \) the number of occurrences of \( t_i \). Let \( \text{bncf}(t_i) \) be the number of occurrences of \( t_i \) in British National Corpus (BNC) [10]. The following formula is used to calculate the weight of a non-stop word \( t_i \).

\[
\text{weight}(t_i) = \frac{\text{maxbncf}(t_i)}{\text{bncf}(t_i)}
\]

A word \( t_i \) with high BNC frequency indicates it may be a common word, and we reduce its weight by dividing \( \text{bncf}(t_i) \). Since the idea of the above formula is borrowed from term frequency-inverse document frequency (tf-idf), we call it the tf-idf-like approach in this work. After ranking all non-stop words, we choose some top words as the related words.

As mentioned above, we don’t need to find the answer from choice candidates for a question for collocations. For other two types of questions, the answer is the one with the biggest Google search result count among answer candidates. If there is no candidate that can be determined correct, that is, the search result count of each answer candidate is zero, we give up this adjective-noun pair and then try to generate a question by using the next pair.

2.5 Collecting Distractors
After determining the answer, the remaining task is to collect distractors. For questions for synonyms or similar words, an important thing is to avoid ambiguities by eliminating distractors that may not be incorrect. In WordNet, two synsets that are “Similar To” each other usually contain some words that have very similar meaning. If more than one of these words is selected as a choice, it will be difficult for examinees to choose one from them. Therefore, we eliminate all synonyms and similar words belonging to the synset that the answer is obtained from. Recall the above example. The 5th synset of the adjective “easy” contains three words: easy, gentle, and soft. If “gentle” is selected as a choice, then “soft” will not be a choice of this question. Since a synset represents a meaning and contains one or more words. If more than one word in the same synset is chosen, it is possible that more than one choice is considered correct. In this case, examinees are confused even if they do understand the meaning of each choice and the original adjective in the sentence.

In our work, a multiple-choice question has three distractors. If the number of distractor candidates is smaller than three, some randomly selected adjectives are chosen to generate this question. Otherwise, we pick three choices from the candidates by choosing those with top three search result count, or more simply, by random.

3. EVALUATION
To evaluate the generated questions, we choose Far East senior high school English textbook Book One, which contains 12 articles (lessons), as the experiment material. Each article is parsed by Link Grammar Parser. According to the parsing result, we extract all adjective-noun pairs and try to generate questions for each pair. The threshold \( th \), which is used to filter out unsuitable choice candidates, is set to zero because we do not drop any choice candidates in the choice candidate filtering stage and only focus on the performance of the methods to determine the answer.

In our proposed approach, an answer candidate with the biggest search result count is considered as the answer even if the count is very small. Here we introduce another threshold \( th^* \), which is used to filter out some questions whose answer is with a very small search result count. When the threshold \( th^* \) is set to 1, we don’t filter out a question as long as the search result count of its answer is greater than or equal to 1.
In order to improve the correctness of the answer, we adopt two strategies to filter out some questions. First, we set the threshold \( \hat{t} \) to 10. This strategy helps remove some noises in web pages. However, since this strategy ignores the difference of the frequency of each word, the second strategy is then developed. We add the original adjective in the article into the choice candidates, and also get a Google search result count by adding the same related words in the step of determining the answer. We filter out questions when the proportion between the search result count of the answer and the search result count of the original adjective is smaller than 0.01.

From the textbook, 50 non-collocation adjective-noun pairs (redundant pairs are counted only one time) are extracted. With the help of WordNet, one human expert examines each pair and all of the corresponding choice candidates our system generates. A pair is marked if it is sufficient for the human expert to generate a multiple-choice question by using the generated choices candidates. In these 50 adjective-noun pairs, 24 of them were marked by the human expert. After the questions are generated and filtered out, the human expert then examines each question and decides if it is a valid one. The generated questions are evaluated by precision, recall, and F1 Score, which are defined as follows.

\[
\text{precision} = \frac{\text{number of valid questions}}{\text{number of questions generated}}
\]

\[
\text{recall} = \frac{\text{number of valid questions}}{\text{number of valid questions marked by the human}}
\]

\[
F1 \text{ Score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

Table 1 shows the experiment results of four proposed approaches to determine the answer. The numbers in the parenthesis behind the precision and recall indicate the number of questions. For example, our system generates 22 questions and 18 of them with a correct answer, and achieves a 0.818 precision when the first-sentence method and strategy 1 is adopted. It is observed that the related words used to determine the answer are useful in improving the precision of the answer. The first-sentence method performs well in terms of precision. As to the tf-idf-like approach, the best precision is obtained when tf-idf-2 is used to determine the answer. However, both precision and recall rate are reduced when we add the third word into the query string (tf-idf-3). As we know, sending too many words to a search engine may result in very few search results, or even no result is returned. This influences the performance of the tf-idf-like approach. In addition, our proposed strategies help improve the precision as well. On the other hand, using the tf-idf-1 approach obtains the best recall rate and F1 score.

### Table 1: Experiment results

(tf-idf-\( n \) indicates the top \( n \) words ranked by the tf-idf-like approach are used)

<table>
<thead>
<tr>
<th></th>
<th>I (baseline)</th>
<th>10 (strategy 1)</th>
<th>0.01 (strategy 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st} sentence</td>
<td>Precision 0.760 (19/25) 0.792 (19/24) 0.776 (19/24) 0.677 (21/31) 0.741 (20/27) 0.784 (20/24)</td>
<td>Recall 0.792 (19/24) 0.750 (18/24) 0.783 (19/24) 0.792 (19/24) 0.842 (16/19) 0.744 (16/24)</td>
<td>F1 Score 0.818 (17/21) 0.784 (17/24) 0.756 (17/24) 0.808 (21/26) 0.750 (18/22) 0.783 (18/24)</td>
</tr>
<tr>
<td>tf-idf-1</td>
<td>Precision 0.677 (21/31) 0.875 (21/24) 0.764 (21/24)</td>
<td>Recall 0.792 (19/24) 0.875 (19/24) 0.792 (19/24)</td>
<td>F1 Score 0.808 (21/26) 0.875 (21/24) 0.840 (21/24)</td>
</tr>
<tr>
<td>tf-idf-2</td>
<td>Precision 0.741 (20/27) 0.833 (20/24) 0.784 (20/24)</td>
<td>Recall 0.842 (16/19) 0.667 (16/24) 0.744 (16/24)</td>
<td>F1 Score 0.818 (18/22) 0.750 (18/24) 0.783 (18/24)</td>
</tr>
<tr>
<td>tf-idf-3</td>
<td>Precision 0.704 (19/27) 0.792 (19/24) 0.745 (19/24)</td>
<td>Recall 0.833 (16/19) 0.625 (15/24) 0.714 (15/24)</td>
<td>F1 Score 0.750 (15/20) 0.625 (15/24) 0.682</td>
</tr>
<tr>
<td>No words are added</td>
<td>Precision 0.647 (22/34) 0.917 (22/24) 0.759 (22/24)</td>
<td>Recall 0.667 (22/33) 0.917 (22/24) 0.772 (22/24)</td>
<td>F1 Score 0.710 (22/31) 0.917 (22/24) 0.800</td>
</tr>
</tbody>
</table>

4. CONCLUSION

In this paper, we propose an automatic multiple-choice generation methodology. In order to encourage the learners to understand the meaning of the adjectives, the answer of the generated questions is a substitute of the adjective being examined. This makes our work very different from previous studies, and learners also have an opportunity to learn more words.

To obtain the choice candidates of the questions, we look up the synonyms, antonyms, and similar words of the adjective being examined. Based on web corpus searching, we check the usage and filter out unsuitable choice candidates. And then, we determine the answer and collect distractors. Experiment results have shown that our proposed answer determination approaches and question filtering strategies are effective in precision.

It is very difficult for questions that are automatically generated to be as good as questions generated by human experts. Currently, our methodology focuses on improving the correctness of the answer. In the future, the quality improvement of distractors will be the next step of this work.
ACKNOWLEDGMENTS
We would like to thank the anonymous reviewers for their useful suggestions and recommendations on the earlier version of this paper. This work is supported partly by NSC of Taiwan under grants 95-2524-S-001-001.

REFERENCES


