Automatic Generation of English Vocabulary Tests

Yuni Susanti[1], Ryu Iida[2] and Takenobu Tokunaga[1]
[1]*Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan*
[2]*National Institute of Informations and Communication Technology, Tokyo, Japan*

Keywords: Automatic Question Generation, English Vocabulary Test, Word Sense Disambiguation, Multiple-choice Question, TOEFL Vocabulary Test.

Abstract: This paper presents a novel method for automatically generating English vocabulary tests using TOEFL vocabulary questions as a model. English vocabulary questions in TOEFL is a multiple-choice question consisting of four components: a target word, a reading passage, a correct answer and distractors. Given a target word, we generate a reading passage from Web texts retrieved from the Internet, and then employ that reading passage and the WordNet lexical dictionary for generating question options, both the correct answer and distractors. Human evaluation indicated that 45% of the responses from English teachers mistakenly judged the automatically generated questions by the proposed method to be human-generated questions. In addition, half of the machine-generated questions were received average rating more than or equals than 3 in 5 point scale. This suggests that our machine-generated questions succeeded in capturing some characteristics of the human-generated questions, and half of them can be used for English test.

1 INTRODUCTION

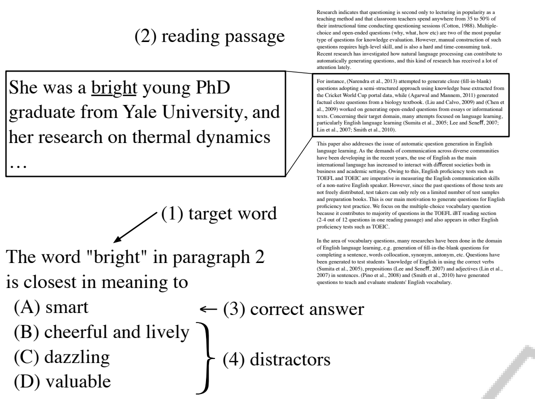
Research indicates that questioning is second only to lecturing in popularity as a teaching method and that classroom teachers spend anywhere from 35% to 50% of their instructional time conducting questioning sessions (Cotton, 1988). Multiple-choice and open-ended questions (why, what, how, etc) are two of the most popular types of questions for knowledge evaluation. However, manual construction of such questions requires a high level of skill, and is also a hard and time-consuming task. Recent research has investigated how natural language processing can contribute to automatically generating questions, and this kind of research has received a lot of attention lately. For instance, (Narendra et al., 2013) attempted to generate *cloze* (fill in the blank) questions adopting a semi-structured approach using knowledge base extracted from the Cricket World Cup portal data, while (Agarwal and Mannem, 2011) generated factual *cloze* questions from a biology textbook. (Liu and Calvo, 2009) and (Chen et al., 2009) worked on generating open-ended questions from essays or informational texts. Concerning their target domain, many attempts have focused on language learning, particularly English language learning (Sumita et al., 2005; Lee and Seneff, 2007; Lin et al., 2007; Smith et al., 2010).

Similarly, this paper also addresses the issue of automatic question generation in English language learning. As the demands of communication across diverse communities have been developing in recent years, the use of English as the main international language has increased to enable this interaction between different societies both in business and academic settings. Owing to this, English proficiency tests such as TOEFL and TOEIC are imperative in measuring the English communication skills of a non-native English speaker. However, since the past questions of those tests are not freely distributed, test takers can only rely on a limited number of test samples and preparation books. Providing test takers a rich resource of English proficiency test questions is the main motivation of this research. We focus on multiple-choice vocabulary questions because it contributes to the majority of questions in the TOEFL iBT[1] reading section (3-5 questions out of a total of 12-14 in one reading passage) and also appears in other English proficiency tests such as TOEIC.

In the area of vocabulary questions, many studies have been done in the domain of English language learning, e.g. generation of fill-in-the-blank

[1] TOEFL iBT is an Internet-based test which is the newest version of the TOEFL test; for more information refer to www.ets.org/toefl

Susanti Y., Iida R. and Tokunaga T..
Automatic Generation of English Vocabulary Tests.
DOI: 10.5220/0005437200770087
In *Proceedings of the 7th International Conference on Computer Supported Education* (CSEDU-2015), pages 77-87
ISBN: 978-989-758-107-6



She was a bright young PhD graduate from Yale University, and her research on thermal dynamics …

Research indicates that questioning is second only to lecturing in popularity as a teaching method and that classroom teachers spend anywhere from 35 to 50% of their instructional time conducting questioning sessions (Cotton, 1988). Multiple-choice and open-ended questions (why, what, how etc) are two of the most popular type of questions for knowledge evaluation. However, manual construction of such questions requires high-level skill, and is also a hard and time-consuming task. Recent research has investigated how natural language processing can contribute to automatically generating questions, and this kind of research has received a lot of attention lately.

For instance, (Narendra et al., 2013) attempted to generate cloze (fill-in-blank) questions adopting a semi-structured approach using knowledge base extracted from the Cricket World Cup portal data, while (Agarwal and Mannem, 2011) generated factual cloze questions from a biology textbook. (Liu and Calvo, 2009) and (Chen et al., 2009) worked on generating open-ended questions from essays or informational texts. Concerning their target domain, many attempts focused on language learning, particularly English language learning (Sumita et al., 2005; Lee and Seneff, 2007; Lin et al., 2007; Smith et al., 2010).

This paper also addresses the issue of automatic question generation in English language learning. As the demands of communication across diverse communities have been developing in the recent years, the use of English as the main international language has increased to interact with different societies both in business and academic settings. Owing to this, English proficiency tests such as TOEFL and TOEIC are imperative in measuring the English communication skills of a non-native English speaker. However, since the past questions of those tests are not freely distributed, test takers can only rely on a limited number of test samples and preparation books. This is our main motivation to generate questions for English proficiency test practice. We focus on the multiple-choice vocabulary question because it contributes to majority of questions in the TOEFL iBT reading section (2-4 out of 12 questions in one reading passage) and also appears in other English proficiency tests such as TOEIC.

In the area of vocabulary questions, many researches have been done in the domain of English language learning, e.g. generation of fill-in-the-blank questions for completing a sentence, words collocation, synonym, antonym, etc. Questions have been generated to test students 'knowledge of English in using the correct verbs (Sumita et al., 2005), prepositions (Lee and Seneff, 2007) and adjectives (Lin et al., 2007) in sentences. (Pino et al., 2008) and (Smith et al., 2010) have generated questions to teach and evaluate students' English vocabulary.

The word "bright" in paragraph 2 is closest in meaning to

(A) smart
(B) cheerful and lively
(C) dazzling
(D) valuable

Figure 1: Four components in a vocabulary question asking for closest in meaning of a word.

questions for completing a sentence, word collocation, synonym, antonym, etc. In previous research, questions have been generated to test students knowledge of English in correctly using the verbs (Sumita et al., 2005), prepositions (Lee and Seneff, 2007) and adjectives (Lin et al., 2007) appearing in sentences. (Pino et al., 2008) and (Smith et al., 2010) have generated questions to teach and evaluate students English vocabulary.

In this research, we adopt TOEFL vocabulary questions as the format. This type of vocabulary question asks the closest option in meaning to a given word. As shown in Figure 1, this type of question is composed of four components: (1) a target word, (2) a reading passage in which the target word appears, (3) a correct answer, and (4) distractors (incorrect options). To generate a question, we need to produce these four components.

One possible approach for generating such questions is using a manually-created lexical knowledge base such as WordNet (Fellbaum, 1998). (Brown et al., 2005) generated multiple-choice questions by taking their components from WordNet. (Lin et al., 2007) also adopted WordNet to produce English adjective questions from a given text. The candidates of options (correct answer and distractors) are taken from WordNet and filtered by Web searching. Unlike previous work, we propose a method for question generation by utilising Web texts from the Internet in addition to information from WordNet. Producing the reading passage from Internet materials enables us to provide learners with fresh, updated, and high-quality English reading passages.

Another focus of this research is generating not only single-word options for a correct answer and distractors, but also multiple-word options that past research did not deal with. Since multiple-word options are often used in actual English vocabulary tests like TOEFL, introducing multiple-word options would make generated questions more natural and closer to human generated questions.

As shown in Figure 1, a multiple-choice vocabulary question consists of four components, thus given a target word with its part-of-speech (noun, verb, adjective or adverb) as an input, the task of generating this kind of question can be broken down into three tasks: (1) reading passage generation, (2) correct answer generation, and (3) distractor generation. In the next three sections, we describe each task in detail followed by an evaluation experiment. Finally we conclude the paper and look at future directions.

2 READING PASSAGE GENERATION

In English proficiency tests such as TOEFL, the reading passage is taken from university-level academic texts with various subjects such as biology, sociology, history, etc. In this work we generate similar passages, but do not limit ourselves to academic texts; the Internet is used as the source for generating the reading passages. In addition, text domains can be controlled by choosing the target sites for retrieving texts from the Internet. Here the users, e.g. English teachers, can choose the sites depending on their interest. For example if the users prefer news articles on the subject of technology, they can choose sites such as www.nytimes.com on "Technology" category. Utilising a reading passage retrieved from the Internet (especially from news portals) gives a lot of benefits because such texts tend to be new and up-to-date, in terms of both content and writing-style. They also come from broad genres and topics, making them well suited for English language learning.

A straightforward approach to generating the questions would be to choose a passage that contains the target word, and then identifying its word sense for generating the correct answer and distractors. In general, a word in a dictionary has several meanings, while the word in a given passage is used for representing a specific one of those meanings. The task of identifying the correct word sense within a context has been studied in natural language processing research under the name of "word sense disambiguation (WSD)".

2.1 Word Sense Disambiguation

Word sense disambiguation (WSD) is the task of identifying the meaning of a word in context in a computational manner (Navigli, 2009). Vocabulary questions in this research ask test takers to select the option closest in meaning to a target word in context; to generate a correct option we need to identify the meaning of the target word in a particular context in the first place. Therefore, WSD is crucial for generating vocabulary questions, especially to generate a correct answer and distractors. The state-of-the-art WSD methods as explained in (McCarthy, 2009) reach around 0.37 in accuracy with a knowledge-based approach, 0.88 with supervised and 0.82 with unsupervised machine learning approaches. Further explanation on WSD can be found in survey papers by (Navigli, 2009) and (McCarthy, 2009). In this research we use the Lesk algorithm (Lesk, 1986) which chooses the sense that shares the highest number of words in its gloss (or example sentence) in a dictionary and the current context. For instance, given two word senses with their glosses for "key" in a dictionary:

1. *Metal device shaped in such a way that when it is inserted into the appropriate lock the locks mechanism can be rotated.*
2. *Something crucial for explaining; "The key to development is economic integration."*

the word sense of "key" in the context "I inserted the **key** and locked the door." should be identified as word sense 1, because its gloss has a three word overlap ("insert" and two "lock"s) with the context, while word sense 2 has no word overlap at all.

Since even with the state-of-the-art WSD method high accuracy is not always available, past attempts avoided use of WSD for generating vocabulary questions by utilising the most frequent word sense with its example sentences as a context in WordNet (Brown et al., 2005). This is, however, obviously not enough to create decent questions because most frequently used senses in WordNet are based on a small corpus [2] and the reading passages are limited.

To remedy insufficient performance of WSD, we propose combining WSD and our *context search* (CS), introduced below. In WSD, given a target word and its context, the task is to identify the correct word sense of the target word in that context. Context search works in reverse; given a target word and one of its word senses, it searches for passages in which the target word is used with the given word sense. To combine both, WSD is applied to the target word in the retrieved passage by CS to confirm that the predicted word sense is the same as the given sense from CS. In our experiment, we used target words from TOEFL iBT sample questions. However in real applications the users can provide the target words or they can be obtained randomly from an article.

2.2 Context Search

Given a target word and one of its word senses, context search (CS) is a threefold process:

(1) query formation from the example sentence,

(2) retrieval of snippets with a search engine,

(3) scoring snippets to choose one of them as an appropriate reading passage.

A query for the search engine is created from the example sentence of the given word sense by taking the target word and its adjacent two words on both sides after removing stop words such as "the", "on", "are" and so on. When the target word is located at the beginning or the end of the sentence, the two following or preceding words of the target word are taken for the query. The created query is submitted to the search engine to retrieve snippets containing the target word possibly with the given sense. The last step selects a snippet which is the most probable snippet containing the target word with the given sense. Plausibility that the word sense of the target word in the snippets is the same as the given word sense is calculated based on the following three scores: (1) S_o: word overlap between the example sentence and the snippet, (2) S_a: the number of adjacent query words to the target words in the snippet after removing the stop words, (3) S_q: the number of query words appearing in the snippet.

The following is a detailed example of the score calculation. Assume our target word is "bright" with *intelligent* sense, and given the example sentence "She was a **bright** young graduate from my university."[3], we have query words "bright", "young" and "graduate".

Suppose that we retrieved the following two snippets where the query words are underlined and the target word is indicated in bold face.

S1. Mary is a **bright** young PhD graduate from Yale University. She was a sophomore in college when she found her true passion in research.

S2. Since she was a child, Mary has been a friendly girl. Mary always gives a **bright** smile to her friends in the university campus.

The first score S_o, the overlap word score, counts the word overlap between the example sentence and

[2] based on WordNet reference manual, accessed through https://wordnet.princeton.edu/wordnet/documentation/

[3] Note that since "she", "was" and "a" are stop words, the target word "bright" is at the beginning of the sentence after stop word removal, thus "young" and "graduate" are used for the query.

the snippets. The scores S_o for these snippets are $S_o(\mathrm{S1}) = 4$ since “bright”, “young”, “graduate” and “university” overlap, while $S_o(\mathrm{S2}) = 2$ for “bright” and “university”.

The second score S_a counts the number of adjacent query words to the target words in the snippet after removing the stop words. Thus, $S_a(\mathrm{S1}) = 1$ for “young”, and $S_a(\mathrm{S2}) = 0$.

The third score S_q counts the number of query words that appear in the snippet. Thus, $S_q(\mathrm{S1}) = 3$ for “bright”, “young” and “graduate”, while $S_q(\mathrm{S2}) = 1$ for only “bright”.

The three scores are combined to provide the final score for each snippet as given by

$$S = 3S_o + 3S_a + 2S_q \tag{1}$$

The weights of the scores were determined experimentally.

The reading passage for a question would be composed of three sentences: a sentence containing the target word, and the two sentences before and after it. However, if the target word is located in the first or last sentence of the retrieved text, the reading passage would be composed of only two sentences: a sentence containing the target word, and a sentence before or after it.

2.3 Preliminary Evaluation on a Reading Passage Generation

We conducted a preliminary experiment on two target word sets: 98 target words from TOEFL iBT sample questions[4] and preparation books for TOEFL iBT[5], and randomly selected 98 target words from Senseval-2 and Senseval-3 data which were prepared for Senseval WSD workshops[6]. These two target word sets share no common words. The Bing Search API[7] was used as the search engine, and we limited the target site to www.nytimes.com. In this experiment, we compare the results of the following three settings.

- **WSD:** We identify the word sense of the target word in a given context by using the Lesk algorithm.
- **CS:** For each target word in the test sets, context search is applied to find the context sentences in which the target word is used with a given sense.
- **CS+WSD:** WSD is applied after CS to confirm that CS has retrieved snippets containing the target word with a given word sense. The snippets with a word sense mismatch are discarded.

Table 1: Accuracy of WSD and CS.

Setting\ Data	TOEFL	Senseval
WSD	0.602	0.296
CS	0.885	0.745
CS + WSD	0.951	0.843

Evaluation was done manually to see if the identified word sense is correct for the WSD setting, and to see if the retrieved snippet with the highest score includes the target word with the given sense for the CS and CS+WSD settings. Thus we evaluated to what extent we could correctly obtain pairs of word senses and their reading passages. Table 1 shows the accuracy of each setting. The accuracy of CS reached 0.885 on the TOEFL data. In addition, by combining with WSD the accuracy improved to 0.951. Although it is still a preliminary evaluation, the proposed CS method combined with WSD shows promising results for continuing to the next step in generating vocabulary questions.

3 CORRECT ANSWER GENERATION

The correct answer in vocabulary questions is the option that has the closest meaning to the target word used in the reading passage. It does not ask for collocation, therefore the correct answer is not necessarily replaceable with the target word used in the reading passage.

In this work we generate two kinds of correct answers: single-word and multiple-word correct answers. The following subsections describe generation of both single and multiple-word correct answer.

3.1 Single-Word Correct Answer

A single-word correct answer is a correct answer composed of one single word. Based on our analysis of TOEFL iBT official sample questions[8], the correct answer for a vocabulary question satisfies the following requirements.

- It has the same part-of-speech as the target word does.
- It shares a similar meaning with the target word.

[4] They are available at www.ets.org/toefl

[5] The preparation books used are TOEFL iBT preparation books published by Longman, Barron, and McGraw-Hill.

[6] www.senseval.org

[7] https://datamarket.azure.com/dataset/bing/search

[8] 38 vocabulary questions available at www.ets.org/toefl

- It does not share any substrings with the target word. Words with a similar meaning often share substrings in their spelling, for example the word “synchronisation” and “synchronism”. Both words are noun and have similar meaning, but these words cannot be a target word-correct answer pair because the test taker can easily estimate a correct option based on their similarity in spelling.

To satisfy the first and second requirements, the candidates for a single-word correct answer are taken from synonyms of the target word in the dictionary, WordNet[9] in our case. After filtering with respect to those three requirements, remaining candidates are ranked according to the order in their WordNet synset[10], and the highest ranked word sense is chosen as a correct answer. Since the order of the WordNet synsets is based on their frequency in a corpus, we can put high priority on frequent word senses. For instance, given the target word “bright” with word sense *bright.s.02*[11], all of its lemmas, “brilliant” and “vivid” are retrieved from WordNet. Then the orders of the word sense in question, *bright.s.02*, within the synset of each candidate word are compared. Suppose we have the synsets for these two candidates as follows.

Candidate	Synset
“brilliant”	*brilliant.s.01, brainy.s.01, brilliant.s.03, bright.s.02, brilliant.s.05, bright.s.08*
“vivid”	*graphic.s.05, vivid.s.02, bright.s.02, intense.s.03*

Since the order of *bright.s.02* in the synsets of the candidates are the fourth and the third for “brilliant” and “vivid” respectively, “vivid” with the highest order is chosen as a correct answer for the target word “bright”. In the case of a word with no synonym in the dictionary, we use its gloss thus creating a multiple-word correct answer.

3.2 Multiple-Word Correct Answer

A multiple-word correct answer is correct answer composed of more than one word. In Figure 1, multiple-word options are shown in option (B). Note that past research on vocabulary question generation did not deal with multiple-word option which actually appears in the real English proficiency tests. Multiple-word options, both for correct answers and distractors, are generated from the gloss in a dictionary.

The multiple-word options in TOEFL iBT sample questions are usually composed by no more than four words. However, sometimes the gloss can be longer than four. In such case, we simplify the long gloss. In the WordNet lexical dictionary as used in this work, a long gloss tends to include disjunctive structures introduced by disjunctive markers like “or”. In simplifying the gloss, we divide the gloss based on its disjunctive markers. We define the disjunctive markers depending on the dictionary. In the case of WordNet, we use “or” and “;” for disjunctive markers.

In generating multiple-word correct answers, we directly use the gloss of the target word if it consists of no more than four words. If it is longer than four words, it is divided by conjunctive markers and the element which has the least number of words is adopted (elements which consist of only one word are excluded). When the numbers of words in the elements are the same, the left most element is selected. The following are some examples of gloss simplification. The target words and their glosses are shown with the result of simplification underlined, which is used for a multiple-word correct answer.

“accepting”: consider or hold as true

“leaked”: tell anonymously

“lived”: inhabit or live in; be an inhabitant of

4 DISTRACTOR GENERATION

Distractors are the incorrect (or less correct) options in a question. Many multiple-choice questions have four options, thus we generate three distractors for a question.

There are two fundamental requirements for distractors. Firstly, they must be hard to distinguish from the correct answer (or target word), and secondly, they cannot be considered as a correct answer or has to be incorrect. These two requirements seem to be contradicting each other, since the first requires distractors should be somehow similar to the target word, while the second requires distractors should be different from the target word. Making a reasonable tradeoff between these two requirements is important.

In this work, distractor generation is a three-fold process:

(1) collecting distractor candidates,

(2) filtering the candidates so that they fulfil several requirements, and

(3) ranking the candidates based on a scoring function.

[9] In this work we use WordNet 3.0, can be downloaded from https://wordnet.princeton.edu/wordnet/

[10] In WordNet, a word sense is represented in the form of a set of its synonyms. The set is called *synset* (synonym set).

[11] Roughly, the second word sense of adjective “bright”.

4.1 Distractor Candidate Collection

Distractor candidates are collected from two sources: (1) the passage generated by CS and WSD for each target word, and (2) the lexical hierarchy in the WordNet taxonomy. We use these two sources because each of them reflects a different aspect of "similar" relations to the target word. The first is the association relation that the words in a passage are somehow related to each other with respect to the topic that the passage describes. Therefore, those co-occurring words with the target word are reasonable to be distractors. We only consider the co-occurring words with the same part-of-speech and tense as the target word. However co-occurring words themselves are not appropriate for the distractors, since they actually appear in the passage. Therefore we collect their synonyms as distractor candidates for the target word.

The second is the relation in the lexical taxonomy that is defined in a dictionary. We focus on words being sibling to the target word in the WordNet taxonomy. Words in the sibling relation share the same hypernym (parent) with the target word, thus they are somehow similar to the target word.

There are cases in which the number of distractor candidates from those two sources is not enough. When that happens, we take additional candidates from WordNet with the same part-of-speech and with close generality to the target word. All of the word senses with the same part-of-speech in WordNet are ordered in generality, from more general to less general word senses. For instance, the first word sense for noun is *entity.n.01* followed by *physical entity.n.01*, *abstraction.n.06*, *thing.n.12*, *object.n.01*, and so on. We select distractor candidates from words located around the order of the target word in this list.

4.2 Distractor Candidate Filtering

According to (Heaton, 1989), there are several requirements for options in multiple-choice questions. The following are Heaton's requirements followed by the descriptions of our implementation of the requirements. All examples mentioned below are taken from (Heaton, 1989).

(1) *Each option should belong to the same word class as the target word.*
Here we implement this by choosing distractors with the same part-of-speech with the target word.

(2) *Distractors should be related with the correct answer, or come from the same general area.*
We implement this by collecting distractor candidates from the synonyms of co-occurring words in a passage and sibling words in the WordNet taxonomy (4.1) and also calculating the similarity between candidates and the target word (4.3).

(3) *Distractors should have similar word difficulty level with the correct answer.*
We implement this by calculating word-frequency scores, assuming that words with similar frequencies have a similar difficulty level. We use the word frequency list provided by Top 20,000 COCA Academic Corpus[12]. We define the word frequency score by the logarithm of the difference of the word frequency between each candidate and the correct answer. The higher the difference between each candidate-correct answer pair, the less they are close to each other and it means that the distractor does not necessarily have a similar word-difficulty level with the correct answer. Here we will only collect candidates with a word frequency score less than 3.86. This threshold is decided based on 22 TOEFL iBT official sample questions. When the resultant number of candidates is less than 5, we do not apply this requirement.

(4) *Each distractor should have approximately the same length.*
Since we are considering single or multiple-word options, this requirement is not necessarily appropriate. When the options are longer unit like sentences or texts, this could be applicable.

(5) *Avoid using pairs of synonyms as distractors. These kind of distractors can be ruled out easily by test takers.*
If there is a pair of synonyms in the candidates, we will remove one of them and leave no pair of synonym in the distractor candidates. For example, given a target word "courteous" in the sentence "*The old woman was always courteous when anyone spoke to her.*", the options "(A) polite, (B) glad, (C) kind and (D) pleased" are not appropriate, since "glad" and "pleased" are synonyms. The test takers will be able to guess that the correct answer should be one of other two, "polite" or "kind".

(6) *Avoid using antonyms of the correct answer as distractors. The test takers can also easily eliminate these kind of distractors.*
We generate a list of the correct answer's antonyms from WordNet, and exclude them from distractor candidates. For example the options "(A) go up, (B) talk, (C) come down and (D) fetch" for the target word "ascend" are not appropriate, since the antonym pair "go up" and "come

[12] corpus.byu.edu/coca

down" immediately stand out, providing a clue for guessing the correct answer.

4.3 Candidate Scoring and Ranking

At this point, we already have distractor candidates filtered by the requirements mentioned in the previous section. Since we only need three distractors for a question, this step chooses the three most appropriate distractors from the candidates. As mentioned in section 4, a good distractor should be related with the correct answer or target word so that it will be hard to distinguish it from the correct answer.

We rank the distractor candidates with respect to their "closeness" to the correct answer by using a combination of the Path and WU-Palmer similarity score calculated in WordNet. The Path similarity (Pedersen et al., 2004) score is calculated from the shortest path length in the WordNet taxonomy (hypernyms and hyponyms), while the Wu-Palmer similarity (Wu and Palmer, 1994) is calculated based on the depth of two senses in the taxonomy and that of their Least Common Subsumer (most specific ancestor node).

In our experiment, first we collect candidates with a path similarity score less than 0.17 and WU-Palmer similarity score less than 0.36. These thresholds are decided based on 22 TOEFL iBT official sample questions. The resultant candidates are sorted in ascending order of the averaged score of these two. The top three candidates in the ranking are selected as the final distractor candidates.

4.4 Single and Multiple Word Distractors

In the distractors generation step, the final distractor candidates are candidates that are already ranked, in the form of single-word distractors. As we generate both single and multiple-word correct answers, we also generate both types for the distractors. Multiple-word distractors are created from the gloss of the final single-word distractor candidates, with the same gloss simplification method explained in section 3.2.

5 FORMING COMPLETE VOCABULARY QUESTION

We now have all four components for a vocabulary question: the target word (input), a reading passage, the correct answer, and distractors. The last step in creating vocabulary questions is, of course, forming the question itself.

The vocabulary question has a fixed form, as we can see in the examples in Figure 1. Therefore, we can use that form as a question template, and put the correct answer and distractors as question options in that form.

As for the composition of single-word and multiple-word options in the generated question, we generate them randomly with these composition alternatives: all singles, two singles and two multiples, and all multiples.

6 EVALUATION

The purpose of current evaluation is to investigate whether the English vocabulary questions generated by our system have no big difference from English vocabulary questions created by humans, considering the human-generated questions as a goal standard.

6.1 Experimental Setting

We created four evaluation sets in which each evaluation set consists of 10 questions, so in total we have 40 questions to evaluate. In these 40 questions, 22 questions were machine-generated (MQ) and 18 were human-generated (HQ). The target sites for machine-generated questions are www.nytimes.com, www.scientificamerican.com and www.sciencedaily.com.

The human-generated questions were taken from TOEFL iBT sample questions[13]. The target words used for machine-generated questions were also taken from TOEFL iBT sample questions. Similar to the machine-generated questions, we also truncated the original reading passage of human-generated question to make its length into three sentences: a sentence containing the target word, and two sentences , one before and one after. If the target word is located in the first or last sentence of the original passage, we took a sentence before or after it in addition to the sentence containing the target word.

We mixed human-generated and machine-generated questions in a evaluation set and asked evaluators to distinguish between two types of questions. The distribution of the number of questions of each type is shown in Table 2. We prepared a small questionnaire for each question as shown in Figure 2.

Seven English teachers (six non-native and one native) participated in the evaluation as evaluator. Not

[13] www.ets.org/toefl

Table 2: Breakdown of evaluation sets.

Eval. set	#HQ	#MQ	#Evaluator
Set 1	5	5	5
Set 2	6	4	4
Set 3	4	6	4
Set 4	7	3	4

(1) Based on your opinion, the question is:
A. Human-generated
B. Machine-generated

(2) What makes you decide the question is belong to either machine or human-generated questions? (more than one is OK)
A. its correct answer
B. its distractors
C. its passage

(3) What are the reasons behind your answer (2)?

(4) In term of decentness and overall quality, what is the score for this question (1-5, for 5 being the best)?

(5) If there is any, please give suggestion on how to improve this question.

Figure 2: A questionnaire for each vocabulary question.

all of them answered all evaluation sets. In total we obtained 170 responses: 93 responses for machine-generated questions and 77 for human-generated questions.

6.2 Experimental Result and Discussion

6.2.1 Distinguishing between Machine-generated and Human-generated Questions

Table 3 shows a confusion matrix representing the relation between question types (HQ: human-generated and MQ: machine-generated) and the evaluators' judgement.

In the total number of 170 responses, the evaluators correctly judged HQ as human-generated in 53 cases, while they mistakenly judged HQ as machine-generated in 24 cases, thus the accuracy for HQs is 69%. On the other hand, the accuracy for MQs is 55%, which is lower than HQs, but from another viewpoint this is a good indicator because 45% of the responses for machine-generated questions were mistakenly judged as human-generated. This means that these questions succeeded in capturing some characteristics of the human-generated questions, that is why the evaluators mistakenly judged them as human-generated.

Table 3: Responses in distinguishing human-generated and machine-generated questions.

	Question type		
Responses	HQ	MQ	total
human-generated	53	42	95
machine-generated	24	51	75
total	77	93	170

We also analysed the ratio of correct and incorrect judgements for all questions. The result is presented in Figure 3. In the figure, the gray coloured bars indicate human-generated questions (HQs) and the black indicates machine-generated questions (MQs). The solid bar indicates correct judgement and the striped bar indicates incorrect judgement.

There are interesting cases where almost all evaluators incorrectly judged MQs as human-generated (e.g. Q.7, Q.25, Q.27, Q.38 and Q.39). According to feedback from evaluators, the reasons for this are because the distractors seem fairly reasonable, the options are good, and the composition of the question is balanced. On the other hand, all evaluators correctly judged Q.17 as machine-generated. According to their feedback, the distractors of the question are irrelevant and unnatural, therefore they judged the question as machine-generated.

6.2.2 Rationale behind Judgement

Question (2) and (3) ask the reason behind the evaluator's judgement, thus they indirectly assess the quality of the question components, i.e. the reading passage, correct answer, and distractors. In question (2) in the questionnaire, we asked what component of the question made the evaluator believe it was human-generated or machine-generated. Among 51 responses which correctly judged the machine-generated questions to be machine-generated, 66% answered "distractors", 24% answered "correct answers" and 10% answered "reading passage".

We can see here that the distractors were the main reason that led the evaluators into the conclusion of marking it as a machine-generated question. In addition, from the question (3) in the questionnaire we got substantial feedback from the evaluators that the distractors were not appropriate with respect to the reading passage, or irrelevant or unnatural. This result suggests that there is much room for improvement in generating options, particularly for the distractors.

As for the correct answer, an evaluator judged that a question has too obvious correct answer therefore it makes the question too easy. Another evaluator judged that a correct answer is not appropriate, not quite right for the question. Other than those two comments, we do not get any feedback regarding correct answer.

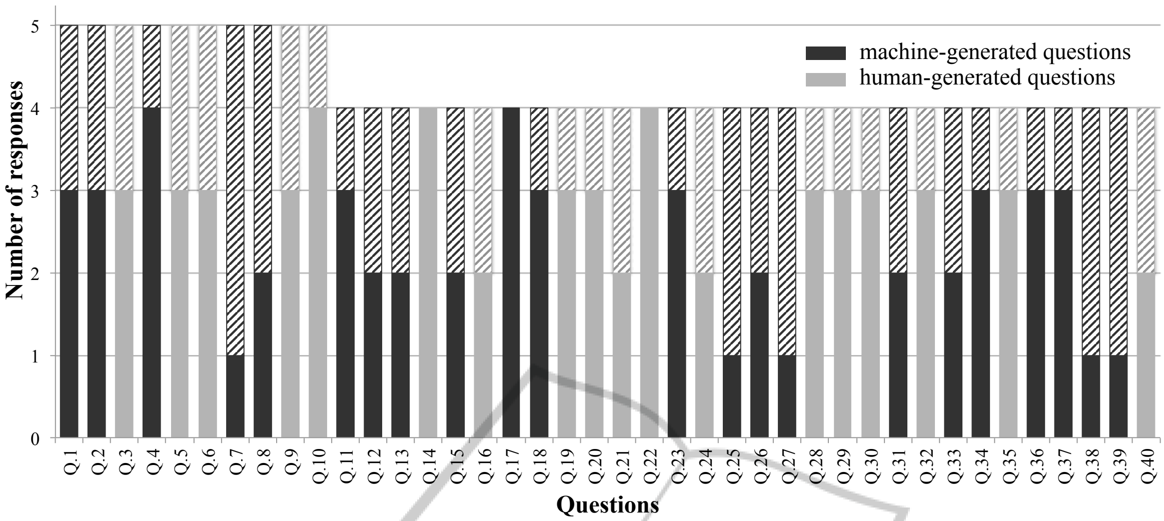


Figure 3: Result of question-wise response. (solid bar: correct response / striped bar: incorrect response).

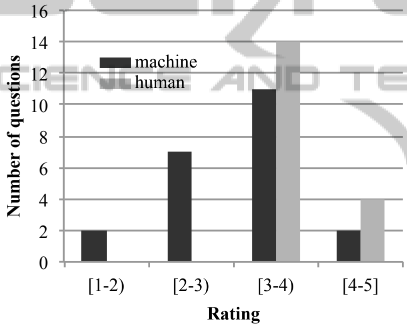


Figure 4: Average rating of questions.

An interesting observation is that multiple-word options could be a reason for judging questions to be both human-generated and machine-generated ones. Some evaluators believed that generating multiple-word options was difficult for a machine, thus they considered questions with multiple-word options to be human-generated. Other evaluators, on the other hand, considered mixture of single and multiple-word options as possible evidence for machine-generated questions. This discrepancy might be related to the expertise of the evaluators. We need further investigation for exploring such relation.

6.2.3 Overall Quality of Questions

In the questionnaire we also asked the evaluators to rate the questions for their decentness and overall quality on a five point scale, 5 being the best. We calculated the average score given by evaluators for each question. The result is presented in Figure 4.

All human-generated questions (gray bars) received average score more than or equals to 3, while about half of the machine-generated questions (13 out of 22) did. This suggests that those machine-generated questions have similarity to questions created by human expert therefore are usable for English tests.

The figure clearly indicates that the human-generated questions are better than the machine-generated questions in term of overall quality. Moreover, half of the machine-generated questions (9 out of 22) received average score less than 3. Further investigation on these poorly rated questions is necessary to see if they have any systematic defect.

There were three machine-generated questions that received particularly low score, only either 1 or 2. Surprisingly, all these three questions include multiple-word options. An evaluator commented that inconsistent length of options was the reason for their judgement. Considering the fact that real TOEFL test uses multiple-word options, it is not always the case that providing various length of options degrades the quality of questions. There could be variance in evaluation criteria among the evaluators we employed.

7 CONCLUSIONS

We have presented a novel method for automatically generating English vocabulary questions, modelling the generated questions after TOEFL vocabulary questions. The vocabulary question asks the test takers to select an option which has closest in meaning to the target word, and consists of four components: a target word, a reading passage, a correct answer and distractors (incorrect options). The proposed method has successfully generated complete vocabulary question in which the passage is generated from the Internet. The options (both correct answer and distractors) are generated from the generated passage and lexical dictionary (we used the WordNet lexical dictionary in our experiment). By producing a passage from Internet materials, we believe that it can provide the learners with a fresh, updated, and high-quality English reading passage so that they are able to learn more.

The evaluation of the machine-generated questions by human experts indicated that 45% of automatically generated questions were mistakenly judged to be human-generated questions. In addition, about half of the machine-generated questions (13 out of 22) were rated more than or equals to 3 in 5 point scale for each question. This suggests that half of the machine-generated questions are similar to those created by human expert therefore quantitatively usable for English tests.

The future work includes improving the poorly rated questions generated by our method through more detailed analysis, and at least in this phase we have already found that generating better distractors is one of the keys. The present evaluation is similar to the Turing test (Turing, 1950), evaluating to what extent our machine-generated questions are similar to those created by human (as a good standard). Our real goal is to generate good questions that can measure test taker's English proficiency precisely. For this goal, evaluation through a real setting with English language learners is indispensable thus we plan to conduct evaluation with the real English learners. We also plan to explore controlling the difficulty of the questions for different levels of English learners.

REFERENCES

Agarwal, M. and Mannem, P. (2011). Automatic gap-fill question generation from text books. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–64.

Brown, J. C., Frishkoff, G. A., and Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826.

Chen, W., Aist, G., and Mostow, J. (2009). Generating questions automatically from information text. In *Proceedings of AIED 2009 Workshop on Question Generation*, pages 17–24.

Cotton, K. (1988). Classroom questioning. *School Improvement Research Series*, pages 1–10.

Fellbaum, C. (1998). *WordNet: A lexical database for English*. A Bradford Book.

Heaton, J. B. (1989). *Writing English Language Tests*. Longman Pub Group.

Lee, J. and Seneff, S. (2007). Automatic generation of cloze items for prepositions. In *Proceedings of Interspeech 2007*, pages 2173–2176.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26.

Lin, Y.-C., Sung, L.-C., and Chen, M. C. (2007). An automatic multiple-choice question generation scheme for English adjective understanding. In *Proceedings of Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007)*, pages 137–142.

Liu, M. and Calvo, R. A. (2009). An automatic question generation tool for supporting sourcing and integration in students' essays. In *Proceedings of the 14th Australasian Document Computing Symposium*.

McCarthy, D. (2009). Word sense disambiguation: An overview. *Language and Linguistics Compass*, 3(2):537–558.

Narendra, A., Agarwal, M., and Shah, R. (2013). Automatic cloze-questions generation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 511–515.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41.

Pino, J., Heilman, M., and Eskenazi, M. (2008). A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems*, pages 22–32.

Smith, S., Avinesh, P., and Kilgarriff, A. (2010). Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, pages 1–6.

Sumita, E., Sugaya, F., and Yamamoto, S. (2005). Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, pages 61–68.

Turing, A. M. (1950). Computing machinary and intelligence. *Mind – A Quarterly Review of Psychology and Philosophy*, LIX(236):433–460.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL 1994)*, pages 133–138.

APPENDIX

Here are some examples of the question generated by our system. Q1 and Q2 are mistakenly judged as human-generated questions by almost all the evaluators (4 out of 5), thus they are considered as acceptable questions. On the other hand, Q3 and Q4 are not so good since many evaluators correctly judged them as machine-generated questions.

Q1 *For example, an opening bid of any sort is usually perceived as a mental anchor, a starting point for the psychological jockeying to follow. If we perceive an opening bid as fundamentally inaccurate or unfair, we reject it by countering with something in another ballpark altogether. But what about less dramatic counter offers?*[14]

The word “reject” in the passage is closest in meaning to

A. hope for,

B. review,

C. refuse to accept,

D. comprehend.

Q2 *Since we now can measure the suns energy output independent of the distorting influence of the atmosphere, we shall see whether the earths temperature trend correlates with measured fluctuations in solar radiation. If volcanic dust is the more important factor, then we may observe the earths temperature following fluctuations in the number of large volcanic eruptions. But if carbon dioxide is the most important factor, long-term temperature records will rise continuously as long as man consumes the earths reserves of fossil fuels.*[15]

The word “fluctuations” in the passage is closest in meaning to

A. a wave motion,

B. the relative position,

C. experimentation,

D. approximation.

Q3 *The article’s big takeaway was that place matters in analyzing relationships between algebra performance and other educational variables. For example, the researchers studied whether a higher percentage of children in poverty was related to lower algebra scores, and whether higher teacher salaries meant higher algebra scores. They found those relationships held true in some districts but not across the board.*[16]

The word “scores” in the passage is closest in meaning to

A. mark,

B. uncovering,

C. determination,

D. deviation.

Q4 *That is, the board no longer provides children in the public schools with sufficient art and music classes. Where will future audiences come from if students are not educated or exposed to the arts in this cultural capital? At John Dewey High School in Brooklyn, a number of English teachers have incorporated the study of art history, music, dance and architecture into our interdisciplinary course curriculums.* [17]

The word “exposed” in the passage is closest in meaning to

A. made into a whole,

B. made accessible to some action,

C. enjoy,

D. divest.

[14] www.scientificamerican.com/article/why-things-cost-1995/

[15] www.scientificamerican.com/article/carbon-dioxide-and-climate

[16] www.sciencedaily.com/releases/2013/07/130701100600.htm

[17] www.nytimes.com/1991/03/08/opinion/l-teachers-help-to-plug-arts-education-holes-281191.html