Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation

Takuya Goto

Graduate School of Information Science, Nagoya University, Japan

Tomoko Kojiri*

Graduate School of Information Science, Nagoya University, Japan
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8603, Japan
E-mail: kojiri@nagoya-u.jp

Toyohide Watanabe

Graduate School of Information Science, Nagoya University, Japan

Tomoharu Iwata

NTT Communication Science Laboratories, Japan

Takeshi Yamada

NTT Science and Core Technology Laboratory Group, Japan

*Corresponding author

Abstract: Since English expressions vary according to the genres, it is important for students to study questions that are generated from sentences of the target genre. Although various questions are prepared, it is still not enough to satisfy various genres which students want to learn. On the other hand, when producing English questions, sufficient grammatical knowledge and vocabulary are needed, so it is difficult for non-expert to prepare English questions by themselves. In this paper, we propose an automatic generation system of multiple-choice cloze questions from English texts. Empirical knowledge is necessary to produce appropriate questions, so machine learning is introduced to acquire knowledge from existing questions. To generate the questions from texts automatically, the system (1) extracts appropriate sentences for questions from texts based on Preference Learning, (2) estimates a blank part based on Conditional Random Field, and (3) generates distracters based on statistical patterns of existing questions. Experimental results show our method is workable for selecting appropriate sentences and blank part. Moreover, our method is appropriate to generate the available distracters, especially for the sentence that does not contain the proper noun.

Keywords: Automatic question generation, multiple-choice cloze question, statistical learning, preference learning, ranking voted perceptron, conditional random field.

Biographical notes: Takuya Goto received the B.E. and M.I. degrees from Nagoya University, Japan, in 2007 and 2009 respectively. His research subject has been the English learning support environment and automatic generation of English exercises. Currently, he works for NTT DOCOMO.

Tomoko Kojiri received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1998, 2000, and 2003 respectively. From 2003 to 2004, she has been a research associate with Graduate School of Information Science, Nagoya University, Japan. From 2004 to 2007, she has been a research associate with Information Technology Center, Nagoya University, Japan. Since 2007, she is an assistant professor with Graduate School of Information Science, Nagoya University, Japan. Her research interests include computer-supported collaborative learning, intelligent tutoring system, and human computer interface. She is a member of IPSJ, JSAI, IEICE, JSET, JSiSE, and APSCE.

Toyohide Watanabe received the B.S., M.E. and Ph.D. degrees from Kyoto University in 1972, 1974 and 1983, respectively. In 1987, he was an Associate Professor in Department of Information Engineering, Nagoya University. He became a Professor in 1994. In 2003, he moved as a Professor to the Department of Systems and Social Informatics, Graduate School of Information Science, Nagoya University. His research interests include knowledge management of personal intelligent activity, computer supported collaborative learning, social environment simulation, spatio-temporal model and geographic information systems and so on. He is a member of the ACM, AAAI, AACE, KES International and the IEEE-CS.

Tomoharu Iwata received the B.S. degree in environmental information from Keio University in 2001, the M.S. degree in arts and sciences from the University of Tokyo in 2003, and the Ph.D. degree in informatics from Kyoto University in 2008. He is currently a researcher at Learning and Intelligent Systems Research Group of NTT Communication Science Laboratories, Kyoto, Japan. His research interests include data mining, machine learning, information visualization, and recommender systems.

Takeshi Yamada received the B.S. degree in mathematics from the University of Tokyo in 1988 and the Ph.D. degree in informatics from Kyoto University in 2003. He was a Leader of Emergent Learning and Systems Research Group of NTT Communication Science Laboratories and is currently a Senior Manager of NTT Science and Core Technology Laboratory Group. His research interests include data mining, statistical machine learning, graph visualization, metaheuristics and combinatorial optimization. He is a member of the ACM and IEEE.

1. Introduction

Spread of e-learning in English enables students to study English with various questions provided on the web. Most of the existing questions have been produced by experts. However, English expression differs in relation to its genre, so it is important for students to tackle questions generated from texts in various genres. In addition, students are highly motivated and willing to study if questions of interesting genres are generated automatically from various texts, such as articles, research papers, and web documents that are selected by the students. A lot of automatic generation systems of various types

of questions were proposed (Funaoi, Akiyama & Hirashima, 2006, Mitkov & Ha, 2003). However, these researches focused on generating questions from single sentence.

In this paper, we propose a system for the automatic generation of multiple-choice cloze questions from texts. For multiple-choice cloze questions, grammatical structures and vocabularies that build the basis of the sentences, determine the appropriateness of the questions. Sentences with a too complicated grammatical structure or a too simple one are not appropriate for a question. Sentences that contain words whose usage is confusing are often selected for questions. In order to avoid inappropriate questions, the selection of sentences that consist of words or word classes appearing frequently in texts (*appropriate sentence*) is important. Furthermore, a blank part of the question indicates the target knowledge to be asked. The appropriate blank part depends on the structure of the sentence. Of course, distracters also represent the target knowledge of the questions. For example, if distracters consist of synonyms, the question asks the meaning of the word. If all distracters are the conjugation of the same verb, the grammatical knowledge for the verb may be asked. Of course, the level of difficulty of the questions varies according to the distracters. If distracters whose word types and meanings are totally different from the correct choice are selected, the question becomes very easy. On the other hand, questions get tricky when distracters have the same word types or similar meanings.

According to their experience, experts usually select appropriate sentences and determine a blank part and distracters that are effective for the sentences. This knowledge depends on the genre the sentence belongs to, so it is difficult to describe the knowledge for all genres. Moreover, parts of this knowledge are heuristics which are too complicated to be explained explicitly. On the other hand, the existing questions may implicitly represent a heuristic knowledge on generating questions. In order to acquire experts' heuristic knowledge on generating questions, our system extracts vocabularies and grammatical features from existing multiple-choice cloze questions based on machine learning and statistical approaches, and applies them to generate new questions from existing texts. By preparing existing questions from different genres, knowledge on generating new questions of those genres is extracted. Therefore, our system can generate multiple-choice cloze questions of any genre automatically.

2. Automatic Generation of Multiple-Choice Cloze Questions

Figure 1 shows the target learning environment. In order to generate and study multiple-choice cloze questions from a particular text, firstly, student inserts text into the system. The text is decomposed into sentences, and multiple-choice cloze questions are generated for each sentence by the system. In order for students to study effectively with the generated questions, appropriate questions need to be selected according to the student's level of understanding. Currently, we focus only on the stage of generating questions and do not consider the effect of the generated questions on a student.

For the purpose of generating multiple-choice cloze questions from texts automatically, the system needs to (1) extract sentences from texts which are appropriate for multiple-choice cloze questions, (2) determine a blank part from the sentence, and (3) generate distracters. Various automatic generation systems of multiple-choice cloze questions have been proposed (Sumita, Sugaya & Yamamoto, 2004, Lin, Sung & Chen, 2007, Brown, Frishkoff, & Eskenazi, 2005, Coniam, 1997). Sumita et al. proposed an automatic generation method of multiple-choice cloze questions for measuring English proficiency (Sumita, Sugaya & Yamamoto, 2004). In this method, leftmost single verb is selected as a blank part. Yi-Chein et al. also constructed an automatic generation system for multiple-choice cloze questions (Lin, Sung & Chen, 2007). They focused on

questions for determining an "adjective" and generated questions whose blank parts are adjectives. In these researches, candidates of distracters are generated using a Thesaurus or WordNet and their appropriateness is verified by searching a corresponding phrase on the web. One of the problems of these researches is that systems do not validate whether given sentences are "appropriate" as multiple-choice cloze questions. Sentences are sometimes too simple or too complicated to represent the questions. In our approach, sentences that are similar to sentences in existing questions are extracted in an "appropriate" order as questions by learning words and grammatical patterns in the existing questions based on machine learning approaches (Figure 2 (1)).

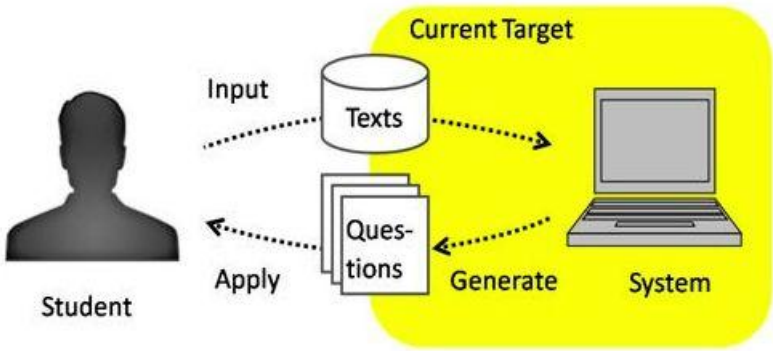

Figure1. The target learning environment

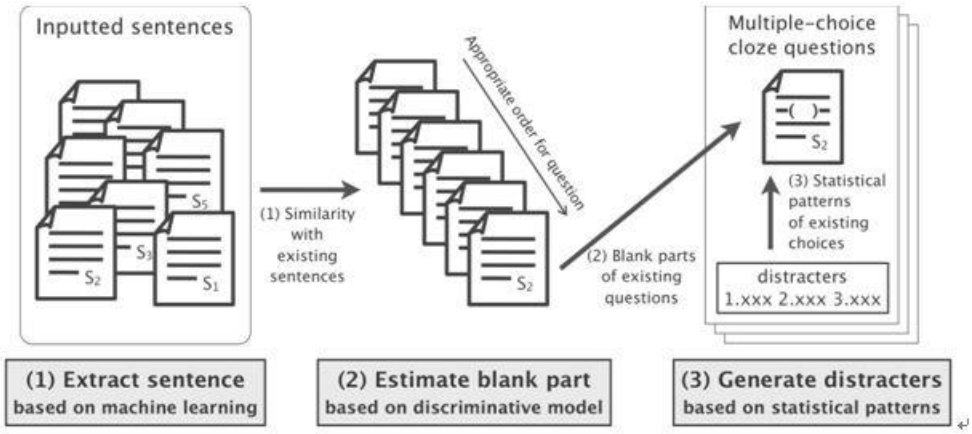

Figure 2. The proposed approach

Moreover, an appropriate blank part of each sentence depends on the structure of this sentence. Therefore, words other than verbs or adjectives should be selected as the blank part as well. In our method, various word classes are determined as the blank part based on a discriminative model, in which blank parts of existing questions are used for specifying those of the sentences being inserted (Figure 2 (2)).

Generating distracters is also an important issue in the automatic generation of multiple-choice cloze questions. Brown et al. proposed automatic generation methods for six types of questions (Brown, Frishkoff, & Eskenazi, 2005). One type of questions is the

multiple-choice cloze question and its distracters are generated by acquiring related words from the WordNet. Coniam developed an automatic generation method for multiple-choice cloze questions which determines words whose part-of-speech (POS) tags and frequencies are the same as those of the correct choice as distracters (Coniam, 1997). In these methods, only questions that ask vocabularies are being generated. Experts select sentences, blank parts, and distracters empirically to produce questions for various word types. Such knowledge can be found in existing questions. In our research, machine learning and statistical patterns are introduced to extract such heuristic knowledge for generating distracters (Figure 2 (3)). Appropriate sentences, blank parts, and distracters for a given text are then determined based on this knowledge.

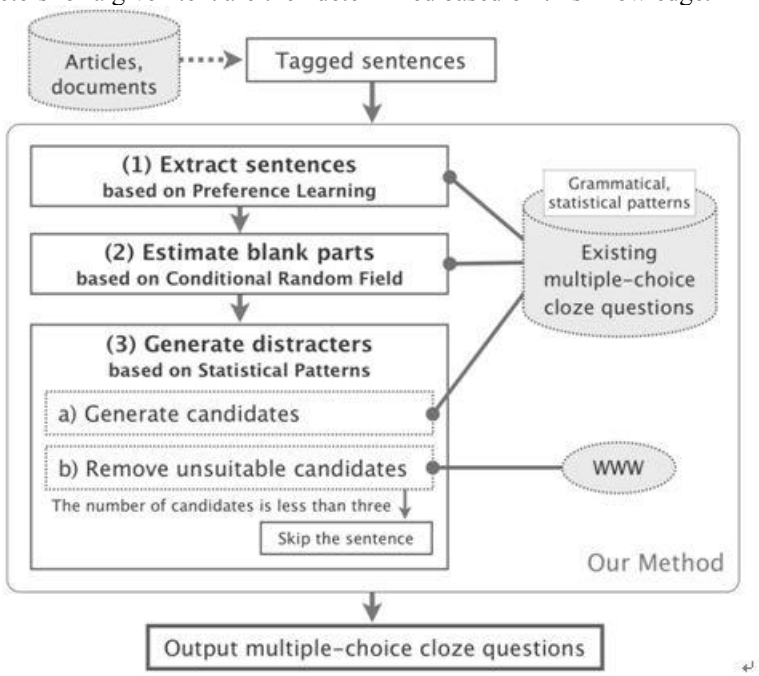

Figure 3. The flow towards generating multiple-choice cloze questions

Figure 3 illustrates a flow for generating multiple-choice cloze questions. Firstly, after *Penn Treebank II tags* were attached to all sentences in the text by *Postagger* (Tsuruoka & Tsujii, 2005), the system extracts some sentences that are appropriate for the multiple-choice cloze questions. In this phase, sentences are extracted from text using Preference Learning. Preference Learning is a method for classifying samples by Preference calculated according to similarity among samples. In our approach, existing questions are defined as positive samples and words and POS tags of existing sentences are learned.

Secondly, the system estimates a blank part using *Conditional Random Field*. Conditional Random Field (CRF) is a framework for building discriminative probabilistic models to segment and label sequence data (Lafferty, McCallum & Pereira, 2001). Hoshino et al. proposed a generation method for multiple-choice cloze questions based on a machine learning approach (Hoshino & Nakagawa, 2005). In their approach, each word which was an original blank part in existing questions was defined as a positive sample and other words in the question were determined as positive/negative samples based on a semi-supervised learning. Positions of positive/negative samples were then learned using a k-nearest neighbor (kNN) classifier. However, their methods cannot learn

the order of words and POS tags. The blank part is usually determined empirically by experts depending on the sequence of the sentence. In our approach, based on the CRF, sequences of words and POS tags and position of blank parts in the sequence are learned.

Thirdly, the system generates distracters. In this phase, the candidates for distracters are generated based on statistical patterns of existing multiple-choice cloze questions. The candidates and their adjacent words are searched through the web for the purpose of finding inappropriate candidates that can form a correct sentence. Based on the search results, the candidates that are often seen in the documents on the web are eliminated. If the number of candidates is less than three, the system gives up using this sentence.

3. Generation Methods of Questions

3.1. Extracting Sentences Based on Preference Learning

In order to extract appropriate sentences from texts based on their structures, words and POS tags in existing multiple-choice cloze questions are learned using *Preference Learning*. For the questions asking the usages of words, the sentences that contain the same words as the exiting questions are required. For the questions asking the grammar knowledge, the sentences that have a similar grammatical structure are appropriate. Therefore, in the training phrase, Preference Learning is carried out using words and POS tags emerging in existing multiple-choice cloze questions. In the generating phase, words and POS tags of each sentence in a text prepared by students are inserted and sentences are returned in the order of appropriateness. We make use of *Ranking Voted Perceptron* proposed by Collins et al. (Collins & Duffy, 2007), which is an online algorithm for Preference Learning.

The training algorithm is shown in Figure 4. $\mathbf{x}_{i0},\ldots,\mathbf{x}_{iN}$ are sentences which characterize existing question i. Sentence $\mathbf{x}_{i0}$ is a positive sample which is an existing question i with its blank part filled with the correct choice and sentences $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN}$ are the candidate samples which are extracted from other texts. *Similarity*($\mathbf{x}_{ij}$, $\mathbf{y}$) indicates the similarity of words and grammatical structures between sentence $\mathbf{x}_{ij}$ and sentence $\mathbf{y}$, which is calculated as Equation 1. *Score*($\mathbf{x}_{ij}$, $\mathbf{y}$) is determined by the ratio of the same words and the same word classes that is defined by the number of the same words in two sentences "*unigram*($\mathbf{x}_{ij}$, $\mathbf{y}$)" and that of the same POS tags "*posunigram*($\mathbf{x}_{ij}$, $\mathbf{y}$)" as Equation 2. If a sentence $\mathbf{y}$ is similar to $\mathbf{x}_{i0}$, the *Preference*($\mathbf{y}$) gets larger. Parameter $\boldsymbol{\alpha}_{ij}$ indicates the weight. If *Preference*($\mathbf{y}$) of candidate sentence $\mathbf{y}$ is larger than those of positive sentences, value of $\boldsymbol{\alpha}_{ij}$ is 1. Therefore, *Preference*($\mathbf{y}$) of each positive sample is adjusted in a manner that it does not make the candidate samples large.

Algorithm 1 Training algorithm of Ranking Voted Perceptron

$Preference(\mathbf{y}) = \sum_{ij} \alpha_{ij}(Similarity(\mathbf{x}_{i0},\mathbf{y}) - Similarity(\mathbf{x}_{ij},\mathbf{y}))$;
Set dual parameter $\alpha_{ij} \leftarrow 0$;
for $i = 0$ to n **do**
  if $\{\arg\max_{j=0\ldots N} Preference(\mathbf{x}_{ij})\} \neq 0$ **then**
    $\alpha_{ij} \leftarrow \alpha_{ij} + 1$;
  end if
end for

Figure 4. The training algorithm of Ranking Voted Perceptron

$$Similarity(\mathbf{x}_{ij}, \mathbf{y}) = Score(\mathbf{x}_{ij}, \mathbf{y}) / \sqrt{Score(\mathbf{x}_{ij}, \mathbf{x}_{ij}) \cdot Score(\mathbf{y}, \mathbf{y})} \quad (1)$$

$$Score(\mathbf{x}_{ij}, \mathbf{y}) = unigram(\mathbf{x}_{ij}, \mathbf{y}) + posunigram(\mathbf{x}_{ij}, \mathbf{y}) \quad (2)$$

When generating questions, *Preference*($\mathbf{z}_k$) for each sentence $\mathbf{z}_k$ ($k = 0 \ldots M$) which forms the text prepared by a student is calculated using trained parameter $\boldsymbol{\alpha}_{ij}$. Sentences are ranked according to the order of *Preference*($\mathbf{z}_k$). Figure 5 shows the sentences from articles in Associated Press. The numbers beside the sentences are their ranks calculated by Ranking Voted Perceptron. A sentence *A* does not form the sentence and a sentence *C* is a conversational sentence, so lower ranks are attached.

Prepared sentences	Rank
A. Umpires : Srinivas Venkataraghavan, India, and Daryl Harper, Australia.	5
B. The strategy has already cost them with Corey Collymore, who has a hamstring problem.	2
C. "Tung hasn't listened enough," said 36-year-old businessman Steve Lee.	4
D. It was not immediately clear whether anyone had claimed responsibility for the attack.	1
E. The plane was scheduled to fly back with passengers from India later Thursday.	3

Figure 5. Execution results of Ranking Voted Perceptron

3.2. Estimating Blank Part Based on Conditional Random Field

A sentence consists of a sequence of words with POS tags. The effective blank part for the sentence is determined by the words and its grammatical sentence. So, the determination of a blank part is interpreted as labeling the "blank part" to sequences of words and POS tags using *named entity extraction*. In the training phrase, sequences of words and POS tags with their named entities in existing multiple-choice cloze questions are learned. In the generating phase, an arbitrary tagged sentence is inserted and marginal probabilities of the named entity for each word are returned.

In our approach, CRF is introduced to attach labels to words of the sentence. A blank part is defined as the named entity in a sequence of words and represented by IOB2 format (Sang & Veenstra, 1999). In IOB2 format, three tags, such as "I", "O", and "B", are prepared. If a word in a sentence is a start of the blank part, "B" tag is given to the word. If the blank part consists of several words and a word is not the first word of the blank part, "I" tag is attached to it. On the other hand, if a word is not included in a blank part, "O" tag is given. For example, if the question *"His doctor urged him to ( ) doing hard exercise."* with its answer *"give up"* is given, IOB2 tags for each word are shown in Figure 6.

In the training phase, sequences of words, POS tags, IOB2 tags, and relations between sequences are trained using *CRF++* (Kudo, 2007). The CRF++ is used for implementation of the CRF. In generating questions, the system determines blank parts by estimating probabilities of their IOB2 tags. Figure 7 presents an example of the sentence *"This is the building where we had our first office."* The third column shows estimated tags and its marginal probability. The fourth, fifth, and sixth columns indicate marginal probability for each IOB2 tag. In this example, the given tag of "where" is "B" tag, so it becomes a blank part. If the estimated IOB2 tags for all words are "O" tag, the word whose marginal probability of "B" tag is the largest is determined as the blank part.

Word	POS tag	IOB2 tag
His	PRP$	O
doctor	NN	O
urged	VBD	O
him	PRP	O
to	TO	O
give	VB	B
up	RP	I
doing	VBG	O
hard	JJ	O
exercise	NN	O
.	.	O

Figure 6.
IOB2 tags for words

			IOB2 tag and marginal probability		
Word	POS tag	Estimated tag	B tag	I tag	O tag
This	DT	O/0.990963	B/0.008806	I/0.000232	O/0.990963
is	VBZ	O/0.983980	B/0.015443	I/0.000577	O/0.983980
the	DT	O/0.989846	B/0.009921	I/0.000233	O/0.989846
building	NN	O/0.997618	B/0.002103	I/0.000279	O/0.997618
where	WRB	B/0.870146	B/0.870146	I/0.000537	O/0.129317
we	PRP	O/0.946453	B/0.002297	I/0.051250	O/0.946453
had	VBD	O/0.901098	B/0.087598	I/0.011304	O/0.901098
our	PRP$	O/0.995546	B/0.000301	I/0.004153	O/0.995546
first	JJ	O/0.953272	B/0.046237	I/0.000492	O/0.953272
office	NN	O/0.991614	B/0.007394	I/0.000992	O/0.991614
.	.	O/0.999785	B/0.000204	I/0.000011	O/0.999785

Figure 7. The Output format of the test data

3.3. Generating Choices Based on Statistical Patterns

In order to generate candidates for distracters, relations between a correct choice and its distracters in existing questions have been investigated. Based on the result, two types of relations have been defined. In *Type I*, possible words in all choices are limited, which can be seen in questions whose blank parts consist of "Preposition or Subordinating conjunction", "Interrogative", "Coordinating conjunction", and "Modal auxiliary verb". For example, most distracters for the questions for "Interrogative" are "which", "what", "who", "when", "where". In this type of questions, candidates for distracters for each type are generated based on the proportion of the distracters' POS tags, and ratios of words in existing distracters. Table 1 shows an example of the proportion of distracters' POS tags and Table 2 indicates an example of frequencies of words in questions for "Interrogative" acquired from 350 multiple-choice cloze questions in TOEIC (Educational Testing Service, 2009) workbooks.

Table 1. The proportion of POS tags for existing distracters for "Interrogative"

All the same	1/3 different	2/3 different	All different
74%	16%	5%	5%

Table 2. An example of frequencies of "Interrogative" words and others

"Interrogative"	Frequency	Others	Frequency
which	14	that	3
what	10	as	1
who	9	because	1
when	5	if	1
where	5	so	1
…	…	though	1

On the other hand, specific patterns exist among choices for *Type II*. Questions for "Verb", "Noun", "Adjective", and "Adverb" correspond to this type. The patterns are classified into four patterns. The patterns and methods for generating distracters concern:

Conjugational word is the pattern in which distracters consist of conjugational words of the correct choice. Conjugational words are defined as the word whose word class is the same but tense or person is different from the original one. For example, if the correct choice is the verb "ask", distracters are "asked", "asking", "asks", etc. The

system obtains conjugational words based on a lexicon in which conjugations of verb are written manually.

Derivative word is the pattern in which distracters consist of derivative words of the correct choice. Derivative words are defined as the word which relates to the original word and whose word class is different from the original one. For example, if the correct choice is noun, "work", "worker", "works", "working", etc. are distracters. Derivative words are acquired by WordNet by searching the first 75% characters of the correct choice from lists of compound words.

Shape of word is the pattern in which string of characters in specific parts, such as prefix or suffix, is similar to that of correct one. For example, if "circulation" is the correct choice, "circumcision", "circumstance", "circus", etc. are the candidates. Such words can be found from WordNet by searching words that have the same prefix or suffix as a correct choice.

Meaning of word is the pattern in which distracters are synonym or antonym to correct choice. Synonym and antonym are acquired easily from WordNet.

Table 3 shows proportions of four patterns in 77 questions of "Verb" from 350 multiple-choice cloze questions in TOEIC workbooks. Based on the result, if the POS tag of a correct choice is "Verb", the pattern of "conjugational words" is applied by 62%.

Table 3. The proportions of the four patterns in "Verb"

Conjugational words	Derivative words	Shapes of words	Meanings of words
62%	-	18%	19%

After the candidates were generated, the unsuitable candidates are eliminated. In multiple-choice cloze questions, sentences with the correct choice should be correct and sentences with the distracters should not form correct sentence. So, the candidates that can form a correct sentence should be removed. In questions for *Type I* and "derivative word", "shape of word" and "meaning of word" of *Type II*, the candidates and adjacent words are searched through the web, which is as proposed in (Sumita, Sugaya & Yamamoto, 2004). The candidates and adjacent two words are searched with the *Google AJAX Search API* (Google 2010), and candidates with non-qero search results are regarded to be inappropriate and eliminated. Figure 8 shows the example of filtering the candidates for the sentence *"This is the building ( ) we had our first office."* whose correct choice is "*where*". In the Figure 8, candidates "which", "what", "who", and "when" are rejected since documents in the web contain these phrases and candidates "whom", "whose", and "how" are determined as distracters.

On the other hand, in questions for "conjugational word" in *Type II*, grammatical relations between the correct choice and the candidates are investigated. If the POS tag of a candidate is the same as that of a correct choice, then the candidate is inappropriate, because it may form the sentence whose structure is grammatically correct.

Sentence with correct word: "This is the building (where) we had our first office."

Search results with candidates:

1	"the building (which) we had" : **Hits**	5	"the building (whom) we had" : **No Hit**
2	"the building (what) we had" : **Hits**	6	"the building (whose) we had" : **No Hit**
3	"the building (who) we had" : **Hits**	7	"the building (how) we had" : **No Hit**
4	"the building (when) we had" : **Hits**	...	...

Generated incorrect choices: {whom, whose, how}

Figure 8. An example of filtering candidates

4. Implementation

The authors have constructed a web-based system for generating multiple-choice cloze questions, which is implemented by PHP and AJAX. Currently, learning data from 1560 questions in TOEIC workbooks are available.

Figure 9 and Figure 10 show the interface of our system. The student inserts the text from which he/she wants to generate questions in the entire text area in Figure 9. After pushing the generation button, the system automatically generates questions. The list of questions is shown in Figure 10. The questions are ordered by the appropriateness of the sentences, namely the question appearing at the top is generated from the most appropriate sentence.

Figure 9. The interface of the system

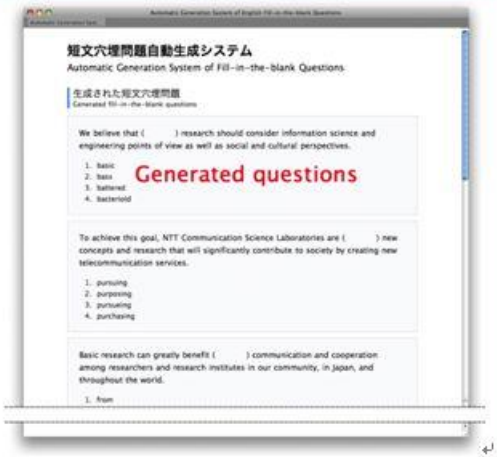

Figure 10. A screenshot on generated questions

5. Experiments

5.1. Correctness of the Extracted Sentences

The authors have evaluated the method for extracting sentences based on the Ranking Voted Perceptron. In this experiment, we have been applied 1560 multiple-choice cloze questions from TOEIC workbooks as positive samples, and 1560 sentences from Associated Press as candidate samples. 10-fold cross-validation was carried out: 1404 (9/10) positive samples and 1404 candidate samples built the basis for the training set and 156 (1/10) positive samples and 156 candidate samples built the basis for the test set.

Table 4 indicates an average proportion that positive samples were successfully ranked in the upper half of the candidate samples, namely the average proportion that candidate samples were ranked higher than 78. Sentences constructed with words or POS tags that were appeared frequently in positive samples were highly ranked. On the other hand, sentences including grammatical errors, conversational sentences and colloquialisms were ranked lower. Therefore, our extraction method is available for selecting appropriate sentences.

Table 4. A proportion that positive samples are ranked as the upper half

# of sentences	Average # of positive samples ranked upper half	Proportion of positive samples ranked in upper half
156	142.8	91.83%

5.2. Correctness of the Estimated Blank Part

The correctness of estimating blank part using Conditional Random Field was evaluated by comparing the detected blank parts with the original blank part of questions. A ten-fold cross-validation was carried out for 1560 questions in TOEIC workbooks: the training set was built on the basis of 1404 questions (9/10), and the test set was built on the basis of 156 questions (1/10). The following two methods were implemented and their results were compared with the result of the proposed system.

Leftmost verb The method determines the verb that appears firstly in a sentence as the blank part, which was proposed by Sumita et al (Sumita, Sugaya & Yamamoto, 2004). This method was applied to 1560 multiple-choice cloze questions.

Frequency of blank part's POS tags The number and the order of each POS tags in the training set are counted. The most frequent tag in the training set is determined as "blanking tag" and sentences in the test set are blanked based on the blanking tag and its position. A ten-fold cross-validation was executed for this method.

Table 5 shows the recall rates of the blank parts estimated by each method. Of all methods, the proposed method appears to be the most effective to estimate the original blank part. Moreover, the blank part that consists of various POS tags or more than two words were successfully selected. Therefore, the method is appropriate to estimate blank parts. Blank parts which are not estimated successfully can also make use of multiple-choice cloze questions although the recall rate of the proposed method is only 18.91%.

Figure 11 shows a failed example which estimated as *"The newscaster provided ( ) commentary on the tragedy during the hour-long broadcast."*, contrary to the original question *"The newscaster provided running ( ) on the tragedy during the hour-long broadcast."* The example suggests that the failed result which is different from the original blank part is also available as the blank part of a multiple-choice cloze question. The availability of the generated blank parts is evaluated in Section 5.3.

Table 5. Estimating methods and recalls

Methods	Recalls
Our method	18.91%
Leftmost verb	10.23%
Frequency of blank part's POS tags	8.46%

Figure 11. Examples on failures in estimation

5.3. Availabilities of Blank Part and Distracters

The authors have asked experts in English language to evaluate the availability of generated multiple-choice cloze questions, especially the blank parts and distracters. The order of sentences was not the target of this experiment, because it is difficult to attach an order to sentences manually. Experts are considered those who are researchers for teaching English language or experts in generating English questions, or who have lived overseas for a long time. All of them have gotten over 900 points in TOEIC. Twenty questions were generated from three types of texts acquired from the official pages: A. story of TV drama (six questions), B. message from the president of Nagoya University (seven questions), and C. history of FBI (seven questions). Experts were asked to answer the questionnaire on the quality of the blank parts and distracters. The questionnaire consists of the following items. In addition, experts were also required to comment their answers, especially if they chose *no*.

1. *Is the word selected as a blank part available for the multiple-choice cloze question? (Choose from yes/no/I cannot judge).*
2. *Is the sentence whose bank part is filled with the distracter grammatically incorrect? (Choose from yes/no/I cannot judge).*

For both items, the authors asked experts to answer only from the grammatical point of view and do not consider the quality of the questions. Three experts answered for the multiple-choice cloze questions for the texts A and B, and four experts answered for questions for the text C.

Table 6. Results for item1

Text type / No. of *no/I cannot judge*	A	B	C
0	4	3	3
1	1	3	2
2	1	1	2

Table 6 shows the results of item 1. The number of experts who answered *no* or *I cannot judge* has been counted. The blank parts of more than a half of the questions were evaluated as available for this question. Experts who answered *no* commented: "It is

difficult to select the answer for the blank part without sentences before and after the sentence." The current proposed method does not consider the meaning of the sentences. If the sentence contains pronouns and targets of pronouns are required for answering the question, the question cannot be answered without the former sentences. One solution for the problem is to eliminate sentences that have pronouns in the stage of selecting sentences.

Table 7. Results for item2

Text type / No. of *no/I cannot judge*	A	B	C
0	**12**	**18**	**6**
1	3	2	11
2	3	1	3
3	0	0	1
4			0

Table 7 shows the result for item 2. Distracters of 66.7% for text A and of 85.7% for text B are proved as grammatically incorrect. In addition, there were five questions whose three distracters are all evaluated as available by all experts. However, the distracters of more than 60 % for text C are decided as inappropriate by more than one of the experts. The sentences in text C contain several proper nouns, such as the name of a person. Such words prevent elimination of grammatically correct candidates as distracters when checking on the web. Namely, because of the proper nouns in text C, grammatical correct words were selected as distracters. In order to avoid such situation, proper nouns should be detected and changed to pronouns or general proper noun, such as John for indicating a person, in searching on the web. Moreover, experts commented: "For some multiple-choice cloze questions, the target knowledge to ask inferred from the blank part and from the distracters are not the same." In order to generate distracters whose target knowledge seems the same as that of the blank part, the types of distracters should be learned with the label of the blank part in the stage of deciding the blank part.

6. Conclusion

In this paper, the authors proposed a statistical method of generating multiple-choice cloze questions automatically. Based on the machine learning and statistical patterns of existing questions, the system is able to select sentences which are appropriate to multiple-choice cloze questions from texts and generate various types of blank parts with distracters. Based on the experimental results, the system is proved to select correct sentences and blank parts. In addition, generated blank parts and distracters are available for multiple-choice cloze questions.

In the future work, the authors firstly need to cope with problems revealed through the experiments, such as to eliminate sentences that contain pronoun in the selected sentence, and to cope with the proper noun in checking the candidates of distracters on the web whether they are grammatically incorrect or not. In addition, the current system is not able to generate distracters for the blank part which consists of more than two words. Further analysis of patterns of distracters in such questions in detail and the development of methods to generate such patterns of distracters are needed.

So far, the authors have focused on generating questions. However, the knowledge and the learning objectives vary among students. The target knowledge for asking a question is able to be characterized by the POS tags and words of the blank parts and distracters. Therefore, by utilizing such information, the authors will work on methods for applying generated questions according to the student's level of understanding and preferences.

References

1. Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005). Automatic Question Generation for Vocabulary Assessment. *Proc. of HLT/EMNLP '05*, 819-826.

2. Collins, M., & Duffy, N. (2007). New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron, *Proc. of 40th Annual Meeting of the Association for Computational Linguistics*, 263-270.

3. Coniam, D. (1997). A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests, *CALICO Journal*, 14 (2-4), 15-33.

4. Educational Testing Service (2009). *Test of English for International Communication*, http://www.ets.org/toeic.

5. Funaoi, H., Akiyama, M., & Hirashima, T. (2006). Automatic Creation of Mis-choices and Comments for Multiple Choice Question Based on Problem Solving Model, *ICCE 2006 Workshop Proc. of Problem-Authoring, -Generation and -Posing in a Computer-Based Learning Environment*, 49-54.

6. Google (2010). *Google AJAX Search API*, http://code.google.com/intl/ja/apis/ajaxsearch/.

7. Hoshino, A., & Nakagawa, H. (2005). A Realtime Multiple-choice Question Generation for Language Testing: A Preliminary Study, *Proc. of the 2nd Workshop on Building Educational Applications Using NLP*, 17-20.

8. Kudo, T. (2007). *CRF++: Yet Another CRF toolkit*, http://crfpp.sourceforge.net/.

9. Lafferty, J., McCallum, A., & Pereira. F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. of ICML 2001*, 282-289.

10. Lin, Y. C., Sung, L. C., & Chen, M. C. (2007). An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding, *ICCE 2007 Workshop Proc. of Modeling, Management and Generation of Problems / Questions in eLearning*, 137-142.

11. Mitkov, R., & Ha, L. A. (2003). Computer-Aided Generation of Multiple-Choice Tests, *Proc. of HLT-NAACL '03 Workshop: Building Educational Applications Using Natural Language Processing*, 17-22.

12. Sang, T. K. & Veenstra, J. (1999). Representing Text Chunks, *Proc. of EACL'99*, 173-179.

13. Sumita, E., Sugaya, F., & Yamamoto, S. (2004). Automatic Generation Method of a Fill-in-the-blank Question for Measuring English Proficiency, *Technical report of IEICE*, 104 (503), 17-22 [in Japanese].

14. Tsuruoka, Y., & Tsujii, J. (2005). Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data, *Proc. of HLT/EMNLP 2005*, 467-474.