Kong, S.C., Ogata, H., Arnseth, H.C., Chan, C.K.K., Hirashima, T., Klett, F., Lee, J.H.M., Liu, C.C., Looi, C.K., Milrad, M., Mitrovic, A., Nakabayashi, K., Wong, S.L., Yang, S.J.H. (eds.) (2009). *Proceedings of the 17th International Conference on Computers in Education [CDROM]*. Hong Kong: Asia-Pacific Society for Computers in Education.

An Automatic Generation of Multiple-choice Cloze Questions Based on Statistical Learning

Takuya GOTO[a], Tomoko KOJIRI[a], Toyohide WATANABE[a], Tomoharu IWATA[b] and Takeshi YAMADA[b]

[a]*Graduate School of Information Science, Nagoya University, Japan*

[b]*NTT Communication Science Laboratories, Japan*

tgoto@watanabe.ss.is.nagoya-u.ac.jp

Abstract: Since English expressions are different according to the genre, it is important for users to study with 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 users want to learn. On the other hand, when producing English questions, enough grammatical knowledge and vocabulary are needed, so it is difficult for inexperts to prepare English questions by themselves. In this paper, we propose an automatic generation system of multiple-choice cloze questions from English texts. Experiential 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 available for selecting appropriate sentences and appropriate for estimating the blank part.

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

1. Introduction

With the spread of e-learning in English, users are able to study with various questions provided through the web. Most of existing questions have been produced by experts. However, expression of English is different according to the genre, so it is important to tackle questions from texts in various genres. In addition, users are highly motivated if questions of interesting genres are generated automatically from various texts, such as articles, research papers, and web documents, that are selected by users. A lot of automatic generation systems of various types of questions were proposed [1, 2]. However, these researches focused on generating questions from single sentence.

In this paper, we propose an automatic generation system of multiple-choice cloze questions from texts. For multiple-choice cloze questions, grammatical structures and vocabularies that compose sentences determine the levels of questions. Therefore, to select sentences which consist of words or word classes appearing frequently from texts (*appropriate sentence*) is important. In addition, a blank part of the questions indicates target knowledge to make user acquire and appropriate blank part depends on the structure of the sentence. Of course, distracters also affect to difficulties of questions. 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. Experts usually select appropriate sentences and determine a blank part and distracters that are effective for the sentences according to their experiences. In order to acquire experts' heuristic knowledge of generating questions, our system extracts vocabulary and grammatical features of existing multiple-choice cloze questions based on machine learning and statistical approaches, and applies them for preparing new questions from existing texts.

2. Automatic Generation of Multiple-choice Cloze Questions

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-

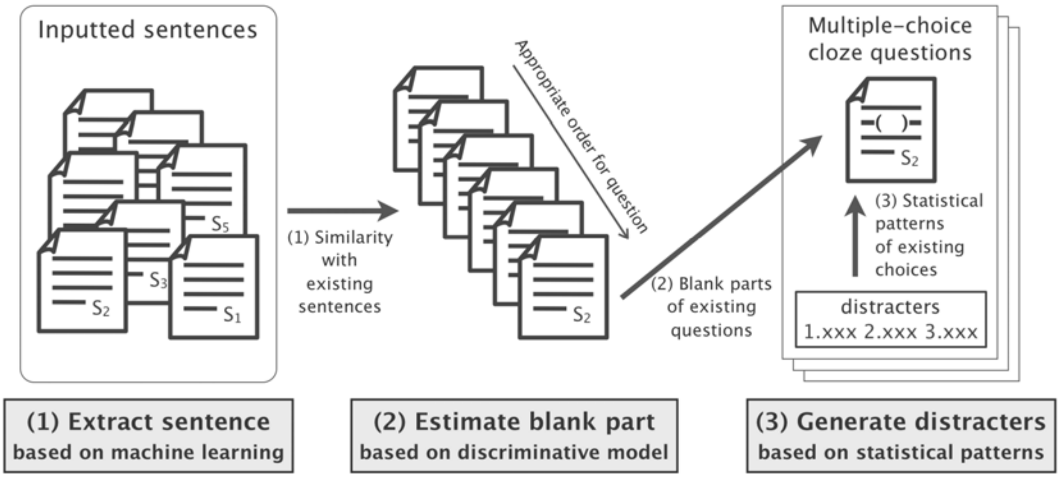

Figure 1. ***Our approach***

choice cloze questions have been proposed [3, 4, 5, 6]. Sumita et al. proposed an automatic generation method of multiple-choice cloze questions for measuring English proficiency [3]. In this method, leftmost single verb is selected as a blank part. Yi-Chein et al. also constructed an automatic generation system of multiple-choice cloze question [4]. They focused on questions for asking "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 corresponding phrase through the web. One of the problems of these researches is that systems do not validate whether given sentences are "appropriate" to multiple-choice cloze questions [3, 4]. Sentences are sometimes too simple or too complicated to be the questions. In our approach, sentences which are similar to sentences in existing questions are extracted in "appropriate" order to questions by learning emerging words and grammatical patterns of existing questions based on a machine learning (Figure 1 (1)).

Moreover, an appropriate blank part of each sentence depends on the structure of the sentence. Therefore, words other than verb or adjective should be selected as the blank part as well. In our method, various word classes are determined as the blank part based on discriminative model, in which blank parts of existing questions are used for specifying those of inputted sentences (Figure 1 (2)).

To generate distracters are also important issue in automatic generation of multiple-choice cloze questions. Brown et al. proposed automatic generation methods of six types of questions [5]. 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 of multiple-choice cloze questions which determines words whose part-of-speech (POS) tags and frequencies are the same as that of the correct choice as distracters [6]. In these methods, only questions that ask vocabularies are 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 1 (3)). Appropriate sentences, blank parts, and distracters for given text are then determined based on the knowledge.

Figure 2 is a flow of generating multiple-choice cloze questions. Firstly, after *Penn Treebank II tags* were attached to all sentences in the text by *Postagger* [9], 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 [7]. Hoshino et al. proposed a generation method of multiple-choice cloze questions based on machine learning approach [8]. In their approach, each word which is an original blank part in existing questions was defined as positive samples 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 kNN classifier. However, their method cannot learn order of words and POS tags. The blank part is usually determined empirically by experts depending on a 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.

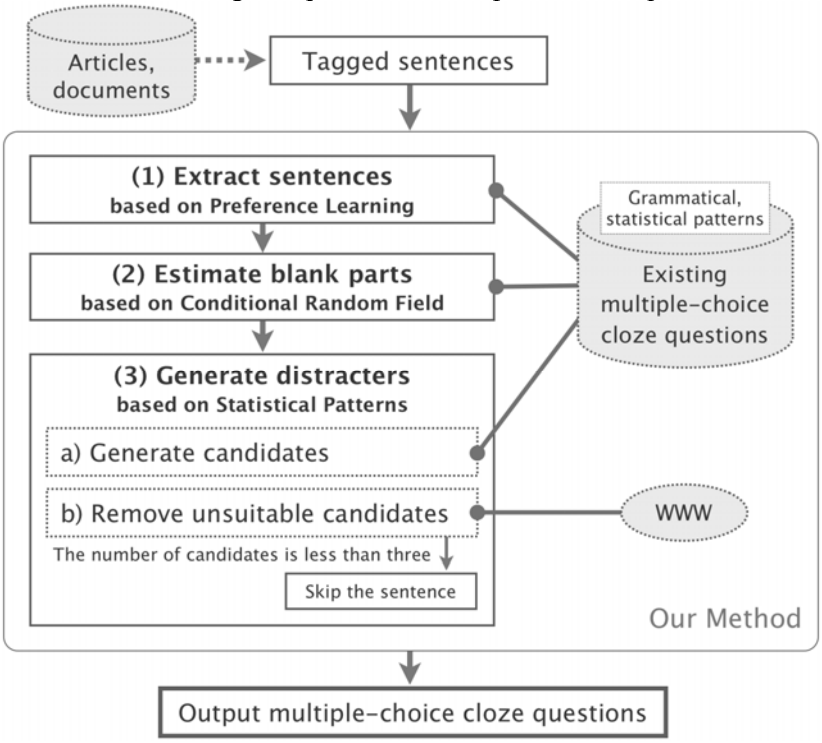

Figure 2. ***Flow of generating multiple-choice cloze questions***

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 $\{\arg\max_{j=0...N} Preference(\mathbf{x}_{ij})\} \neq 0$ **then**
    $\alpha_{ij} \leftarrow \alpha_{ij} + 1$;
  end if
end for

Thirdly, the system generates distracters. In this phase, candidates of distracters are generated based on statistical patterns of existing multiple-choice cloze questions. Candidates and their adjacent words are searched through the web for the purpose of finding candidates that can form a correct sentence. Based on the search results, candidates that are often seen in the documents on the web are eliminated. If the number of candidates is less than three, the system abandons to make question using the 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*. 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 users are inputted and sentences are outputted in the order of appropriateness. We make use of *Ranking Voted Perceptron* proposed by Collins et al. [10], which is online algorithm of Preference Learning.

The training algorithm is shown in Algorithm 1. $\mathbf{x}_{i0},\ldots,\mathbf{x}_{iN}$ are sentences for learning characteristic of existing question *i*. Sentences $\mathbf{x}_{i0}$ 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 candidate samples which are extracted from other texts. *Similarity*($\mathbf{x}_{ij}$, $\mathbf{y}$) indicates 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

Kong, S.C., Ogata, H., Arnseth, H.C., Chan, C.K.K., Hirashima, T., Klett, F., Lee, J.H.M., Liu, C.C., Looi, C.K., Milrad, M., Mitrovic, A., Nakabayashi, K., Wong, S.L., Yang, S.J.H. (eds.) (2009). *Proceedings of the 17th International Conference on Computers in Education [CDROM]*. Hong Kong: Asia-Pacific Society for Computers in Education.

word types 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

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 3. Execution result of Ranking Voted Perceptron

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

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

larger. Parameter $\boldsymbol{\alpha}_{ij}$ indicate weight. If *Preference*($\mathbf{y}$) of candidate sentence $\mathbf{y}$ is larger than those of positive sentences, values of $\boldsymbol{\alpha}_{ij}$ are 1. Therefore, *Preference* of each positive sample is adjusted so as not to make those of candidate samples large.

When generating questions, *Preference*($\mathbf{z}_k$) for each sentence $\mathbf{z}_k$ ($k = 0 \ldots M$) which forms the text prepared by a user is calculated using trained parameter $\boldsymbol{\alpha}_{ij}$. Sentences are ranked according to the order of *Preference*($\mathbf{z}_k$), and sentence $\mathbf{z}_k$. Figure 3 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.

3.2 Estimating Blank Part Based on Conditional Random Field

A sentence consists of a sequence of words with POS tags. 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 inputted and marginal probabilities of the named entity for each word are outputted.

In our approach, CRF is introduced to attach label to words of the sentence. A blank part is defined as the named entity in a sequence of words and represented by IOB2 format [11]. In IOB2 format, three tags, such as "I", "O", and "B", are prepared. If a word in a sentence is 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 4.

In the training phase, sequences of words, POS tags, IOB2 tags, and relations between sequences are trained using *CRF++* [12]. 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 5 is 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 forth, fifth, and sixth column indicates marginal probability for each IOB2 tag. In this example, given tag of "where" is "B", 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.

3.3 Generating Choices Based on Statistical Patterns

In order to generate candidates of distracters, relations between a correct choice and distracters of existing questions are investigated. Based on the result, two types of relations are defined. In *Type I*, emerging 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 of the questions for "Interrogative" are "which", "what", "who", "when", "where". In this type of question, candidates of distracters for each type are generated based on proportion of POS tags and ratios of words in existing distracters. Table 1 shows an example of proportion of POS tags and Table 2

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 4. ***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 5. ***Output format of test data***

Table 1. ***Proportion of POS tags of questions for "Interrogative"***

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

Table 2. ***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

Table 3. ***Proportions of four patterns in "Verb"***

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

indicates an example of frequencies of words in questions for "Interrogative" acquired from 350 multiple-choice cloze questions in TOEIC [13] workbooks.

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

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 original one. For example, if the correct choice is 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 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 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 candidates. Such words can be found from WordNet by searching words that have same prefix or suffix as 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.

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 6. ***Example of filtering candidates***

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 correct choice is "Verb", the pattern of "conjugational words" is applied by 62%.

After candidates are generated, unsuitable candidates are eliminated. In multiple-choice cloze questions, sentences with the correct choice should be correct and sentences with distracters must not form correct sentence. So, the candidates that could be a correct sentence should be removed. In questions for *Type I* and "derivative word", "shape of word" and "meaning of word" of *Type II*, candidates and adjacent words are searched through the web, which is proposed in [3]. Candidates and adjacent two words are searched with the *Google AJAX Search API* [14], and candidates whose search results are not 0 are regarded to be inappropriate and eliminated. Figure 6 shows the example of filtering candidates for the sentence *"This is the building (   ) we had our first office."* whose correct choice is "*where*". In the Figure 6, 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 candidates are investigated. If the POS tag of a candidate is the same as that of correct choice, that candidate is inappropriate, because it may form the sentence whose structure is grammatically correct.

4. Implementation

We have constructed the web-based system for generating multiple-choice cloze questions, which is implemented by PHP and AJAX. Currently, we have prepared learning data from 1560 questions in TOEIC workbooks.

Figures 7 and 8 show the interface of our system. The user inputs the text from which he/she wants to generate questions in the inputting text area in Figure 7. When he/she pushes generation button, the system automatically generates questions and show lists of questions in Figure 8. In the list, questions are ordered by their appropriateness, namely the first question is the most appropriate sentence.

5. Experiments

5.1 Experiment of Extracting Sentences

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

Table 4 indicates an average proportion that positive samples were successfully ranked in upper half of the candidate samples, namely the average proportion that candidates 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.

Figure 7. ***Interface of the system***

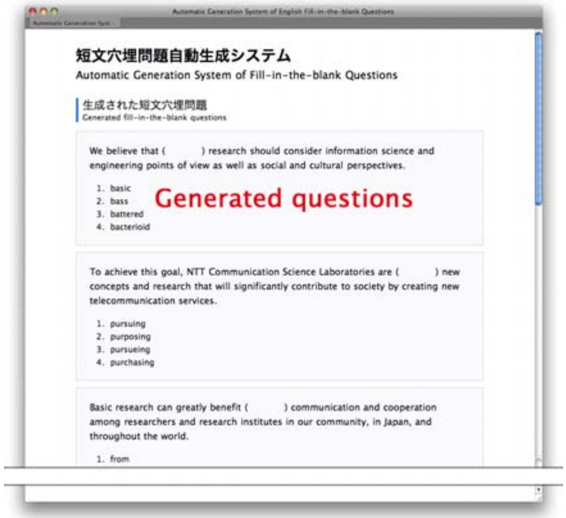

Figure 8. ***Screenshot after generating questions***

5.2 Experiment of Estimating Blank Part

Effectiveness of estimating blank part using Conditional Random Field was evaluated by comparing detected blank parts with original blank part of questions. 10-fold cross-validation was carried out for 1560 questions in TOEIC workbooks: training set was 1404 questions (9/10) and test set was 156 questions (1/10). Following two methods were implemented and their results were compared with that of our 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 [3]. This was applied to 1560 multiple-choice cloze questions.

Frequency of blank part's POS tags The number and the order of each POS tag in training set are counted. The most frequent tag in training set is determined as “blanking tag” and sentences in test set are blanked based on the blanking tag and its position. 10-fold cross-validation was executed for this method.

Table 5 shows the recall rates of blank parts estimated by each method. Of all methods, our method is the most effective to estimate original blank part. Moreover, the blank part that consists of various POS tags or more than two words were successfully selected. Therefore, our 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 our method is only 18.91%. Figure 9 is 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.

6. Conclusion

In this paper, we proposed a statistical method of generating multiple-choice cloze questions automatically. Based on the machine learning and statistical patterns of existing questions, our system is able to select sentences which are appropriate to multiple-choice cloze questions from texts and generate various types of blank part with distracters.

Currently, the system is not able to treat the blank part which consists of more than two words. We must analyze patterns of distracters in such questions in detail. In addition, an evaluation of generated multiple-choice cloze questions is not conducted. We need to ask native speakers to evaluate qualities of generated questions.

We have focused on generating questions. However, acquired knowledge and learning objective are different among users. Therefore, in our future, POS tags of blank parts and distracters should be determined according to users so as to generate questions that are appropriate for users' understanding levels and preferences.

Table 4. ***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%

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%

Word	POS tag	IOB2 tag (original)	IOB2 tag/% (estimated)
The	DT	O	O/0.99
newscaster	NN	O	O/0.99
provided	VBD	O	O/0.92
running	**VBG**	**O**	**B/0.62**
commentary	**NN**	**B**	**O/0.94**
on	IN	O	O/0.96
the	DT	O	O/0.99
tragedy	NN	O	O/0.98
during	IN	O	O/0.91
the	DT	O	O/0.99
hour-long	JJ	O	O/0.98
broadcast	NN	O	O/0.98
.	.	O	O/0.99

Figure 9. ***Failed example in estimation***

References

[1] Funaoi, H., Akiyama, M., and 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*, pp. 49-54.

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

[3] Sumita, E., Sugaya, F., and Yamamoto, S. (2004). Automatic Generation Method of a Fill-in-the-blank Question for Measuring English Proficiency, *Technical report of IEICE*, Vol. 104, No. 503, pp. 17-22 [in Japanese].

[4] Lin, Y. C., Sung, L. C., and 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*, pp. 137-142.

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

[6] Coniam, D. (1997). A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests, *CALICO Journal*, Vol. 14, No. 2-4, pp. 15-33.

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

[8] Hoshino, A., and 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*, pp. 17-20.

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

[10] Collins, M., and 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*, pp. 263-270.

[11] Sang, T. K. and Veenstra, J. (1999) Representing Text Chunks, *Proc. of EACL'99*, pp. 173-179.

[12] Kudo, T. (2007). CRF++: Yet Another CRF toolkit, http://crfpp.sourceforge.net/.

[13] Test of English for International Communication, http://www.ets.org/toeic.

[14] Google AJAX Search API, http://code.google.com/intl/ja/apis/ajaxsearch/.