Hirashima,T. et al. (Eds.) (2010). Workshop Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.

Statistical Learning-based Approach for Automatic Generation System of Multiple-choice Cloze Questions

Tomoko KOJIRI*, Takuya GOTO, Toyohide WATANABE^{a*}, Tomoharu IWATA^{b} & Takeshi YAMADA^{c}

[a]*Graduate School of Information Science, Nagoya University, Japan*
[b]*NTT Communication Science Laboratories, Japan*
[c]*NTT Science and Core Technology Laboratory Group, Japan*
*kojiri@nagoya-u.jp

Abstract: 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.

Keywords: automatic question generation, multiple-choice cloze question

1. Introduction

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, most of the existing questions have been produced by experts based on their heuristic knowledge. Therefore, it is difficult to define the generation knowledge of multiple-choice cloze questions for individual genres.

In this paper, we propose a system for the automatic generation of multiple-choice cloze questions from texts. For generating questions from texts, system 1) selects sentences that are appropriate to form questions, 2) determines blank part which is target knowledge to ask, and 3) generates distracters that characterize difficulties of questions. The generation knowledge for each generation stage can be observed by the existing questions. Therefore, in our approach, characteristics of existing questions for each generation stage are extracted based on statistical learning. Then, by applying extracted characteristics to the inputted text, the system selects the sentence, determines words for blank part, and selects words from various dictionaries as distracters.

2. Approach

Various automatic generation systems of multiple-choice cloze questions have been proposed [1] [2]. One of the problems of these researches is that systems do not validate whether given sentences are “appropriate” as multiple-choice cloze questions. In our approach, sentences that are similar to sentences in existing questions are extracted in an “appropriate” order as questions using Preference Learning. Preference Learning is a method for classifying samples by Preference calculated according to similarity among samples. Existing questions are defined as positive samples, and words and Part-of-speech (POS) tags of existing sentences are learned. Based on the approach, sentences whose words and grammatical patterns are similar to the existing questions are selected as appropriate sentences.

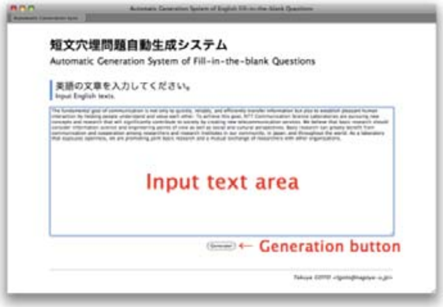
An appropriate blank part of each sentence depends on the structure of this sentence. In our approach, the system estimates a blank part using Conditional Random Field (CRF). CRF is a framework for building discriminative probabilistic models to segment and label sequence data [3]. From existing questions, sequences of words and POS tags, and position of blank parts in the sequence are learned based on CRF. According to this approach, various word classes are determined as the blank part.

Generating distracters is also an important issue in the automatic generation of multiple-choice cloze questions. Relations between a correct choice and its distracters in existing questions have been investigated statistically and methods for generating each type of distracters are defined. The methods include searching appropriate types of words from WordNet or a lexicon of conjugations of verb. The candidates of distracters 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.

3. Prototype System

Figure 1 and Figure 2 are the interface of our system[6]. The system is constructed as a web-based system, which is implemented by PHP and AJAX. Currently, learning data are 1560 questions from TOEIC workbooks. Therefore, questions that have similar characteristics to TOEIC questions are generated.

The user inserts the text from which he/she wants to generate questions in the entire text area in Figure 1. After pushing the generation button, the system automatically generates questions and displays list of questions as shown in Figure 2. The questions are ordered by the appropriateness of the sentences, namely the question appearing at the top is generated from the most appropriate sentence.



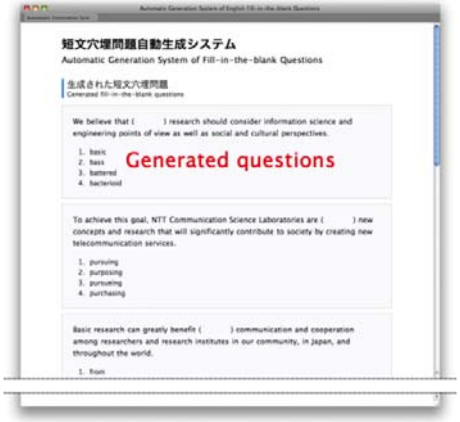
短文穴埋問題自動生成システム

Automatic Generation System of Fill-in-the-blank Questions

英語の文章を入力してください。

Input English texts.

Figure 9. The interface of the system



短文穴埋問題自動生成システム

Automatic Generation System of Fill-in-the-blank Questions

生成された短文穴埋問題

Generated fill-in-the-blank questions

We believe that (   ) research should consider information science and engineering points of view as well as social and cultural perspectives.

1. basic
2. bass
3. battered
4. bacteriod

To achieve this goal, NTT Communication Science Laboratories are (   ) new concepts and research that will significantly contribute to society by creating new telecommunication services.

1. pursuing
2. purposing
3. pursueing
4. purchasing

Basic research can greatly benefit (   ) communication and cooperation among researchers and research institutes in our community, in Japan, and throughout the world.

1. from

Figure 10. A screenshot on generated questions

References

[1] 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.

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

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

[6] The URL of the system: http://www.watanabe.ss.is.nagoya-u.ac.jp/ja/pickup/magic/