Automatic Generation of English Cloze Questions Based on Machine Learning

Tomoharu Iwata[†], *Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, and Takeshi Yamada*

Abstract

We are developing a system that automatically generates multiple-choice cloze questions in English, which we call MAGIC. By using machine learning techniques, MAGIC extracts the characteristics of manually generated questions and selects appropriate sentences for questions, determines words to be blanked, and generates multiple choices. This project is part of collaborative research and development between NTT and Nagoya University.

1. Introduction

With the development of information and communications technology (ICT), e-learning has become popular. One of its main benefits is that each user can learn in accordance with his or her own achievement level, interests, and pace. However, we need to prepare a huge quantity of learning materials in order to suit the levels and interests of a wide variety of users.

To resolve this problem, we are researching and developing systems that generate learning materials automatically. By fusing e-learning technology studied in Nagoya University and machine learning technology studied in NTT, we have developed a system that automatically generates multiple-choice cloze questions for English, which we call MAGIC (multiple-choice automatic generation system for cloze questions) [1]. Some examples of multiple-choice cloze questions are shown in **Fig. 1**. There were several reasons for focusing on this type of English questions. First, a lot of Japanese learn English, and English learning is in demand not only in Japan but all over the world. Second, this question type is common and used in TOEIC (test of English for international communication) and university entrance examinations.

MAGIC takes English sentences as input. Users may input just one sentence or multiple sentences from newspaper articles and novels. MAGIC sorts the sentences in order of appropriateness for English questions and outputs sentences containing blanks and four options for each blank. Figure 1 shows an example of MAGIC's input and output.

MAGIC can be used for a variety of ways. First, by entering English sentences that match their own interests, users can learn while having fun. For example, users who like football can learn English using football articles, and users can learn using novels by authors that they like. Second, users can learn a type of English that suits their study purpose. For example, users can learn business English by using business news articles, scientific English by using scientific papers, and travel English by using travel guidebooks. Third, users can overcome their weak points by generating and learning questions that test them.

The interests and goals for learning depend on users, and they may change over time. Therefore, it is difficult to prepare learning materials that meet the diverse interests and goals of all users.

† NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan

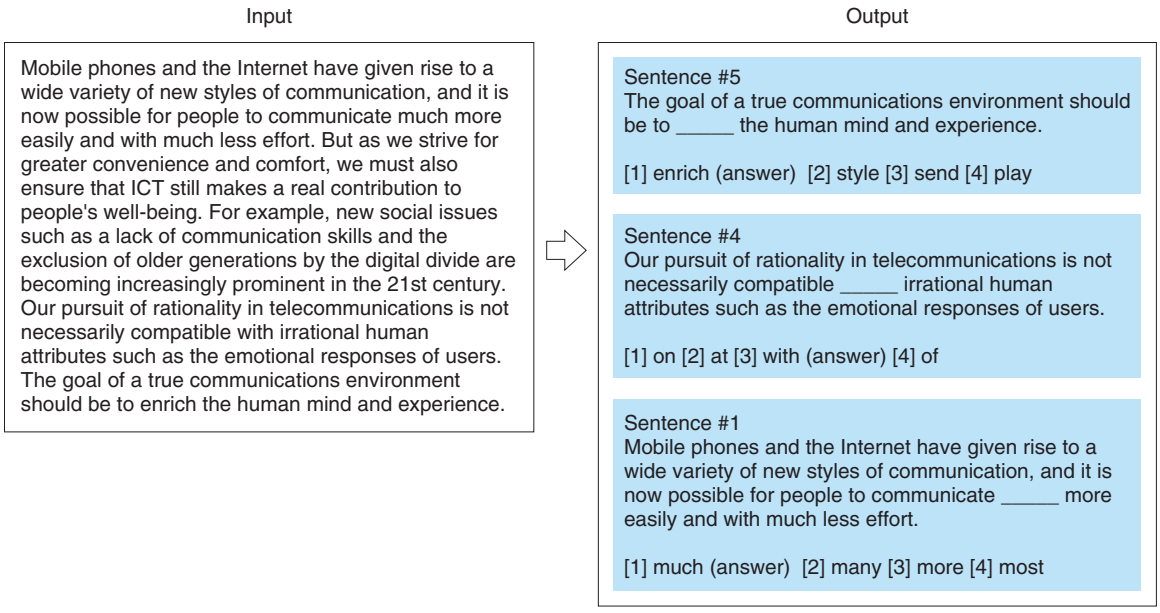


Fig. 1. Input and output of MAGIC.

2. Rule extraction using machine learning

MAGIC has three components, which 1) sort sentences in order of appropriateness for English questions, 2) determine words to be blanked, and 3) generate multiple choices. The procedures of MAGIC are shown in **Fig. 2**. For these three components, we need rules for evaluating the appropriateness of English questions, selecting words to be blanked, and generating choices, respectively. However, such rules are not available, and humans do not generate questions by following explicit rules.

MAGIC can automatically extract these rules by using machine learning techniques, which find statistical rules from given training data by using computers. The training data for MAGIC are manually generated English questions. By analyzing hand-made questions, we can extract rules for sorting sentences and generating blanked sentences with their multiple choices.

By changing the training data, we can change the question generating characteristics. For example, by using TOEIC questions for the training data, we can generate questions that are similar to TOEIC questions. We can also increase the level of difficulty by using high-level questions for the training.

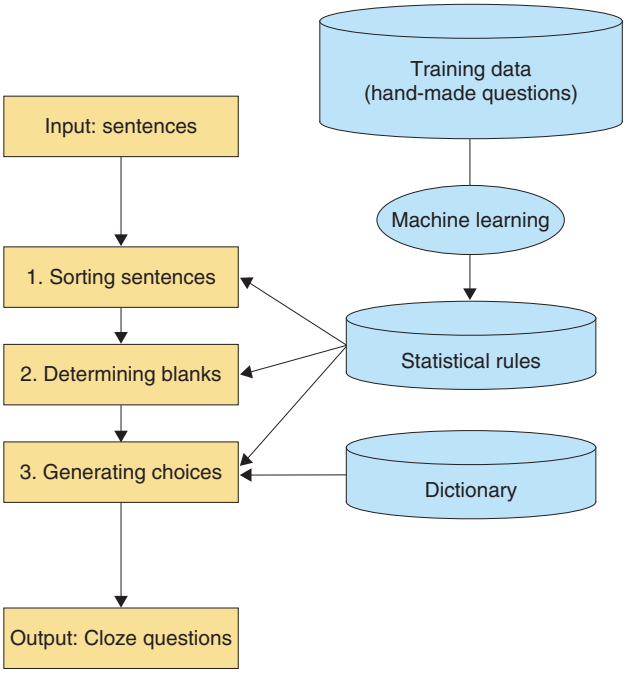


Fig. 2. Procedures of MAGIC.

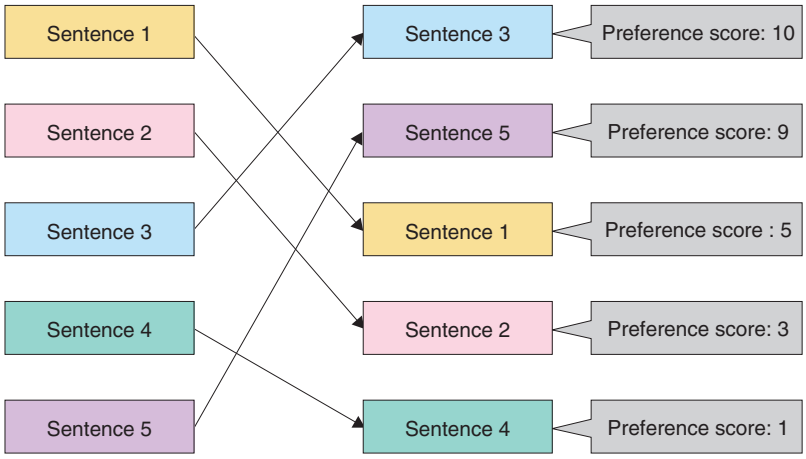


Fig. 3. Sorting sentences on the basis of preferences.

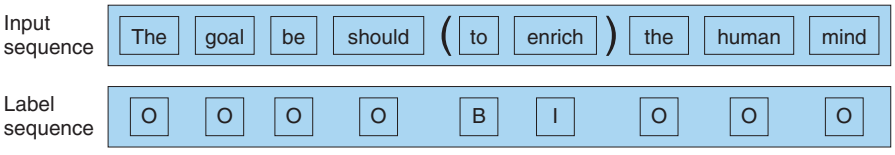


Fig. 4. Determining blanks by using sequence labeling.

3. Question generation procedures

In this section, we explain each component in detail.

3.1 Sorting sentences

Some of the input sentences might be appropriate and some of them might be inappropriate for English questions. For example, sentences that contain important idioms and have important grammar structures are more appropriate for questions. MAGIC sorts English sentences by comparing hand-made questions (training data) and standard English sentences by using preference learning [2]. An example is shown in **Fig. 3**. Preference learning enables us to assign high preference scores to sentences that are similar to the training data. The similarities can be calculated by using word appearance frequencies and parts of speech (POSs) in the sentences. By sorting input sentences in preference score order, users can learn English with appropriate sentences for questions.

3.2 Determining blanks

Determining words to be blanked can be regarded as a sequence labeling problem in machine learning. The sequence labeling problem is to estimate an optimal label sequence for a given input sequence. In our case, the input sequence is a sentence (word sequence), and the output label sequence is a sequence that represents the blank's position. The label sequence can be represented as shown in **Fig. 4**, where B, I, and O are standard IOB2 tags indicating the blank's beginning and words inside and outside the blank, respectively. With MAGIC, we use Conditional Random Field (CRF) [3], which has achieved high performance in sequence labeling problems. As features for CRF, we used words around the blank and their POSs. We confirmed that by using these features, we could select the words to be blanked more accurately than by using conventional methods.

3.3 Generating choices

MAGIC generates multiple choices on the basis of statistical information and patterns obtained from the training data. We classify a set of choices into two

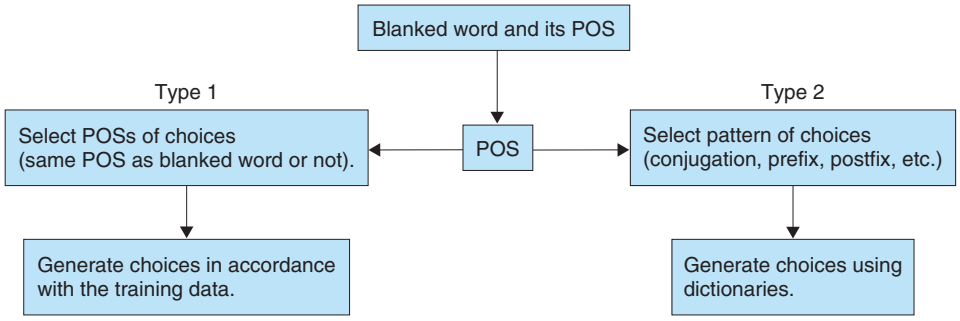


Fig. 5. Flow of choice generation.

types by the blank's POS. The first type is a POS that restricts words that can be chosen. For example, when the preposition *of* is blanked, the alternative choices are likely to be other prepositions such as *to*, *in*, and *at*. The words of an interrogative and auxiliary verb are also included in this type as well as prepositions. With this type, we can generate choices by considering the POS and word probabilities. The second type is a part of speech that has patterns in conjugation, orthography, or meaning. This type includes verbs, adjectives, and nouns. For example, patterns include various conjugations of the same base word (ask, asked, asking, asks), the same prefix or postfix (defective, elective, emotive, active), and similar meanings (told, said, spoke, talked). This type lets us select a pattern in accordance with the POS and generate choices using dictionaries. The flow of choice generation is shown in **Fig. 5**.

4. Conclusion

We are developing a system for automatically generating English cloze questions. We would like to sort sentences in order of user interests or difficulties [4], and this leads to personalized learning. We would also like to extend the system so that it can generate questions for learning other languages besides English and to generate learning materials in other subjects such as history and mathematics.

References

[1] T. Goto, T. Kojiri, T. Watanabe, T. Iwata, and T. Yamada, "Automatic Generation System of Multiple-choice Cloze Questions and Its Evaluation," Knowledge Management & E-Learning: An International Journal (KM&EL), Vol. 2, No. 3, pp. 210–224, 2010.

[2] M. Collins and N. Duffy, "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron," Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 263–270, 2001.

[3] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. of the 18th International Conf. on Machine Learning, pp. 282–289, 2001.

[4] T. Iwata, T. Kojiri, T. Yamada, and T. Watanabe, "Recommendation for English Multiple-choice Cloze Questions Based on Expected Test Scores," International Journal of Knowledge-based Intelligent Engineering Systems, Vol. 15, No. 1, pp. 15–24, 2011.

Tomoharu Iwata
Research Scientist, Learning and Intelligent Systems Research Group, NTT Communication Science Laboratories.

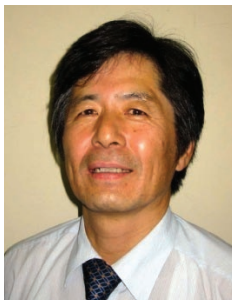
He received the B.S. degree in environmental information from Keio University, Tokyo, the M.S. degree in arts and sciences from the University of Tokyo, and the Ph.D. degree in informatics from Kyoto University in 2001, 2003, and 2008, respectively. His research interests include data mining, machine learning, information visualization, and recommender systems. He received the IPSJ (Information Processing Society of Japan) Best Paper Award, FIT (Forum on Information Technology) Young Researcher's Award, and Funai Best Paper Award. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and IPSJ.

Takuya Goto
NTT DOCOMO.

He received the B.E. and M.I. degrees from Nagoya University, Aichi, in 2007 and 2009, respectively. His research subject was an English learning support environment and automatic generation of English questions. He is currently interested in human-centered design.

Tomoko Kojiri
Associate Professor, Faculty of Engineering Science, Kansai University.

She received the B.E., M.E., and Ph.D. degrees from Nagoya University, Aichi, in 1998, 2000, and 2003, respectively. From 2003 to 2004, she was a research associate with the Graduate School of Information Science, Nagoya University. From 2004 to 2007, she was a research associate with the Information Technology Center, Nagoya University. From 2007 to 2011, she was an assistant professor with the Graduate School of Information Science, Nagoya University. Since 2001, she has been an associate professor with the Faculty of Engineering Science, Kansai University, Osaka. In 2011, she stayed at the Technical University in Graz as a guest researcher for two months. Her research interests include computer-supported collaborative learning, creative learning support, intelligent tutoring systems, and human-computer interfaces. She has received several academic awards including the Outstanding Paper Award of ICCE/ICCAI 2000 (International Conference on Computers in Education, International Conference on Computer-Assisted Instruction), the Best Paper Award of KES 2005 (Knowledge-Based & Intelligent Information & Engineering Systems), and the Outstanding Poster Presentation Award of ICCE2007. She is a member of IPSJ, the Japan Society for Artificial Intelligence (JSAI), IEICE, the Japanese Society for Educational Technology, the Japan Society for Information and Systems in Education (JSiSE), and the Asia-Pacific Society for Computers in Education.



Toyohide Watanabe
Professor, Graduate School of Information Science, Nagoya University.

He received the B.S., M.E., and Dr.Eng. degrees from Kyoto University in 1972, 1974, and 1985, respectively. Since 1975, he has worked in Kyoto University and Nagoya University: as a research associate in the Data Processing Center, Kyoto University, from 1975 to 1987; as an associate professor in the Faculty of Engineering, Nagoya University, from 1987 to 1994; as a full professor in the same faculty from 1994 to 1997; as a professor in the Graduate School of Engineering, Nagoya University from 1997 to 2003; and as a professor in the Department of Systems and Social Informatics, Graduate School of Information Science, Nagoya University, since 2003. In addition, from 2004 to 2008 he was concurrently the head director at the Information Technology Center, Nagoya University. His current research interests include intelligent tutoring systems, computer-supported collaborative learning, knowledge management, and intelligent activity-support. He is a member of the Association for Computing Machinery, IEEE-CS, the Association for the Advancement of Artificial Intelligence, Association for the Advancement of Computing in Education, KES International, IEICE, IPSJ, the Institute of Electrical Engineers of Japan, JSAI, the Japan Society for Software Science and Technology, JSiSE, etc. He is also currently Editor in Chief of the International Journal of Knowledge and Web Intelligence. He has been a Fellow of IEICE since 2004.

Takeshi Yamada
Senior Manager, Research Planning Department, NTT Science and Core Technology Laboratory Group.

He 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 joined NTT in 1988. His research interests are in machine learning, data mining, and combinatorial optimization. He is a member of IEEE, IEICE, IPSJ, the Association for Computing Machinery, and the Scheduling Society of Japan.