Automatic Scoring Method of Short-Answer Questions in the Context of Low-Resource Corpora

Tao-Hsing Chang
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan
changth@nkust.edu.tw

Ju-Ling Chen
Research Center for Indigenous Education
National Academy for Educational Research
Taipei, Taiwan
juling@mail.naer.edu.tw

Hui-Min Chou
Research Center for Indigenous Education
National Academy for Educational Research
Taipei, Taiwan
mayaw@mail.naer.edu.tw

Ming-Hong Bai
Research Center for Translation, Compilation and Language Education
National Academy for Educational Research
Taipei, Taiwan
mhbai@mail.naer.edu.tw

Fu-Yuan Hsu
Research Center for Psychological and Educational Testing
National Taiwan Normal University
Taipei, Taiwan
kevin@rcpet.ntnu.edu.tw

Yu-Chi Chen
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan
F110151101@nkust.edu.tw

2021 International Conference on Asian Language Processing (IALP) | 978-1-6654-8311-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/IALP54817.2021.9675160

***Abstract*—Short-answer questions offer a method to evaluate whether students have acquired a domain of knowledge. Because it is very time-consuming to evaluate the answer of the questions, many automatic scoring methods have been proposed. However, most of these models require large quantities of training data. This study proposes a method for automatic creation of a concept map based on small quantities of corpora. The concept map verified by experts is available and reliable. In addition, this study proposes a method to score the correctness of students' answers to short-answer questions using a machine-generated concept map. The preliminary experiment shows that the correctness of scoring results of the proposed method is close to that of the BERT-based model. Since the proposed method does not require large quantities of training data, it is suitable for the instruction system only using low-resource corpora.**

Keywords—short-answer question, concept map, automatic scoring, automatic construction

I. Introduction

The cultural inheritance of indigenous peoples is an important issue worldwide, and the teaching of knowledge about indigenous peoples offers an important means of cultural inheritance [1]. To facilitate the teaching of knowledge, some studies recommend that concept maps be used to evaluate whether students have acquired the framework for a particular domain of knowledge [2]. This evaluation method requires that students draw a concept map for a domain of knowledge and then compare the concept map with that drawn by experts. The larger the overlapping area between the two concept maps, the higher the score the students are given. Many studies have shown that the evaluation method is very effective [3]-[6]. However, this method is scarcely used in teaching. This is mainly for two reasons: 1) this method is based on the premise that the students are able to draw a concept map proficiently, but they may not be able to do so – even if they are able to draw a concept map, the process is very time-consuming; 2) it is very time-consuming for experts to draw a concept map and check and correct the students' concept map.

Short-answer questions offer another method to evaluate whether students have acquired a domain of knowledge. Teachers can ask students a question about certain knowledge, and students can answer the question using several sentences or one or two paragraphs. It is very time-consuming to check and correct short-answer questions, so many scoring methods for short-answer questions have been proposed. Because of rapid progress in natural language processing (NLP) and machine learning in recent years, many studies [7]-[10] have been conducted on automatic check and correction of short-answer questions with the use of deep-learning models. However, these supervised models require large quantities of training data, so they cannot be used for the teaching of knowledge on indigenous people, which lacks large quantities of students' answer data for short-answer questions.

This study aims to develop an automatic scoring method for short-answer questions in the context of low-resource corpora. First, an expert concept map is created based on small quantities of corpora. Then, a projected concept map is generated based on a combination of students' answers and the expert concept map. Last, several indices of node importance are calculated, and they are used as the basis for calculating the score of the projected concept map. The score can be used to evaluate whether students have acquired the knowledge within the specific domain.

The rest of this paper is organized as follows. Section II conducts a review of studies about the evaluation method proposed in this study. Section III describes the technical details of the evaluation method. Section IV analyzes the effectiveness of the method in scoring the short-answer questions. Section V draws conclusions and points out the orientations of subsequent studies.

II. Related Works

So far, many studies have discussed how to automatically create concept maps. Early studies [11]-[15] mainly focused on the automatic creation of knowledge ontology. For example, Lee et al. [13] used the TF-IDF method to identify important nouns, conducted a term analysis on these nouns to determine their semantic parameters, and input these parameters into a self-organizing map model to generate a concept map. In recent years, related studies have concentrated on the automatic creation of concept maps for specific domains of knowledge. Atapattu et al. [16] extracted the concept relation concept patterns in the projected maps for

speeches, and organized these patterns into a hierarchical concept map. Shao et al. [17] proposed an automatic creation method known as "TA-ARM." In addition, researchers have proposed many automatic creation methods for knowledge graphs, which are quite similar to concept maps and are used for semantic searches on the Internet. Most of these methods analyze and calculate the relationship between concepts using many different deep-learning models [18][19]. As far as we know, these methods require large quantities of training corpora.

C-rater [20] is the first technology that makes great progress in the automatic scoring of short-answer questions. The subsequent methods are mainly classified into two types. The first type of method comprises the following steps: 1) the training data of each score grade are converted into semantic vectors; 2) students' answer content is compared with these vectors; and 3) the score grade of the vector with the highest similarity is the score grade of the students' answer content. For a short-answer question, Heilman and Madnani [21] first analyzed the correct answer content, identified the influence of each word on the correctness of answers, and used the scikit-learn method to create an SVM prediction model for the question. In recent studies, deep-learning models (e.g., LSTM, CNN, and BERT) [7]-[10] were used as automatic scoring models for short-answer questions. Likewise, these methods require large quantities of training data.

III. Methodology

The automatic scoring method for short-answer questions herein includes two steps: 1) automatically create an expert concept map using small quantities of knowledge corpora; 2) based on the students' answer content combined with the expert concept map, create a projected concept map for scoring and calculate the score of the projected concept map for use as the score of the answer content. The following sections describe the two steps in detail.

A. Generating a concept map automatically by using low-resource corpora

In this study, the automatic construction for concept maps is derived from the method specified by Chang et al. [22]. The method comprises three steps: 1) choose common nouns and proper nouns from corpora, use such nouns as concepts, and calculate the occurrence frequency of each concept and co-occurrence frequency of paired concepts; 2) identify the dependency relationship between paired concepts according to the two frequencies above; 3) calculate the importance degree of each concept according to the dependency relationship between it and other concepts, identify the most important concepts according to the importance degree of each concept, and create a concept map accordingly; and 4) use the grammar parser to identify main verbs that are capable of connecting two concepts, and use such verbs as conjunctions between paired concepts.

In a corpus with numerous words, important concepts have a certain occurrence frequency, so the method proposed by [22] can perform well. In a low-resource corpus, however, the method is confronted with two problems. First, the occurrence frequency of important concepts is not necessarily higher than that of common words, so important concepts are prone to be regarded as unimportant words in the original method. The traditional TD-IDF method for identifying important words is not suitable for low-resource knowledge corpora, because important concepts appear in each document. Second, Chinese documents must be subjected to word segmentation before analysis. A Chinese sentence is composed of Chinese characters, which are not separated by blank spaces, so the Chinese sentence must be decomposed into Chinese words by a word segmentation tool. However, there are a few errors in the results of word segmentation, and some incorrect Chinese words may appear in the analysis results of the Chinese document because of word segmentation.

For the two questions, we used the following equation to calculate the importance of concept *c*.

$$IV(c) = {fq_K(c)}\Big/{\log(fq_B(c) + \max(fq_B(l), fq_B(r)))} \quad (1)$$

where *B* denotes a large balance corpus; *K* denotes the corpus about specific knowledge; *l* is the first character of concept c; *r* is the last character of concept *c,* and $fq_D(x)$ denotes the frequency of the concept *x* in corpus *D*.

The concepts with low IV values are not shown in concept maps. The equation (1) mainly is derived from the following idea. If a concept seldom appears in a general language environment but often appears in a knowledge corpus, and its combining characters are not common single words, the concept may be an important concept. In addition, a traditional concept map is a tree-shaped hierarchical concept map. However, the concept map generated by the method herein is a directed graph. In other words, the concept map comprises nodes and edges; specifically, a node stands for a concept, and an edge stands for a connection between two concepts.

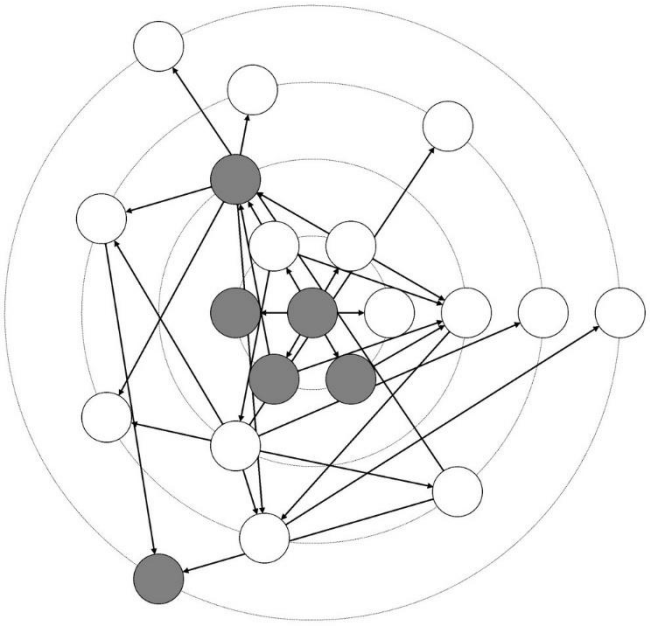

Fig. 1 An Example of Projected Concept Map

B. Scoring of short-answer questions

As mentioned in subsection III-A, a concept map for specific knowledge can be created based on a low-resource corpus. This concept map can play the role of the expert concept map in the traditional concept map–scoring method. For a student's answer, we extract common nouns and proper nouns as identified concepts and use the grammar parser to

identify the main verbs between paired concepts in each sentence.

Fig. 1 is an example of projected concept map. The map is the one that is automatically created in subsection III-A, and the grey part pertains to the concepts mentioned in the students' answer content and connections between different concepts. Evidently, the larger the grey part, the higher the degree of overlap between a student's answer content and the expert concept map, and the higher the score. In Fig. 1, some concepts are obviously more important, and a few concepts located in the outer layer are not incorporated without affecting the correctness of the student's answer. Therefore, the correctness of a student's answer *a* can be calculated using equation (2).

$$Sco(\mathrm{a}) = \frac{1}{2p}\sum_{g\in G}\frac{Con(g)}{Lev(g)} \qquad (2)$$

where G is the set of identified concepts; function $Con(g)$ represents the number of connections of concept g; function $Lev(g)$ is the level of concept g, which can be calculated by the algorithm in [22]; p is the number of connections in the map. The equation reveals that both a higher importance grade and a larger number of connections with other concepts lead to a higher score.

The score calculated by the equation is in the value range of 0 to 1. Sometimes, the scoring results are classified into several categories. For example, the scoring results of short-answer questions include such categories as "completely correct," "partially correct," and "incorrect." Therefore, the score of an answer should be converted to the category the answer belongs to.

Assume a set A is the collection of the answers for an question, and answer *a* is an element in A. Moreover, a set M, the collection of answers for a previous question, consists of several subsets M_1, M_2,…, M_n. M_i represents the collection of the answers which were identified category i by human experts, and the correctness of answers in M_i is higher than that in M_{i+1}. The category of answer *a* can be identified by the following equation.

$$Cat(a) = \underset{1\le p\le n}{\arg}\left(\frac{S(M_{p-1})}{R(M)} \le \frac{Rank(a)}{R(A)} \le \frac{S(M_p)}{R(M)}\right) \qquad (3)$$

where $Rank(a)$ is the number of answers in the set A whose scores are greater than the score of the answer a; $R(U)$ represents the number of the answers in set U; M_0 is an empty set; $S(M_i)$ can be calculated by the following equation:

$$S(M_i) = \sum_{k=0}^{i} R(M_k) \qquad (4)$$

IV. Experiments

This section describes the results of two experiments. As exemplified by the knowledge about traditional slate houses of Taiwan's Paiwan and Rukai Tribes, subsection IV-A describes the automatic construction of a concept map based on small quantities of corpora. Because of a lack of data sets about answers to questions on indigenous knowledge, we apply the automatic scoring method with concept map to a well-known data set of short-answer questions about scientific knowledge, to verify its effectiveness in subsection IV-B.

A. Example of machine-generated concept map

Academic dissertations offer the most accessible source of corpora in the endangered culture and are feasible in practice. Hence, we collected 11 academic dissertations on slate houses of Taiwan's Paiwan and Rukai tribes as a source of corpora. The 11 academic dissertations were published from 2003 to 2016 and were all written in Chinese, containing a total of 950 thousands Chinese characters. Using the method specified in subsection III-A, we created a concept map for the knowledge of slate houses. This concept map was very large, so we first drew a concept map for the 20 most important concepts, as shown in Fig. 2.

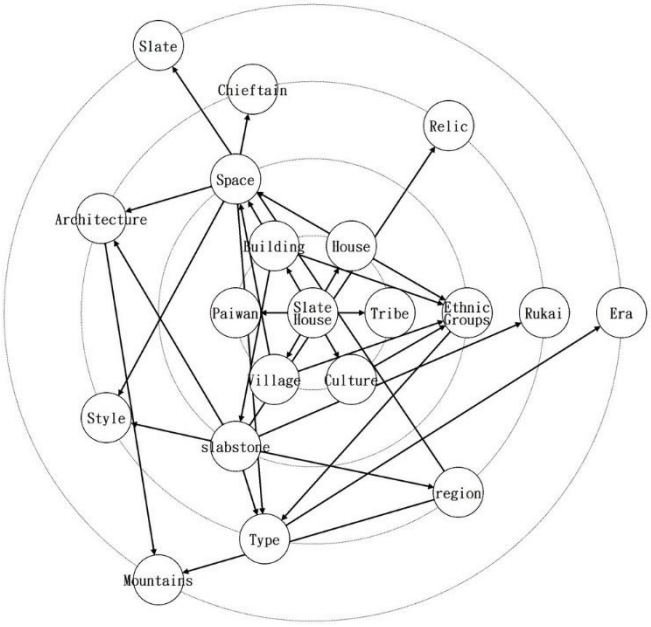

Fig. 2 An Example of Concept Map Generated by the Proposed Method

Experts with aboriginal knowledge have interpreted the concept map, considering that 19 concepts and 29 connections are correct. However, experts consider that concept "buildings" is not among the 20 most important concepts. We think that concept "buildings" is a method commonly used to check buildings, so it is considered by machines as an important concept. In addition, experts consider that there is no connection between concepts "tribe" and "Paiwan". Moreover, there is also no connections among concepts "slabstone" and the set of concepts "slate house", "architecture" and "type". Overall, experts consider that the concept map is still highly correct.

B. Effectiveness of the automatic scoring method

In this study, the answers to three short-answer questions and their scoring results in a study published in the Trends in International Mathematics and Science Study 2003 [23] are selected as a data set. The three short-answer questions are as follows:

Q1. Advantage of having two ears: what are the benefits to our auditory sense if we listen with two ears rather than one ear?

Q2. How glasses/contact lenses work: please briefly describe how glasses (including contact lenses) help people see more clearly.

Q3. Nail pulled out of a wooden board: when we pull a nail out of a plank, the nail will get hot. Please explain the reason.

Table I describes the quantity of answers with different scoring results (i.e., completely correct (CC), partially correct (PC), and incorrect (IC)) in the data set of each question. To compare with existing common depth model methods, we used the BERT model employed by Lun et al. [8] as the standard of comparison. For the BERT-based model, each answer is converted into a semantic vector through BERT [24], and then the correctness of each answer is judged by a full-connection classifier.

TABLE I. NUMBER OF ANSWERS IN DATASETS

Questions	*# of CC*	*# of PC*	*# of IC*	*Total*
Q1	80	88	457	625
Q2	116	81	496	693
Q3	597	0	77	674

Each corpus is divided into a training set and a test set, because the BERT-based model requires training data and the proposed method herein requires no training data. In this study, the effectiveness of the proposed method and the BERT-based model is measured in terms of two indices, exact agreement rate (EAR) and serious error rate (SER). The equations for calculating the two indices are as follows:

$$EAR = \sum_{i=1}^{n}\sum_{j=1}^{n} e_{i,j} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} \tag{5}$$

$$SER = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{i,j} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} \tag{6}$$

where n is the number of categories; $c_{i,j}$ represents the number of students' answers which was identified category i by experts and category j by machines; $e_{i,j}$ and $f_{i,j}$ are as follows:

$$e_{i,j} = \begin{cases} c_{i,j}, & if\ i = j \\ 0, & otherwise \end{cases} \tag{7}$$

$$f_{i,j} = \begin{cases} c_{i,j}, & if\ |i-j| > 1 \\ 0, & otherwise \end{cases} \tag{8}$$

Table II describes the performance of the BERT-based model and the proposed method in terms of various indices. For all questions, the BERT-based model performs better in the exact agreement, whereas the proposed method performs better in the severe error rate. For questions Q2 and Q3, the BERT-based model outperforms (but slightly) the proposed method in the exact agreement. In addition, the exact agreement provided by the proposed method is also as high as 0.68-0.90. This is quite close to the exact agreement between two human raters in the condition of manual scoring.

TABLE II. PERFORMANCE OF TWO METHODS

Questions	BERT-based method		The proposed method	
	EAR.	*SER*	*EAR*	*SER*
Q1	0.77	0.05	0.59	0.00
Q2	0.69	0.12	0.68	0.03
Q3	0.95	0.05	0.90	0.10

V. CONCLUSIONS

This study mainly makes two contributions. First, this study proposes a method for automatic creation of a concept map based on small quantities of corpora. The concept map verified by experts is available and reliable. Second, this study proposes a method to score the correctness of students' answers to short-answer questions using a machine-generated concept map. The preliminary experiment shows that the correctness of scoring results of the proposed method is very close to that of the BERT-based model. The method proposed in this study does not require large quantities of training data, so it is suitable for the teaching of low-resource corpora.

Because of a lack of data sets of answers to low-resource short-answer questions, the proposed method herein, or, specifically, whether the method can be applied to the answers to low-resource short-answer questions, is yet to be further verified. Moreover, the concept map generated by the proposed method contains a few minor errors, which are yet to be further corrected. Overall, the proposed method offers a feasible solution to the automatic scoring of short-answer questions in the teaching of low-resource corpus knowledge.

ACKNOWLEDGMENTS

This study was partially supported by the Ministry of Science and Technology, Taiwan, under the Grant MOST 110-2420-H-992-001 and 107-2511-H-992-001-MY3.

REFERENCES

[1] M. G. Stevenson, "Indigenous knowledge in environmental assessment," Artic, vol. 49, no. 3, pp. 278-291, 1996.

[2] J. D. Novak, and A. J. Canas, "Theoretical origins of concept maps, how to construct them, and uses in education," Reflecting Education, vol. 3, no. 1, pp. 29-42, 2007.

[3] K. E. Chang, Y. T. Sung, and S. F. Chen, "Learning through computer-based concept mapping with scaffolding aid. Journal of Computer Assisted Learning," vol. 17, no. 1, pp. 21-33, 2001.

[4] K. E. Chang, Y. T. Sung, and I. D. Chen, "The effect of concept mapping to enhance text comprehension and summarization," The Journal of Experimental Education, vol. 71, no. 1, pp. 5-23, 2002.

[5] P. Chularut, and T. K. DeBacker, "The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language," Contemporary Educational Psychology, vol. 29, no. 3, pp. 248-263, 2004.

[6] M. Elhelou, "The use of concept mapping in learning science subjects by Arab students," Educational Research, vol. 39, no. 3, pp. 311-317, 1997.

[7] S. Kumar, S. Chakrabarti, and S. Roy, "Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading. IJCAI, pp. 2046-2052, 2017.

[8] J. Lun, J. Zhu, Y. Tang, and M. Yang, "Multiple data augmentation strategies for improving performance on automatic short answer scoring," Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 09, pp. 13389-13396, 2020.

[9] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. Lee, "Investigating neural architectures for short answer scoring," Proceedings of the 12th

Workshop on Innovative Use of NLP for Building Educational Applications, pp. 159-168, 2017.

[10] C. Sung, T. Dhamecha, S. Saha, T. Ma, T. Reddy, V., and R. Arora, "Pre-training BERT on domain resources for short answer grading," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6071-6075, 2019.

[11] S. M. Bai, and S. M. Chen, "Automatically constructing concept maps based on fuzzy rules for adapting learning systems," Expert Systems with Applications, vol. 35, no. 1, pp. 41-49, 2008.

[12] N. S. Chen, P. Kinshuk, C. W. Wei, and H. J. Chen, "Mining e-learning domain concept map from academic articles," Computers & Education, vol. 50, pp. 1009-1021, 2008.

[13] C. S. Lee, Y. F. Kao, Y. H. Kuo, and M. H. Wang, "Automated ontology construction for unstructured text documents," Data and Knowledge Engineering, vol. 60, pp. 547-566, 2007.

[14] C. H. Lee, G. G. Lee, and Y. Leu, "Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning," Expert Systems with Applications, vol. 36, no. 2, pp. 1675-1684, 2009.

[15] S. S. Tseng, P. C. Sue, J. M. Su, J. F. Weng, and W. N. Tsai, "A new approach for constructing the concept map," Computers & Education, vol. 49, no. 3, pp. 691-707, 2007.

[16] T. Atapattu, K. Falkner, and N. Falkner, "A comprehensive text analysis of lecture slides to generate concept maps," Computers & Education, vol. 115, pp. 96-113, 2017.

[17] Z. Shao, Y. Li, X. Wang, X. Zhao, and Y. Guo, "Research on a new automatic generation algorithm of concept map based on text analysis and association rules mining," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 2, pp. 539-551, 2020.

[18] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4762-4779, 2019.

[19] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records. Scientific reports, vol. 7, no. 1, 1-11, 2017,

[20] C. Leacock, and M. Chodorow, "C-rater: Automated scoring of short-answer questions. Computers and the Humanities," vol. 37, no. 4, pp. 389-405, 2003.

[21] M. Heilman, and N. Madnani, "The impact of training data on automated short answer scoring performance," Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 81-85, 2015.

[22] T. H. Chang, H. P. Tam, C. H. Lee, Y. T. Sung, "Automatic Concept Map Constructing Using Topic-specific Training Corpus," Proceedings of APERA 2008, Singapore, 2008.

[23] I. V. Mullis, M. O. Martin, E. J. Gonzalez, and S. J. Chrostowski, "TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades," International Association for the Evaluation of Educational Achievement. Amsterdam, Netherlands,2004.

[24] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, vol. 1, 2019.