Automated Prediction of Item Difficulty in Reading Comprehension Using Long Short-Term Memory

Li-Huai Lin, Tao-Hsing Chang
Dept. of Computer Science and Information Engineering
National Kaohsiung Univ. of Science and Technology
Kaohsiung, Taiwan.
{1105108143, changth}@nkust.edu.tw

Fu-Yuan Hsu
Institute for Research Excellence in Learning Sciences
National Taiwan Normal University
Taipei, Taiwan.
kevinhsu@ntnu.edu.tw

***Abstract*—Standardized tests are an important tool in education. During the test preparation process, the difficulty of each test item needs to be defined, which previously relied on expert validation or pretest for the most part, requiring a considerable amount of labor and cost. These problems can be overcome by using machines to predict the difficulty of the test items. In this study, long short-term memory (LSTM) will be used to predict the test item difficulty in reading comprehension. Experimental results show that the proposed method has a good prediction for agreement rate.**

Keywords-sentence Reading comprehension; Item difficulty estimation; Long short-term memory

I. Introduction

The item response theory (IRT) is one of the main theories of modern educational assessment. Many tests such as Test of English as a Foreign Language (TOEFL) and Scholastic Assessment Test (SAT) are developed based on this method. The utilization of IRT involves an item bank, which must include the difficulty of each item. In the past, test item difficulty was estimated by one of two methods. In the first one, the difficulty of each item is collected in small-scale experiments, these experiments are also called pretest. However, the pretest method is very costly. The second method adopts expert evaluation, which costs less but is not as accurate as the pretest.

Using a computer program for automated item difficulty estimation solves the two problems above. There has been extensive research on automated machine evaluation for English content, but most of them classify the difficulty of test items based on manually defined features. By virtue of the development of deep learning, many classification models have had manually defined features replaced with those learned by neural networks. Long short-term memory (LSTM) [1] is a deep learning model commonly used in natural language processing (NLP). Compared with other deep learning models, LSTM features the utilization of context to improve the accuracy of prediction. Therefore, the purpose of this paper is to propose an automated item difficulty estimation method based on LSTM. However, because the test items may belong to various categories of subjects, and some contain pictures and tables, etc., that cannot be easily converted into a single data format, this study focuses only on Chinese reading comprehension items consisting only of words.

II. Related Works

Freedle and Kostin [2]-[3] pioneered the studies on difficulty evaluation for reading comprehension. The study assumed many variables of the test items, such as those defining whether an item tests the examinees' understanding of the main idea or expecting the examinee to deduce a result from the item statement, and so on. The study used the calculation results based on these variables to evaluate the item difficulty. Using the method above, Chon and Shin [4] first selected a series of predictors, such as response time and paragraph length, that may affect the difficulty of test items based on relevant studies and statistical data obtained from the College Scholastic Ability Test (CSAT). After classification, the predictors were then compiled into a scale. As validated by the actual application in the preliminary CSAT in September 2009, the methods of the study can be used to effectively estimate the overall average score of the English section of the CSAT. The methods are also readily applicable to similar tests.

Boldt and Freedle [5] proposed a model based on a neural network to predict the difficulty of test items and applied a genetic algorithm (GA) to obtain improved prediction results. Based on the above research, Loukina et al. [6] proposed to predict the difficulty of listening tests according to the text complexity in language proficiency tests. The study used the TextEvaluator, a system for predicting the complexity [7]-[9], to extract the text complexity features of the items. Also, the linear regression classifier in SciKit-Learn Laboratory (SKLL) was used to evaluate the item difficulty.

Hsu et al. [10] utilized the word embedding technique proposed by [11]. First, the semantic space was constructed using the learning corpus. Then, the items and options were projected onto the semantic space to obtain the corresponding semantic vectors. Next, they computed the semantic similarities among the stem, answer, and distractors, which were then input to a support vector machine (SVM) for training to predict the item difficulty. Among the methods mentioned above, only the estimation model achieved automated calculation, and the predictors still required manual definition or evaluation.

LSTM can achieve automated extraction of the features in a text that can be used as the bases for prediction. Wu et al. [12] used LSTM as the computation unit and constructed an eight-layer LSTM encoder-decoder model for machine translation. Adopting a gated recurrent unit (GRU) [13], Song et al. [14] proposed a text-emotion detection system integrating a four-layer model. This system starts from the bottom layer. First, it converts the words into word vectors. Then it inputs each word vector of a sentence into a sentence level GRU layer, whose output is equivalent to the sentence vector of the corresponding

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

sentence. Next, it inputs the sentence vectors into the discourse level BiGRU layer. Finally, a multi-label layer outputs the results of emotional labeling. This method offers considerable assistance in essay analysis and scoring of compositions, etc. This study also adopts LSTM in designing the item difficulty prediction model.

III. Method

This paper focuses solely on the difficulty evaluation of Chinese reading comprehension items. Fig. 1 shows an example of the test items involved in this study. Unlike English, Chinese words and phrases are not separated from each other by spaces. Therefore, Chinese sentences must first undergo the word segmentation procedure so that the word boundaries within the sentences can be identified. This paper adopts WECAn [15] for word segmentation, which utilizes both the lexicography-based bidirectional maximum matching and the conditional random field (CRF) as a second word segmentation method.

「藏書室是一處有著許多迷睡靈魂的神奇陳列室，
當我們呼叫他們，這些靈魂就甦醒過來。」
這句話的意旨，與下列何者最接近？
(“The library is a magical gallery of many sleepy souls.
When we call to them, these souls awaken.”
Which of the following is closest to the meaning of this sentence?)

(A) 許多作家在藏書室找到心靈的歸屬
(Many writers find their spiritual home in the library.)

(B) 好的作品可以喚醒讀者迷睡的靈魂
(Good writing can awaken the sleepy souls in readers.)

(C) 迷失在書海的讀者需要適當的指引
(Readers who are lost in a sea of books need proper guidance.)

(D) 讀者可透過閱讀與作者的心靈相通
(Readers can connect mentally with writers through reading.)

Fig. 1. Example of an item from the reading comprehension test, retrieved from the Chinese section of comprehensive assessment program (CAP) test in 2018.

This study adopts word embedding to transform words into semantic vectors. This technique is able to identify the coordinates of words in the semantic space, namely, the word vector. In other words, by examining the distance between two-word vectors, the degree of semantic similarity between two words can be determined. This study utilizes word embedding proposed by Mikolov et al. to train the transformation model to perform word-semantic vector transformation. When the training is completed, a word can be transformed into a word vector through this model. This study uses the Sinica Balance Corpus as the data source for model training. When the training is completed, each word is converted into a vector of 300 dimensions.

In this study, the LSTM proposed by Hochreiter and Schmidhuber is adopted as the prediction model, as shown in Fig. 2. Using this as an example, C_{t-1} represents the current memory, h_{t-1} represents the previous LSTM output, and x_t represents the current input value. Through the forget gate and input gate, the decision can be made on whether to update the memory C_t thereby affecting the output value h_t.

If the current word or phrase is related to a new topic, the forget gate will filter out the previous memory. The input gate determines whether the current input and the newly-generated memory cell candidate should be added to the long-term memory. The two gates are usually sigmoid functions that indicate whether the previous memories have been forgotten or retained. The output gate is also a sigmoid function, determining whether to add the current word or phrase to the output.

This method is one of the improved models of RNN. Through appropriate gate design, the problems of vanishing and exploding gradients generated by RNN are avoided while retaining the same input and output methods as those for RNN.

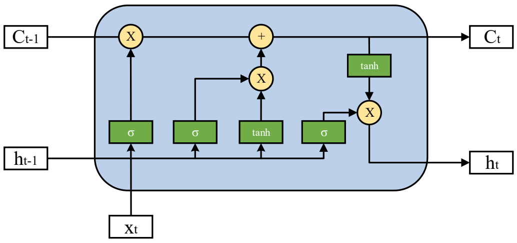


Fig. 2. LSTM

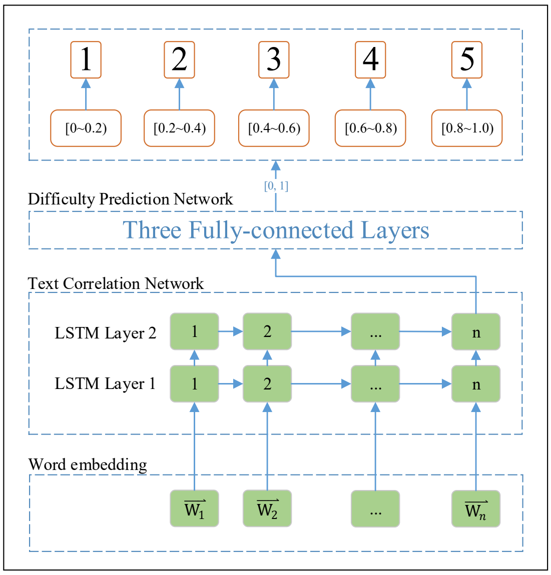


Fig. 3. The architecture of the proposed model.

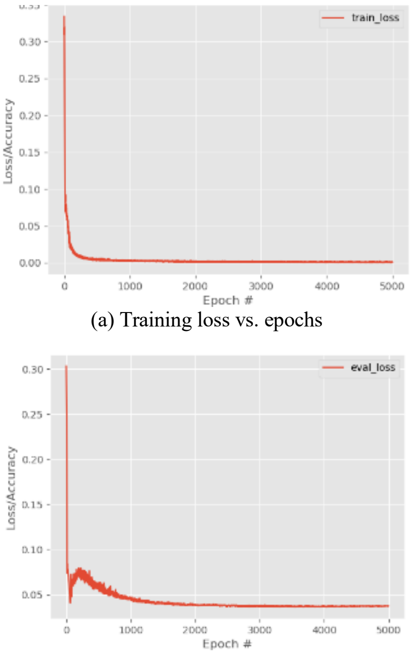
Fig. 3 shows a multi-layer LSTM prediction model used in this study. This model is based on the architecture of model proposed by [14]. Our model consists of two networks: text correlation network (TCN) and difficulty prediction network (DPN). First, the proposed model converts the words of an item into semantic vectors using word embedding technique. For instance, the word W_1 in Fig. 3 is converted into the semantic vector $\overrightarrow{W_1}$ which consists 300 dimensions. These vectors are input into the TCN.

The TCN contains two uni-directional LSTM layers to establish the correlation between the words of an item. Then, the output derived from the TCN are input into the DPN. DPN is composed of three fully-connected (FC) layers. It employs the features of inputs to evaluate the difficulty of an item. The DPN output a value which is between 0 and 1 finally. Since the difficulty of an item is defined from level 1 to level 5, the value is converted into a level by a conversion rule.

IV. Experiment

The training and test data collected for this experiment included 80 items from the basic competence test (BCTEST) and CAP TEST during 2013-2017. Due to the shortage of items, another 254 items designed by experts were added, resulting in a total of 334 items. The experts were professionals with over ten years' experience for analyzing items. The item difficulty was divided into five levels, with "difficulty 1" being the easiest item and "difficulty 5" the most difficult. The difficulty of the BCTEST and CAP TEST items were determined by the actual test results, whereas the items designed by experts were determined by the experts.

In the experiment, Adam was adopted as the optimizer of the proposed neural networks. The loss function was mean squared error (MSE). The number of repetitions of training was 5000 epochs. The initial learning rate was 0.0001, which decreased as the number of repetitions of training increased. Due to the small amount of experimental data, the 10-fold cross validation method was adopted for the experiment. Fig. 4 shows the learning curves of the proposed model. It suggests that the model have no overfitting.



(a) Training loss vs. epochs

(b) Tuning test loss vs. epochs.

Fig. 4. Learning Curves

Because the automated difficulty evaluation methods in the previous studies were not applicable to Chinese reading comprehension items, the experiment only presents the results obtained via the method proposed in this study. Table 1 shows the confusion matrix produced by the experiment. From Table 1, three indices representing predictive effectiveness can be further calculated. The first is the exact agreement rate (EAR), which indicates the proportion of items whose predicted item difficulty exactly equals the actual difficulty. The second is adjacent agreement rate (AAR), which indicates the proportion of the items whose predicted item difficulty has a difference of one level from the actual difficulty. The serious error rate (SER) indicates the proportion of items whose predicted difficulty has a difference of over three levels from the actual difficulty.

As shown in Table 1, the EAR, AAR, and SER were 0.37, 0.84, and 0.01, respectively. According to the analysis of the difference between the predicted difficulty and the actual difficulty proposed by [10], the validity of the current results is very close to that given by pretests. This suggests that the model is indeed effective in assessing item difficulty.

TABLE I. Confusion Matrix

		Actual difficulty				
		1	***2***	***3***	***4***	***5***
Predicted difficulty	1	19	5	2	0	0
	2	34	27	13	12	0
	3	14	24	30	16	9
	4	3	9	19	21	33
	5	0	1	2	13	28

V. Conclusions and Future Work

This study presents a difficulty prediction model for test items based on LSTM, which can effectively predict the item difficulty in Chinese reading comprehension as demonstrated by preliminary experiments. There are four potential topics for future work.

Firstly, the test items used in this experiment were in Chinese, thus the validity of the method in an English context cannot be confirmed. However, many studies have shown that word embedding and LSTM, which constitute the proposed method in this paper, perform identically in different languages. Therefore, it is appropriate to apply them to English items. Secondly, the method proposed in this paper was tested using the items in reading comprehension only. Its validity for other types of test items needs to be further tested and evaluated. Thirdly, the reasons for the difficulty in confirming the item features captured by LSTM still require investigation.

Therefore, it may be possible to further improve the accuracy of prediction by combining the manually defined characteristics that are known to be valid and whose results can be obtained by automated calculations. Finally, other deep learning models can be used as prediction models to explore the most suitable one for test item difficulty prediction.

Acknowledgment

This study was partially supported by the Ministry of Science and Technology (MOST), under the grant 107-2511-H-992-001-MY3 and 108-2813-C-992-017-H, and also financially supported by the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

References

[1] S. Hochreiter and J. Schmidhuber. "Long Short-term Memory." Neural computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.

[2] R. Freedle and I. Kostin. "The prediction of SAT reading comprehension item difficulty for expository prose passages." ETS Research Report, vol. 1991, no. 1, pp. i-52, Jun. 1991.

[3] R. Freedle and I. Kostin. "The prediction of TOEFL reading item difficulty: implications for construct validity." Language Testing, vol. 10, no. 2, pp. 133-167, Jul. 1993.

[4] Y. V. Chon and T. Shin. "Item Difficulty Predictors of a Multiple-choice Reading Test," ENGLISH TEACHING (영어교육), vol. 65, no. 4, pp. 257-282, 2010.

[5] R. F. Boldt, and R. Freedle. "Using a neural net to predict item difficulty." ETS Research Report, vol. 1996, no. 2, pp. i-19, Dec. 1996.

[6] A. Loukina, S. Yoon, J. Sakano, Y. Wei and K. Sheehan, "Textual complexity as a predictor of difficulty of listening items in language proficiency tests," in Proc. COLING, 2016, pp. 3245–3253.

[7] K. M. Sheehan, M. Flor, and D. Napolitano. "A two-stage approach for generating unbiased estimates of text complexity," in Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility, 2013, pp. 49–58.

[8] K. M. Sheehan, I. Kostin, D. Napolitano, and M. Flor. "The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment." The Elementary School Journal, vol. 115, no. 2, pp. 184–209, Dec. 2014.

[9] D. Napolitano, K. M. Sheehan, and R. Mundkowsky. "Online Readability and Text Complexity Analysis with TextEvaluator," in Proc. NAACL-HLT, 2015, pp. 96–100.

[10] F. Y. Hsu, H. M. Lee, T. H. Chang, and Y. T. Sung, "Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques." Information Processing and Management, vol. 54, no. 6, pp. 969-984, Nov. 2018.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems 26, 2013, pp. 3111-3119.

[12] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, ... and J. Klingner. "Google's neural machine translation system: Bridging the gap between human and machine translation." Oct. 2016. Internet: https://arxiv.org/abs/1609.08144.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling," in NIPS 2014 Deep Learning and Representation Learning Workshop, 2014. Internet: https://arxiv.org/abs/1412.3555.

[14] W. Song, D. Wang, R. Fu, L. Liu, T. Liu, and G. Hu. "Discourse Mode Identification in Essays," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 112-122.

[15] T. H. Chang. "The Development of Chinese Word Segmentation Tool for Educational Text," in Proceedings of the 7th International Conference on Information, 2017, pp. 179-182.