Automatic Chinese Short Answer Grading with Deep Autoencoder

Xi Yang[1], Yuwei Huang[2,3], Fuzhen Zhuang[4,5(✉)], Lishan Zhang[1], and Shengquan Yu[1]

[1] Beijing Advanced Innovation Center for Future Education, Beijing Normal University, Beijing 100875, China
{xiyang85,lishan,yusq}@bnu.edu.cn
[2] Sunny Education Inc., Beijing 100102, China
huangyw95@foxmail.com
[3] Beijing University of Chemical Technology, Beijing 100029, China
[4] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
zhuangfuzhen@ict.ac.cn
[5] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Short answer question is a common assessment type of teaching and learning. Automatic short answer grading is the task of automatically scoring short natural language responses. Most previous autograders mainly rely on target answers given by teachers. However, target answers are not always available. In this paper, a deep autoencoder based algorithm for automatic short answer grading is presented. The proposed algorithm can be built without expressly defining target answers, and learn the lower-dimensional representation of student responses. For the sake of reducing the influence of data imbalance, we introduce the expectation regularization term of label ratio into the model. The experimental results demonstrate the effectiveness of our proposed method.

Keywords: Automatic grading · Short Answer · Deep autoencoder
Text classification

1 Introduction

Grading short-answer questions with a natural language response automatically has been extensively studied for a long time, due to the fact that Automatic Short Answer Grading (ASAG) systems can overcome some limitations of human scoring [1,2]. C-rater [3] is probably the most well-known system. With the development of machine learning techniques, various machine learning algorithms have been applied to ASAG task, such as Logistic Regression (LR) [4], Decision Tree [5,6], k-Nearest Neighbor [7], Naive Bayes [8], Support Vector Machine (SVM) [3,9], Deep Belief Network [10] and so on.

The traditional ASAG methods based on machine learning have the following limitations. Firstly, much of the prior researches grade the student responses

C. Penstein Rosé et al. (Eds.): AIED 2018, LNAI 10948, pp. 399–404, 2018.
https://doi.org/10.1007/978-3-319-93846-2_75

based on target answers provided by teachers. However, the target answers are not always available. Secondly, the representations of student responses extracted from natural language processing techniques are always high-dimensional and high-sparse. Finally, for most traditional machine learning models, one of the basic assumptions is that the distribution of class ratio on data should be balanced. But this assumption is not satisfied in most cases.

Based on the analysis above, this paper is aiming at presenting an algorithm for Chinese ASAG by only using graded student responses and without any target answers. The algorithm needs to be able to get the lower-dimensional representation of student responses, and overcome the imbalance of data distribution. Prompted by recent advances in learning more robust and higher-level representations in deep learning, especially deep autoencoder [11], we proposed the use of deep autoencoder for ASAG, named Deep Autoencoder Grader (DAGrader). Both accuracy and Quadratic weighted Kappa (QWKappa) are used to measure the grading model.

2 Grading Model with Deep Autoencoder

In this paper, we consider ASAG task as a text classification problem. The classification algorithm based on deep autoencoder [11] is employed, which is shown in Fig. 1. The deep autoencoder consists of two encoding and decoding layers. The first and second hidden layer are the encoding layers. The first encoding layer is the embedding layer, where the lower-dimensional representations of student responses are learnt. The lower-dimensional representations of student responses can retain the most salient information of the input data. The second encoding layer is the label encoding layer, where the label information (i.e., the score of the student response) is encoded using a softmax regression [12]. In addition, the

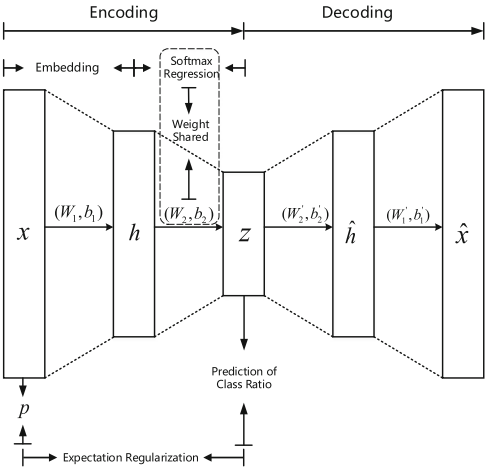

Fig. 1. Framework of the proposed model.

encoding weights in the second hidden layer are also used for the final prediction model. The third and fourth layer are the decoding layers, where the outputs of the first and second hidden layers are reconstructed respectively.

In order to train the deep autoencoder model, there are four factors to be considered in the loss function, i.e., the reconstruction error, the loss function of softmax regression, the expectation regularization of class ratio and the model parameter regularization. Specifically, the loss function of softmax regression can incorporate the label information of student responses into the embedding space; the expectation regularization of class ratio is introduced for reducing the influence of data imbalance.

After the parameters of deep autoencoder are learnt, we can obtain the lower-dimensional representation of student responses from the first encoding layer. Two methods can be utilized to construct the text classifiers for automatic grading task, which are named as DAGrader_1 and DAGrader_2. The first method DAGrader_1 is to use the second hidden layer's output $\boldsymbol{z}$. The corresponding label of the maximum element of $\boldsymbol{z}$ is the predicted label of the input instance. The second method DAGrader_2 is to use the lower-dimensional representation of student responses to train a classifier by applying standard classification algorithms. We use random forest with 100 trees in this paper.

3 Data Description and Preprocessing

Our corpus consists of five data sets. Each of the data sets was generated from a reading comprehension question. All responses were written by students in Grade 8. In order to ensure the reliability of the label, all responses were hand graded by two experienced human raters. The details of all data sets are showed in Table 1.

Table 1. Overview of all datasets

Item ID	#Samples	Grading scheme	#Avg-words	#Unigram features	#QWKappa
1	2579	5-point	39	1071	0.9847
2	2571	3-point	33	1644	0.9723
3	2382	4-point	26	618	0.9427
4	2458	5-point	27	655	0.9733
5	2538	4-point	31	768	0.8319

To obtain the input of the proposed algorithm, a series of standard natural language preprocessing methods are conducted on data sets, including punctuation removal, stop word removal and tokenization. The student responses are preprocessed by using a parser called "jieba"[1], which is a Python Chinese word

[1] https://github.com/fxsjy/jieba.

segmentation module. Then, we utilize a Python module named scikit-learn[2] to extract the n-gram features of the student responses.

4 Results and Discussion

We compare our model, denoted as DAGrader_1 and DAGrader_2, with several automatic scoring models. LR (Logistic Regression) and SVM (Support Vector Machine) are two efficient and well-known ASAG models. Yang et al. proposed an ASAG model based on LSTM without grading rubrics in [13]. We utilize continuous bag-of-words model(CBOW) [14] to expand Yang's model. CBOW_a and CBOW_w are trained on our corpus and Chinese wikipedia corpus, respectively.

All the results of these five data sets are shown in Table 2, and we have the following observations,

(1) DAGrader is significantly better than LR on every data set, which indicates the efficiency of our proposed ASAG framework.
(2) SVM performs better than LR, which demonstrates the grading results can be improved by applying a better text classifier. Yang's expanded model CBOW_w has higher performance than the corresponding figures of LR and SVM, indicating the importance of extracting deep semantic feature of student answers.
(3) DAGrader_1 and DAGrader_2 outperforms all the baselines in term of accuracy, which shows that our proposed model can combine the merits of conventional bag-of-words models and deep learning models.

Table 2. Accuracy and QWKappa on all data sets

Item ID	LR	SVM	CBOW_a	CBOW_w	DAGrader_1	DAGrader_2
Accuracy (%)						
1	55.86	54.82	57.17	62.33	**64.65**	62.23
2	66.47	65.85	71.06	71.84	71.60	**71.84**
3	81.82	88.62	84.84	88.92	**89.50**	88.19
4	57.53	58.67	62.21	61.87	63.21	**66.67**
5	76.40	76.60	74.31	80.77	**81.50**	80.54
Avg.	67.62	68.91	69.92	73.15	**74.09**	73.89
QWKappa						
1	0.3697	0.4015	0.2213	0.4431	0.5185	**0.5221**
2	0.3915	0.4254	0.3752	0.4825	0.4915	**0.4940**
3	0.7913	**0.8680**	0.7276	0.8364	0.8539	0.8407
4	0.5142	0.5789	0.5693	0.5612	0.6257	**0.6599**
5	0.6270	0.6522	0.4214	0.6754	**0.7360**	0.7056
Avg.	0.5387	0.5852	0.4630	0.5997	**0.6451**	0.6445

[2] http://scikit-learn.org/stable/index.html.

5 Conclusions

In this paper, we tackle the ASAG task by using a deep autoencoder. Our method does not rely on any target answer due to the fact that target answers are not always available. Specifically, there are two layers for encoding in the deep model, one is for embedding and the other is for label encoding. In the embedding layer, we can get the lower-dimensional representations of student responses, which can be used for text classifier construction. In the label encoding layer, we can easily incorporate the label information into the text representation. Additionally, to reduce the impact of data imbalance, we introduce an expectation regularization of class ratio term into the loss function of the deep autoencoder. Experiments on five Chinese data sets demonstrate the effectiveness of the proposed method.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61773361, 61473273), the Youth Innovation Promotion Association CAS 2017146, the China Postdoctoral Science Foundation (No. 2017M610054).

References

1. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**(1), 60–117 (2015)
2. Sung, K.H., Noh, E.H., Chon, K.H.: Multivariate generalizability analysis of automated scoring for short answer items of social studies in large-scale assessment. Asia Pac. Educ. Rev. **18**(3), 425–437 (2017)
3. Liu, O.L., Rios, J.A., Heilman, M., Gerard, L., Linn, M.C.: Validation of automated scoring of science assessments. J. Res. Sci. Teach. **53**(2), 215–233 (2016)
4. Madnani, N., Burstein, J., Sabatini, J., O'Reilly, T.: Automated scoring of a summary-writing task designed to measure reading comprehension. In: BEA@ NAACL-HLT, pp. 163–168 (2013)
5. Jimenez, S., Becerra, C.J., Gelbukh, A.F., Bátiz, A.J.D., Mendizábal, A.: Soft-cardinality: hierarchical text overlap for student response analysis. In: SemEval@ NAACL-HLT, pp. 280–284 (2013)
6. Dzikovska, M.O., Nielsen, R.D., Brew, C.: Towards effective tutorial feedback for explanation questions: a dataset and baselines. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 200–210. Association for Computational Linguistics (2012)
7. Bailey, S., Meurers, D.: Diagnosing meaning errors in short answers to reading comprehension questions. In: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, pp. 107–115. Association for Computational Linguistics (2008)
8. Zesch, T., Levy, O., Gurevych, I., Dagan, I.: UKP-BIU: similarity and entailment metrics for student response analysis, Atlanta, Georgia, USA, p. 285 (2013)
9. Hou, W.-J., Tsao, J.-H., Li, S.-Y., Chen, L.: Automatic assessment of students' free-text answers with support vector machines. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds.) IEA/AIE 2010. LNCS (LNAI), vol. 6096, pp. 235–243. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13022-9_24

10. Zhang, Y., Shah, R., Chi, M.: Deep learning + student modeling + clustering: a recipe for effective automatic short answer grading. In: EDM, pp. 562–567 (2016)
11. Zhuang, F., Cheng, X., Luo, P., Pan, S.J., He, Q.: Supervised representation learning: transfer learning with deep autoencoders. In: International Conference on Artificial Intelligence, pp. 4119–4125 (2015)
12. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
13. Yang, X., Zhang, L., Yu, S.: Can short answers to open response questions be auto-graded without a grading rubric? In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 594–597. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_72
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Comput. Sci. (2013)