Attention-Based Bidirectional Long Short-Term Memory Neural Network for Short Answer Scoring

Linzhong Xia(✉), Mingxiang Guan, Jun Liu, Xuemei Cao, and Dean Luo

Shenzhen Institute of Information Technology, Shenzhen 518172, China
xialz@sziit.edu.cn

Abstract. The automatic short answer scoring by using computational approaches has been considered the best way to release the workload of human answer raters. In this paper, we designed a novel neural network architecture which is attention-based bidirectional long short-term memory to implement the task of automatic short answer scoring. We evaluate our approach on the Kaggle Short Answer dataset (ASAP-SAS). Our experiment results indicate that our model can scoring short answers more accurately in terms of the quality of the results. Meanwhile, our experiment results demonstrate that our model is more effective and efficient than other baseline methods in most cases.

Keywords: Natural language processing · Short answer scoring · Long Short-Term memory · Attention mechanism · Quadratic Weighted Kappa

1 Introduction

The short answer examinations are considered as an essential part in the educational processes. It can help students to check the mastery of problem. But now, there are a large of short answers which are produced by students need to be examined. It is a large workload for human answer raters [1].

To solve above problem, many answer examination computational approaches have been invented. Those approaches are regarded as a machine learning approach, such as classification, or linear regression. Most of machine learning approaches make use of various statistical features, such as total words number, different words number, words spelling errors, average of words syllables, sentences frequency, essay's length, term frequency-inverse document frequency (TF-IDF), and so on. Short answer correction by this type of methods are usually seen as a linear regression problem. The commercial correction systems such as Project Essay Grader (PEG) [2] and E-rater [3] are the typical Representatives. However, the obvious drawback of the PEG and E-rater is that they both can't extract the semantic features of answers.

In order to extract the semantic features from answers, latent semantic analysis (LSA) [4] had been invented. The key idea of LSA is the analysis of the underlying semantic of answers. LSA solves the problem of synonymy or polysemy by mapping the same answers or words into a different space and doing the comparison in the space

M. Guan and Z. Na (Eds.): MLICOM 2020, LNICST 342, pp. 104–112, 2021.
https://doi.org/10.1007/978-3-030-66785-6_12

with the method of Singular Value Decomposition (SVD) of term-essay matrix. In the process of mapping, the high dimensional answer vector will be transformed to low dimensional vector. The commercial correction systems such as intelligent essay assessor (IEA) [5] is the typical Representatives. However, there are some limitations for LSA. First, the computational cost of the SVD is expensive. Second, word order is not taken into consideration in the space representation.

The limitations of LSA had been solved by the approach of probabilistic LSA (pLSA) [6] or latent Dirichlet allocation (LDA) [7]. The different with LSA is that the LDA can capture the exchangeability of both words and answers. LDA is a generative probabilistic model of a corpus. The answers of corpus are represented as random mixtures over latent topics. Each latent topic of answers is characterized by a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well.

In a word, the final representation of an answer consists of a features vector that have been selected and tuned by human experts to assess a score in a marking scale [8]. Although those approaches have achieved performance comparable to human raters, there is essential manual effort involved in obtaining these results on different fields. It is hard to select or tune suitable linguistic features by human experts in specific fields. In order to perform well on specific data, separate models with distinct feature sets are needed.

But now, the most effective approach to solve the problems mentioned above paragraph is the technology of deep neural network (DNN) [9]. And DNN has achieved remarkable advances in the field of automatic answer scoring (AAS). The ability of AAS systems which based on DNN have surpass state-of-the-art models in similar areas [10]. In a sense, answer scoring is like answer classification. In answer classification field, most of the studies are focus on how to learn word vector representations by neural language models [11] and how to classify the answers based on the learned word vectors [12]. Therefore, the basic of answer classification is find a highly efficient word vector representations. At present, the most popular word vector representation models are Word2Vec [13], C&W [14], GloVe [15], and so on. Before starting answer classification, we can use the word vectors which trained on external big dataset [16, 17] or trained on local [18] dataset to represent answers. Then taking the word vectors as the input of DNN models which are used to execute the task of answer classification. There are two types of DNN models for answer classification: recurrent neural network (RNN) and convolutional neural networks (CNN). In recent years, much great results have been achieved in text classification by using RNNs or CNNs [19–22]. CNN is good at handling spatial data but not good at handling sequence data. In contrast of CNN, RNN is good at handling sequence data which makes it a more 'natural' approach when dealing with textual data since answer is naturally sequential. The answer classification can be considered as a sequential modelling problem. Due to the characteristic of RNN, we usually use RNN to process the task of answer classification. However, RNN can capture the correlations of short-term dependencies but the ability to learn the long-term dependencies is weakly. Long short-term memory (LSTM) [23] is a special kind of RNN and it is powerful to learn the correlations of long-term dependencies. Now, LSTM has been widely used in text classification field [24, 25]. Moreover, bidirectional long short-term memory (BiLSTM) [26] can run inputs in two

ways, one from past to future and one from future to past. Therefore, you are able in any point in time to preserve information from both past and future by combining the two hidden states [27].

Ideally, we hope to obtain the word vectors which can retain all contextual information of answer. In fact, represent ability of different word vectors which are obtained by training in different datasets is different. Therefore, well-suited dataset is the key point to obtain the powerful word vectors. Google word vector (GWV) which pre-trained on roughly 100 billion words from a Google News dataset includes a vocabulary of 3 million words and phrases has advantages over local-trained word vector (LWV) which trained on local dataset in most cases. Although GWV can obtain powerful representation of the contextual information of the essay, it is not possible to focus on the important information in the obtained contextual information. Therefore, attention mechanism (AM) [28, 29] which is used to focus on the important information of the contextual information has been recommended. The combination of GWV and AM can further improve the ability of answer classification.

In this paper, we propose a novel DNN architecture for short answer scoring. This new architecture is an enhanced BiLSTM by using GWV and AM, referred to as attention-based BiLSTM Neural Network with GWV as the input layer (GA-BiLSTM). The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 introduces the propose approach in detail. Section 4 shows the result of the experiment. Section 5 is conclusions.

2 Related Work

The AAS has been pushed into a new stage by the invention of DNN. In recent years, there are many researchers have proposed many approaches to improve DNN. Zhang et al. [30] proposed a deep belief networks (DBN) to tackle the task of AAS. Their experimental results show that DBN is the best model than other models which mentioned in the article. Xiang et al. [31] apply a temporal convolutional network to various large-scale datasets, including ontology classification, sentiment analysis, and text categorization. Their model can achieve astonishing performance without the knowledge of words, phrases, sentences and any other syntactic or semantic. Walia et al. [32] proposed a BiLSTM to achieve an efficient AAS for Punjabi language. Surya K et al. [33] compared some DNN techniques for AAS task. The results show that BERT (bidirectional encoder representations from transformers) performs better than CNN and LSTM.

3 Automatic Short Answer Scoring Model

In this section, we introduce our AAS model as shown in Fig. 1. The designed novel architecture is named attention-based BiLSTM with GWV as the input layer (GA-BiLSTM). The model consists of four components, word embedding layer, BiLSTM, attention layer, as well as Softmax layer.

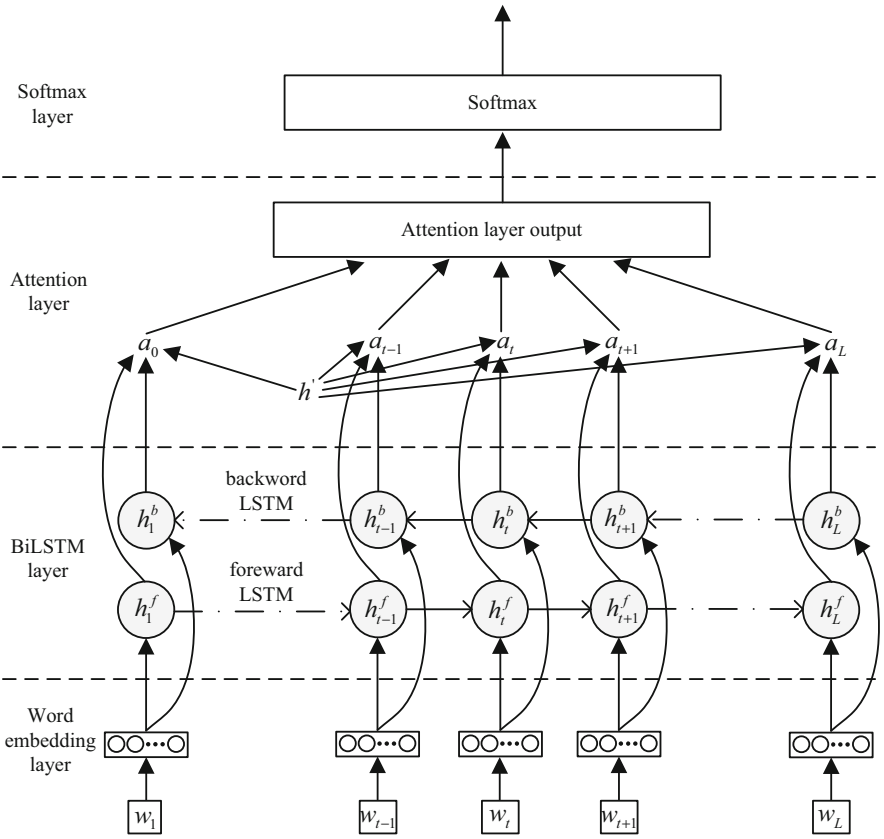

Fig. 1. The architecture of the GA-BiLSTM

The word embedding layer converts words of answer to word vectors and takes them as the input of the BiLSTM layer. In natural language processing field, how to represent words of text is important. Usually, there are two ways to represent words: one-hot representation and distributed representation. Compared to one-hot representation, distributed representation of word embedding not only reflect the characteristics of low dimension but also reflect the orderly relationship between words of text. In this paper, we use the word vector which trained on Google News dataset in experiments.

The BiLSTM layer accesses the contextual features. The contextual features extraction capability of BiLSTM is more powerful than LSTM. Because the BiLSTM can access to both history and future of the given time frame throughout backward and forward layers together.

The Attention layer has been used to pay more attention to information units which strongly related to the score of the answer. Because AM can focus on the important words to reduce the impact of unimportant words. In this paper, we provide an AM layer after the output of BiLSTM to learn a weight combination of hidden states across all time steps and produce a global feature vector.

The Softmax layer is to classify the output of the AM layer.

4 Experiments

In this section, the experiments setup, baseline methods and experiment results are described.

4.1 Experiments Setup

Dataset. In this paper, the used dataset is come from Automated Student Assessment Prize (ASAP). The dataset consists of 10 prompts as listed in Table 1. There are total of 17207 training answers. The prompts scored on either a scale of 0–2 or 0–3. There are no released labeled test data from the ASAP competition. Therefore, we just divided the released dataset into three parts: 60% for training set, 20% for the validation set and 20% for the test set.

Table 1. Description of ASAP dataset

Prompt	Subject	Answers	Avg length	Scores
1	Science1	1672	50	0-3
2	Science2	1278	50	0-3
3	English Language Arts1	1891	50	0-2
4	English Language Arts2	1738	50	0-2
5	Biology1	1795	60	0-3
6	Biology2	1797	50	0-3
7	English1	1799	50	0-2
8	English2	1799	50	0-2
9	English3	1798	40	0-2
10	Science3	1640	60	0-2

Evaluation Metrics. In our experiment, we use Quadratic Weighted (QW) Kappa to evaluate our results. At the same, the QW Kappa is also used in ASAP competition. The QW Kappa is an index to measure the agreement between the raters. The result of QW Kappa equal to 1 when the scores of raters are identical. The result of QW Kappa equal to 0 when the scores of raters are totally inconsistent.

Parameter Settings. We adopt the Adam optimizer as the optimization algorithm. The best hyper-parameters obtained through multiple experiments are shown in Table 2. The initial learning rate is set to 0.01. The learning rate is reduced by 0.05 times once every 1 epoch.

Table 2. The hyper-parameter settings

Layer	Parameter Name	Value
Word Embedding	GWV dimension	300
BiLSTM layer	Hidden units	128
Dropout	Dropout rate	0.5
Others	Epochs	30–100
	Batch size	32
	Initial learning rate	0.01

4.2 Baselines

We use several baseline methods as the benchmarks, they are effective methods for AAS. CNN: the approach uses a CNN to score the student answers [33]. BiLSTM: the approach uses a BiLSTM to score the student answers [33]. BERT: the approach uses a BERT to score the student answers [33]. We name our model described in Sect. 3 as GA-BiLSTM. The results of those baseline models are listed in Table 3.

Table 3. Experiment results of all compared models on the ASAP dataset. Best result is in bold.

Model	Prompts										Aver.
	1	2	3	4	5	6	7	8	9	10	
CNN	0.68	0.67	0.27	0.55	0.75	0.74	0.58	0.50	0.64	0.67	0.61
BiLSTM	0.70	0.66	0.28	0.59	0.78	0.74	0.60	0.54	0.70	0.71	0.63
BERT	0.79	0.70	0.37	0.69	0.75	0.84	0.66	0.60	0.80	0.74	0.69
GA-BiLSTM	0.81	0.71	0.47	0.67	0.79	0.85	0.61	0.58	0.78	0.76	0.70

4.3 Results

The comparison results are presented in Table 3. The experiment results are evaluated by QWK. The font which is bold presents the best experiment results as the Table 3 shown. All approaches, including our approach, are deep neural network approaches. In Table 3, we can find that our approach of GA-BiLSTM has achieved the best result than other approaches on the average QWK value. Among the four approaches mentioned in Table 3, our approach outperforms other baseline models on all prompts except prompt4, prompt7, prompt8, and prompt9. The results of GA-BiLSTM are 0.81, 0.71, 0.47, 0.79, 0.85, and 0.76 for prompt1, prompt2, prompt3, prompt5, prompt6, and prompt10. Compared to BERT, the QWK value of GA-BiLSTM obtains the relative improvements of 2.5%, 1%, 27%, 5.3%, 1.2%, 2.7%, respectively. In a word, the results of our model outperform most of the published baseline models. In Table 3, we can conclude that the overall performance of GA-BiLSTM is better than other approaches in term of the average QWK value.

As the results show, the combination of GWV, BiLSTM architecture and attention mechanism has remarkably improved the performance of AAS. For most of the benchmark prompts, GA-BiLSTM can obtain better results than other baseline models. And GA-BiLSTM obtains the best result for the average QWK value.

5 Conclusions

In this paper, we augmented the input by using GWV as the representation of the contextual information of answers. It is expected to learn more key information of answers by using BiLSTM which can in any point in time to preserve information from both past and future by combining two hidden states. The themes of some prompts of ASAP dataset are clear. To this kind of short answers, the technique of AM is an efficient choice to extract the key themes information. Therefore, word vector, neural network architecture, and the design of classifier are the key point for AAS. Our experiment results indicate that our model can scoring short answers more accurately in terms of the quality of the results. Meanwhile, GA-BiLSTM demonstrates that our model is more effective and efficient than other baseline methods in most cases.

Acknowledgement. This work is supported by Engineering Applications of Artificial Intelligence Technology Laboratory of Shenzhen Institute of Information Technology (Number: PT201701), the Guangdong Province higher vocational colleges & schools Pearl River scholar funded scheme (2016), and The Scientific and Technological Projects of Shenzhen (No. JCYJ20190808093001772).

References

1. Dikli, S.: An overview of automated scoring of essays. J. Technol. Learn. Assess. **5**(1), 1–35 (2006)
2. Page, E.B.: The imminence of grading essays by computer. Phi Delta Kappan **48**, 238–243 (1966)
3. Claudia, L., Martin, C.: C-rater: Automated scoring of short-answer questions. Comput. Humanit. **37**(4), 389–405 (2003)
4. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
5. Landauer, T., Laham, D., Foltz, P.: Automated scoring and annotation of essays with the intelligent essay assessor. In: Automated Essay Scoring: A Cross-Disciplinary Perspective, pp. 87–112 (2003)
6. Hofmann T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. Association for Computing Machinery ACM, Berkeley (1999)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
8. McNamara, D., Crossley, S.A., Mccarthy, P.M.: Linguistic features of writing quality. Written Commun. **27**(1), 57–86 (2010)

9. Gomaa, W.H., Fahmy, A.A., Ans2vec: a scoring system for short answers. In: Hassanien, A., Azar, A., Gaber, T., Bhatnagar, R., F. Tolba, M. (eds) AMLTA 2019, vol. 821, pp. 586–595. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-14118-9_59
10. Tang, D.: Sentiment-specific representation learning for document-level sentiment analysis. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 447–452. Association for Computing Machinery (ACM), Shanghai (2015)
11. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 115–124. Association for Computational Linguistics (ACL), Ann Arbor (2005)
12. Lee, K., Han, S., Myaeng, S.-H.: A discourse-aware neural network-based text model for document-level text classification. J. Inf. Sci. **44**(6), 715–735 (2018)
13. Mikolov T., Chen K., Corrado G., Dean J.: Efficient estimation of word representations in vector space. arXiv:1301.3781[cs.CL], 1–12 (2013)
14. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)
15. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543. Association for Computational Linguistics (ACL), Doha (2014)
16. Zhang, H., Litman, D.: Co-attention based neural network for source-dependent essay scoring. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 399–409. Association for Computational Linguistics (ACL), New Orleans (2018)
17. Ali, M.N.A., Tan, G.Z., Hussain, A.: Bidirectional recurrent neural network approach for Arabic named entity recognition. Future Internet **10**(12), 123 (2018)
18. Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 7–12. Association for Computational Linguistics (ACL), Berlin (2016)
19. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1746–1751. Association for Computational Linguistics (ACL), Doha (2014)
20. Liao, S., Wang, J., Yu, R., Sato, K., Cheng, Z.: CNN for situations understanding based on sentiment analysis of twitter data. In: Proceedings of the 8th International Conference on Advances in Information Technology, Elsevier B.V., pp. 376–381. Macau (2016)
21. Zhang, Y., Wallace, B.C.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv:1510.03820[cs.CL], pp. 1–18 (2016)
22. Zhang, Y., Er, M.J., Venkatesan, R., Wang, N., Pratama, M.: Sentiment classification using comprehensive attention recurrent models. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1562–1569. IEEE, Vancouver (2016)
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
24. Ran, X., Shan, Z., Fang, Y., Lin, C.: An LSTM-based method with attention mechanism for travel time prediction. Sensors **19**(4), 861 (2019)
25. Nowak, J., Taspinar, A., Scherer, R.: LSTM recurrent neural networks for short text and sentiment classification. In: Proceedings of the 16th International Conference on Artificial Intelligence and Soft Computing, pp. 553–562. Springer Verlag, Zakopane (2017)
26. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)

27. Bin, Y., Yang, Y., Shen, F., Xie, N., Shen, T., Li, X.: Describing video with attention-based bidirectional LSTM. IEEE Trans. Cybern. **49**(7), 2631–2641 (2019)
28. Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421. Association for Computational Linguistics (ACL), Lisbon (2015)
29. Yin, W., Ebert, S., Schütze, H.: Attention-based convolutional neural network for machine comprehension. In: Proceedings of the Workshop on Human-Computer Question Answering, pp. 15–21. Association for Computational Linguistics (ACL), San Diego (2016)
30. Zhang, Y., Shah, R., Chi, M.: Deep learning + student modeling + clustering: a recipe for effective automatic short answer grading. In: Proceedings of the 9th International Conference on Educational Data Mining, pp. 562–567. International Educational Data Mining Society (IEDMS), Raleigh (2016)
31. Zhang, X., LeCun, Y.: Text Understanding from Scratch. arXiv:1502.01710 [cs.LG] (2016)
32. Walia, T.S., Josan, G.S., Singh, A.: An efficient automated answer scoring system for Punjabi language. Egyptian Inf. J. **20**, 89–96 (2019)
33. Surya, K., Ekansh, G., Nallakaruppan, K.: Deep learning for short answer scoring. Int. J. Recent Technol. Eng. **7**(6), 1712–1715 (2019)