Automatic Chinese Reading Comprehension Grading by LSTM with Knowledge Adaptation

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

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

Abstract. Owing to the subjectivity of graders and the complexity of assessment standard, grading is a tough problem in the field of education. This paper presents an algorithm for automatic grading of open-ended Chinese reading comprehension questions. Due to the high complexity of feature engineering and the lack of consideration for word order in frequency based word embedding models, we utilize long-short term memory recurrent neural network to extract semantic feature in student answers automatically. In addition, we also try to impose the knowledge adaptation from web corpus to student answers, and represent the students' responses to vectors which are fed into the memory network. Along this line, the workload of teacher and the subjectivity in reading comprehension grading can both be reduced obviously. What's more, the automatic grading methods for Chinese reading comprehension will be more thorough. The experimental results on five Chinese and two English data sets demonstrate the superior performance over compared baselines.

Keywords: Automatic grading · Knowledge adaptation
Reading comprehension · LSTM · Text classification

1 Introduction

It is a tough problem in the field of education when grading student answers because of the subjectivity of graders and the complexity of assessment standards. Particularly, with the evolution of e-learning and online examination, the demand for assessment is increasing. It is apparent that hiring a great number of teachers is not a cost-efficient way. A growing number of researchers

D. Phung et al. (Eds.): PAKDD 2018, LNAI 10937, pp. 118–129, 2018.
https://doi.org/10.1007/978-3-319-93034-3_10

have engaged in automatic grading, especially automatic reading comprehension grading [1,10,16–18]. Nevertheless, there is little research on automatic open-ended Chinese reading comprehension grading. Therefore, it is fairly essential to develop automatic grading methods for open-ended Chinese reading comprehension questions.

Table 1. Sample of open-ended Chinese reading comprehension

Text Information
黄阿二的酒酿在古庙镇上老老少少都翘起大拇指 (…) 他那极有韵味的吆喝可以说已成了古庙镇的一种文化风景 (…) 黄阿二听后浑身一震, 他撑起身子说:“你们这一声吆喝, 对我来说, 比吃啥药都强, 这不, 毛病好了一半.”
Text Information Translation
People in Gumiao town thumb up for the sweet ferment rice of Huang A'er(…) his lasting appeal of yo-heave-ho has became a kind of cultural landscape of Gumiao town, after hearing that(…) Huang A'er shocks his body and raises body to say: “your shouting is better than any other medicine for me, look, illness has been improved half.”
Question
结合 “他那极有韵味的吆喝可以说已成了古庙镇的一种文化风景” 这句话, 结合你的生活体验, 请说一说你是怎么理解 “酒酿王的吆喝声” 的?
Question Translation
Combine with the sentence of “his lasting appeal of yo-heave-ho has became a kind of cultural landscape of Gumiao town” and your life experience, please express your understanding about the “yo-heave-ho of sweet ferment rice king”.
Example of Full-Score Answer
我认为酒酿王的吆喝声是人的勤劳, 人与人间无形的热情. 生活中, 卖东西的小贩沿街叫卖, 无论天气季节, 都坚持和做交易时的样子, 都使人感到暖意.
Example of Full-Score Answer Translation
I think his yo-heave-ho shows diligent and intangible enthusiasm among people. Vendors' shouting along street and their insistence on trading no matter how is the weather and what is the season can make people feel warm-hearted in everyday life.

The reading comprehension question is the one that students should read a text and answer the questions about it. The questions can be designed with different openness. In this paper, we concentrate on the open-ended reading comprehension questions. As the example shown in Table 1, after reading a text, the students may be asked to express their understanding about the given text according to their actual experience. The students may answer this question differently since their different experiences and understandings. Moreover, the sentence semantic and words used by different students maybe divergent and diverse, which make the open-end reading comprehension actually no reference answers.

In the past few years, most of methods for automatic reading comprehension grading are based on two assumptions: (1) Reading comprehension is close-ended and can be graded by recognizing some specific words without considering word

orders in student answers. (2) Reading comprehension questions would provide graders with reference answer. However, a large proportion of Chinese reading comprehension questions is open-ended and there is few specific words among student answers. Under this circumstance, automatic grading by specific words based on bag-of-words models without word orders would be invalid. What's worse, most of Chinese comprehension lack of reference answer, thus the automatic grading methods based on reference answers would not work.

Based on the analysis above, we have the following motivations to propose a new automatic reading comprehension grading model:

(1) These are enormous demands for automatic reading comprehension grading.
(2) It is significant to propose a framework for grading reading comprehension without reference answers.
(3) It is crucial to take word orders into consideration for automatic reading comprehension grading methods.

Along this line, we try to formalize the automatic open-ended Chinese reading comprehension grading problem as text classification. In this way, the algorithm can be conducted without reference answers. For word level, we represent each word as a vector trained by continuous bag-of-words model (CBOW) [20]. For sentence level, long-short term memory recurrent neural network (LSTM) [9] is used to model student answers by calculating the word embedding based on CBOW. LSTM is well approved to model sequence data, which can be used to extract word order features in student answers. We validate our model on seven data sets, including Chinese and English reading comprehension questions. The extensive experimental results demonstrate the superiority of our method over several state-of-the-art baselines in terms of QWKappa (Cohen's kappa with quadratic weight), accuracy, precision, recall, and F1-score.

The remainder of this paper is organized as follows. Section 2 introduces the framework and its solution details. The experimental results are reported in Sect. 3. Section 4 discusses the related work and finally Sect. 5 concludes.

2 Model for Automatic Open-Ended Chinese Reading Comprehension Grading

Figure 1 shows the framework and the training process of our proposed model, which is constructed based on continuous bag-of-words model (CBOW) [19] and long-short term memory recurrent neural network (LSTM) [9]. CBOW is a word embedding model, which represents each word as a vector. These word vectors are fed into LSTM in sequence. LSTM is a variant of recurrent neural network (RNN), and we take the advantage of LSTM to extract semantic information of each student answer. We use Adam in [14] to minimize the cross-entropy [4] loss function. In order to train the CBOW model more quickly and accurately, we utilize knowledge adaptation to transfer the external knowledge from web corpus to student answers (target corpus). Next, we will introduce the details of our model.

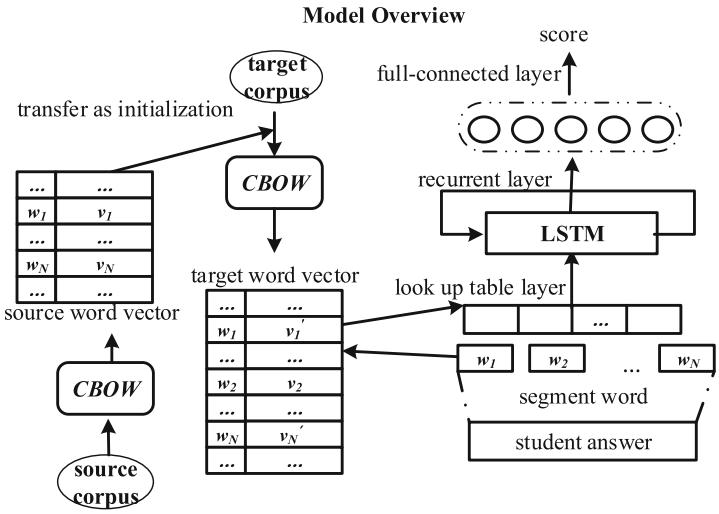

Fig. 1. Framework of our proposed model.

2.1 Negative Sampling Based Continuous Bag-of-Words Embedding

One hot encoding is a prevalent word representation method for neural network based natural language processing tasks. It encodes each word as a vector by marking at its index in vocabulary. However, encoding words in this way can not measure the distances among words. What's worse, it may be high-dimensional as the growth of vocabulary. Hence many word representation models are proposed for estimating continuous representations of words, including the prominent Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) [19]. In this paper, we use distributed representations of word learned by deep learning. Distributed representations proved to perform better than LSA for preserving linear regularities among words [19,21,28].

According to [19], the basic CBOW model is similar to the feed-forward neural network language model (NNLM) in [3], where the non-linear hidden layer is removed and the projection layer is shared for all words. In this paper, we also utilize negative sampling proposed in [20] for CBOW model that results in faster training and better vector representation for frequent words. Next, we would like to introduce the CBOW model more clearly.

Input Layer: For one of the target word w_k in a sentence, there are $2c$ words including c precedent words $w_{k-c}, ..., w_{k-2}, w_{k-1}$ and c posterior words $w_{k+1}, w_{k+2}, ..., w_{k+c}$ are used as context words of target word w_k, and we initialize the vector for each word in d-dimension randomly. These vectors will be updated by back propagation and used as input in look-up table layer.

Project Layer: Averaging all word vectors in each position to a vector v_k according to Eq. (1)

$$v_k = \frac{1}{2c} \sum_{i=k-c, i \neq k}^{k+c} v_i, \tag{1}$$

where v_i is the vector of word w_i.

Output Layer: For all words in the corpus, the CBOW model try to predict each target word w_k based on the context information v_k. The loss function is as Eq. (2):

$$\mathcal{L}_{CBOW} = log \prod_{k=1}^{|D|} \{\delta(v_k^T \theta^k) \prod_{j=1}^{|NEG_k|} [1 - \delta(v_k^T \theta^j)]\}, \tag{2}$$

where D is the whole word vocabulary for a corpus, NEG_k is the negative sampling set of target word w_k. Although the noise words in negative sampling set for each word is diverse, the total number of noise words $|NEG_k|$ are the same. $\delta(v_k^T \theta^k)$ is the likelihood of correct prediction, while $\delta(v_k^T \theta^j)$ is the likelihood of negative prediction and j is the word index in negative sampling set. The model maximizes the $\mathcal{L}_{CBOW}$, so that the likelihood of correct predication would be maximized and the likelihood of negative predication would be minimized at the same time. The back propagation would update context information v_k and the random initialized word vector during optimizing the CBOW model.

2.2 Knowledge Adaptation for Continuous Bag-of-Word Embedding

According to [25], the parameters in earlier layers of neural network which trained on large data sets are general to different tasks. Therefore, using existing parameters for initialization can benefit performance improvement and time-saving. Knowledge adaptation is a technique that aims to adapt pre-trained models to new natural language processing tasks.

As we have mentioned in the last subsection, the continuous bag-of-words (CBOW) model contains three layers, and weights of each layer are updated by back propagation. The word vectors initialized on input layer would be updated as well. Hence we first train CBOW model on large scale corpora (wikipedia) and transfer the existing word vectors to input layers, and add new initialized word vectors to target automatic grading task. In this way, the existing vectors transferred from large corpora would complete training quickly and the vectors for new words on target tasks would be trained precisely with accurate pre-trained context word vectors. Finally, the external knowledge learned from large scale corpora would be adapted for new automatic grading tasks.

2.3 Recurrent Layer

Frequency based word embedding models are well-known for statistic machine learning based automatic grading methods. Boolean vectorization is a model

that reflects whether a word occurs in a document or not. And term frequency-inverse document frequency (TF-IDF) [13] is a numerical statistic that reflects how important a word is to a document in a collection or corpus. Since these models ignore context information in student answers, therefore we utilize the excellent sequence modeling technique LSTM to learn the semantic features of student answers.

LSTM contains special units called memory blocks in the recurrent hidden layer of RNN, which further contains memory cells with self-connections to store the temporal state of the network, and special multiplicative units called gates to control the information flow. Every block in the architecture contains an input gate and an output gate. The input gate controls the flow of input activations into memory cell, and the output gate controls the output flow of cell activations into the rest of the network [9]. The LSTM functions are as follows,

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ c'_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c), \quad c_t = i_t \circ c'_t + f_t \circ c_{t-1} \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad h_t = o_t \circ \tanh(c_t) \end{aligned} \tag{3}$$

where x_t and h_t are the input and output vectors at time t, W_i, W_f, W_c, W_o, U_i, U_f, U_c, and U_o are weight matrices and b_i, b_f, b_c, and b_o are bias vectors, $\circ$ is the element-wise multiplication, and σ represents the sigmoid function.

2.4 Fully-Connected Layer with Softmax Activation

After learning sequence features from student answers, we utilize a fully-connected layer and softmax activation [6] to calculate the output probability of each score. Assuming that there are K possible scores for each answer, and the output is a K-dimension vector as follows,

$$h_\theta(x_i) = [P(y_i = 0|S_i;\theta), P(y_i = 1|S_i;\theta), \cdots, P(y_i = K-1|S_i;\theta)]^\top \tag{4}$$

where $P(y = k|S;\theta)(k = 0, 1, \cdots, K-1)$ is the probability of each score for a student answer, and θ indicates the parameter in fully-connected layer. Finally, we use the Adam optimization algorithm [14] to minimize the cross-entropy [4] loss function on training data.

3 Experiments

To validate the effectiveness of our proposed model, we conduct experiments on seven data sets and compare it with several state-of-the-art baselines. Moreover, we also compare our proposed model with CBOW without knowledge adaptation, TCBOW train on student answers and SCBOW train on web corpus.

3.1 Data Sets and Preprocessing

The details of all data sets are showed in Table 2, where "Avg#word" denotes the average number of words for student answers, "#samples" denotes the total number of student answers for each data set, and "QWKappa" denotes the grading consistency between two graders. For Chinese data sets, we utilize Chinese word segmentation system jieba[1] to segment Chinese words. For English data sets, we use scikit-learn[2] to tokenize and extract answer features.

Table 2. Overview of all datasets

Problem	Avg#word	#samples	#Score level	Language	QWKappa
CRCC1	39	2579	0–4	Chinese	0.9847
CRCC2	33	2571	0–2	Chinese	0.9723
CRCC3	26	2382	0–3	Chinese	0.9427
CRCC4	27	2458	0–4	Chinese	0.9733
CRCC5	31	2538	0–3	Chinese	0.8319
ASAP-SAS3	47	2297	0–2	English	–
ASAP-SAS4	40	2033	0–2	English	–

CRCC Data Set: To evaluate our model in Chinese answers, we construct a Chinese Reading Comprehension Corpus (CRCC). In order to ensure the reliability of the grading label, two Chinese teachers were asked to grade the answer individually. The consistency of the two teachers' scoring is evaluated by QWKappa. The QWKappa scores of CRCC are shown in Table 2, which demonstrates the label of each data set is reliable (i.e., the value of QWKappa is required to larger than 0.8). Higher value of QWKappa indicates a higher consistency between the two teachers' scores. At last, two teachers will discuss to make agreements on those answers that with different scorings.

ASAP-SAS Data Set: There are ten data sets in Kaggle Automatic Student Assessment Prize: Short Answer Scoring (ASAP-SAS)[3] (denoted as ASAP-SASx, $x \in \{1, 2, \cdots, 10\}$). We combine the training and test data from the leaderboard solution to a complete data set. According to the data description on that competition web site, ASAP-SAS3, 4, 7, 8, 9 are normal reading comprehension questions, which belong to "English language arts" and "English" subjects. Students should read a text in these questions and extract information from it. However, only the openness of ASAP-SAS3 and 4 fit our openness definition. Therefore, we select them for experiments, which have about 2000 samples in each question and average number of words is from 40 to 50.

In total, we construct 7 automatic open-ended reading comprehension grading tasks for each automatic grading model, 5 for Chinese and 2 for English.

[1] https://github.com/fxsjy/jieba.
[2] http://scikit-learn.org/stable/index.html.
[3] https://www.kaggle.com/c/asap-sas/data.

3.2 Baselines

We compare our model, denoted as KAGrader, with several state-of-the-are baseline automatic scoring models, including MLR [17], ZNB [26], DJDT [7,12], HSVM [11].

Implementation Details: All these baselines of MLR, HSVM, DJDT and ZNB are implemented by scikit-learn. We utilize gensim [22][4] to train CBOW model and the size of window for posterior and precedent words is 5, and the words whose frequency below 5 are ignored. The size of negative sampling sets is set to 5, and the wikipedia corpus[5] is used for knowledge adaptation.

For TCBOW, SCBOW and KAGrader, LSTM are implemented by the deep learning library keras [5]. For CRCC and ASAP-SAS data sets, the numbers of nodes in recurrent layers are {128, 128} and batch size are {512, 1024}, respectively. Finally, we all use 5-fold cross validation to evaluate our approach and baselines.

3.3 Results and Analysis

Except for four baselines, we also investigate the performance of our model without knowledge adaptation. Specifically, "TCBOW" only uses corpus from student answers (target corpus), and "SCBOW" uses Chinese or English wikipedia corpus (source corpus). While our model "KAGrader" considers both wikipedia corpus and student answers for knowledge adaptation, and all the results are reported in Tables 3 and 4. From these results, we have the following observations,

(1) HSVM is still a strong baseline which outperforms the other baselines many times, indicating that bag-of-words models can be improved by carefully selecting competitive classifier.
(2) TCBOW has worse performance than SCBOW. We conjecture that the volume of corpus has great influence on CBOW training. Therefore, it is significant for us to utilize the large source corpus to help train the target corpus vectors.
(3) KAGrader outperforms all the baselines in terms of QWKappa and average accuracy, which indicates that our model can combine the advantages of TCBOW and SCBOW.

It is worth mentioning that for neural approaches, sometimes it is limited to use the source corpus word vectors since some keywords may not appear in large source corpus, which may lead to the loss of important information in target student answers and output pure performance. Also the vector training performance is influenced by the volume of data set.

To further compare the behavior between bag-of-words models and our proposed model, we choose several student answers for analyzing.

[4] https://radimrehurek.com/gensim/index.html.
[5] https://dumps.wikimedia.org/.

Table 3. QWKappa on all data sets

	MLR	ZNB	DJDT	HSVM	TCBOW	SCBOW	KAGrader
CRCC1	0.3697	0.1970	0.2959	0.4015	0.2213	0.4431	**0.4520**
CRCC2	0.3915	0.1729	0.2556	0.4254	0.3752	0.4825	**0.4983**
CRCC3	0.7913	0.6340	0.8108	0.8680	0.7276	0.8364	**0.8694**
CRCC4	0.5142	0.2954	0.4333	0.5789	0.5693	0.5612	**0.5911**
CRCC5	0.6270	0.4465	0.6288	0.6522	0.4214	0.6754	**0.7058**
ASAP-SAS3	0.5604	0.5046	0.4558	0.5905	0.5947	0.6126	**0.6430**
ASAP-SAS4	0.5482	0.5644	0.4433	0.5695	0.5655	0.5717	**0.6103**
Average	0.5432	0.4021	0.4748	0.5837	0.4964	0.5976	**0.6230**

answer 1. “酒酿王的吆喝声” 质朴、[勤劳]，在我的生活中，妈妈每天早上叫我起床，就是朴实的声音。

answer 2. “酒酿王的吆喝声” 代表了他的 [勤奋]，值得传承。

answer 3. [扁鹊] 进谏的方式太直白，[邹忌] 是以家事比作国事来委婉进谏，以喻设理，形象的比喻了，让人欣然接受。

answer 4. [邹忌] 进谏的方式太直白，[扁鹊] 是以家事比作国事来委婉进谏，以喻设理，形象的比喻了，让人欣然接受。

For these answers, HSVM and KAGrader grade the answer 1 successfully. However, HSVM failed in answer 2 because it lack of some specific words such as 质朴、勤劳、朴实 While KAGrader score it in a right way because 勤劳 and 勤奋 are closely in CBOW word vector space and KAGrader can recognize them to output a correct score. Furthermore, the figures of 扁鹊 and 邹忌 are exchanged in sentence 4, which leads to the failure of HSVM. In a word, HSVM can not recognize the wrong sequence order, while KAGrader can take advantage of LSTM to address this issue.

Table 4. Average accuracy, precision, recall and F1 on seven data sets

	MLR	ZNB	DJDT	HSVM	TCBOW	SCBOW	KAGrader
Accuracy	0.6724	0.6645	0.6367	0.6868	0.7000	0.7256	**0.7375**
Precision	0.6628	0.6436	0.6334	0.6871	0.6821	0.7144	**0.7281**
Recall	0.6724	0.6645	0.6367	0.6868	0.7000	0.7256	**0.7375**
F1	0.6634	0.6304	0.6334	0.6844	0.6745	0.7125	**0.7255**

3.4 Parameter Sensitivity

In this section, we discuss how the number of nodes in recurrent layer and the training batch size impact on the performance of our model. To tune the hyper-parameters, we randomly selected two Chinese and all English problems.

The number of nodes are sampled from {16, 32, 64, 128, 256}, and the training batch size is sampled from {64, 128, 256, 512, 1024}. From these results in Fig. 2, we finally set the numbers of nodes and training batch size to {128, 128, 128, 128} and {512, 512, 1024, 1024} for these four data sets. Furthermore, we tune the size of negative sampling sets NEG_k in CBOW models, and the size is sampled from {1, 5, 10, 15, 20}. From the results, we finally set the same size to 5 for both Chinese and English data sets.

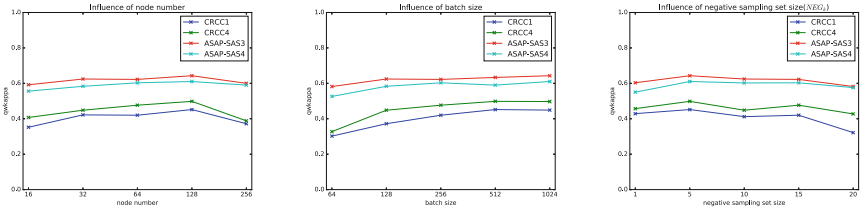

Fig. 2. Parameter sensitivity

4 Related Work

Short Answer and Reading Comprehension: NLP techniques are often used to extract various features from student answer to measure the similarity between the reference answer and student answers. Content Assessment Module (CAM) used features to measure the overlap of content on various linguistic levels [1]. The types of overlap include word unigrams, trigrams, text similarity thresholds etc. Madnani et al. in [17] used eight features based on the rubric (BLEU, coherence etc.) for summary assessment. After feature extracting process, these features are used to train various classification, regression or clustering models for grading new student answers automatically. Different machine learning models are utilized in ASAG task in [1,2,7,11,12,17,26]. Particularly, Zhang et al. [27] introduced Deep Belief Network (DBN) into short answer.

Automatic reading comprehension grading is regarded as an exception of ASAG, due to the fact that reading comprehension task need students to "understand" the reading text assuredly, not just "recall" the external knowledge. Meures et al. [18] considered that answers might show variation on different levels (lexical, morphological etc.). Horbach et al. [10] demonstrate that the use of text-based features can promote performance. Automatic reading comprehension grading was also investigated by Liu et al. [16] and Wang et al. [24].

Neural Network and Text Classification: A growing number of researchers applied neural network techniques in text classification which is a relevant topic for automatic grading. Graves et al. [8] applied LSTM into speech recognition. Tang et al. utilized GRU in sentiment classification [23]. Lai et al. used R-CNN in text classification [15]. Previous works reveal that neural network techniques perform well in natural language processing, which may have significant implications to automatic open-ended reading comprehension grading.

5 Conclusions

In this paper, we propose to combine continuous bag-of-words model (CBOW) and long-short term memory recurrent neural network (LSTM) for automatic open-ended Chinese reading comprehension grading. Our method does not rely on any reference answer due to the fact that reference answer is not always available for most open-ended reading comprehension questions. Based on CBOW and LSTM, our framework can extract semantic information automatically and effectively by considering the word orders in student response. Additionally, through knowledge adaptation, the external knowledge is transfered to present corpus. Experiments on seven data sets, including Chinese and English, 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. 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)
2. Basu, S., Jacobs, C., Vanderwende, L.: Powergrading: a clustering approach to amplify human effort for short answer grading. Trans. Assoc. Comput. Linguist. **1**, 391–402 (2013)
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**(6), 1137–1155 (2003)
4. Boer, P.T.D., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**(1), 19–67 (2005)
5. Chollet, F., et al.: Keras (2015). https://github.com/fchollet/keras
6. Dunne, R.A., Campbell, N.A.: On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In: Proceedings of the 8th Australian Conference on the Neural Networks, Melbourne, vol. 185, p. 181 (1997)
7. Dzikovska, M.O., Nielsen, R.D., Brew, C.: Towards effective tutorial feedback for explanation questions: a dataset and baselines. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 200–210. Association for Computational Linguistics (2012)
8. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649. IEEE (2013)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Horbach, A., Palmer, A., Pinkal, M.: Using the text to evaluate short answers for reading comprehension exercises. In: * SEM@ NAACL-HLT, pp. 286–295 (2013)

11. 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
12. 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)
13. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **60**(5), 493–502 (2013)
14. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. AAAI **333**, 2267–2273 (2015)
16. Lui, A.K.-F., Lee, L.-K., Lau, H.-W.: Automated grading of short literal comprehension questions. In: Lam, J., Ng, K.K., Cheung, S.K.S., Wong, T.L., Li, K.C., Wang, F.L. (eds.) ICTE 2015. CCIS, vol. 559, pp. 251–262. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-48978-9_23
17. 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)
18. Meurers, D., Ziai, R., Ott, N., Bailey, S.M.: Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. Int. J. Contin. Eng. Educ. Life Long Learn. **21**(4), 355–369 (2011)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
21. Mnih, A., Hinton, G.: Three new graphical models for statistical language modelling. In: Proceedings of the 24th International Conference on Machine Learning, pp. 641–648. ACM (2007)
22. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta, May 2010
23. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: EMNLP, pp. 1422–1432 (2015)
24. Wang, H.C., Chang, C.Y., Li, T.Y.: Assessing creative problem-solving with automated text grading. Comput. Educ. **51**(4), 1450–1466 (2008)
25. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328 (2014)
26. Zesch, T., Levy, O., Gurevych, I., Dagan, I.: UKP-BIU: similarity and entailment metrics for student response analysis, Atlanta, Georgia, USA, p. 285 (2013)
27. 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)
28. Zhila, A., Yih, W.t., Meek, C., Zweig, G., Mikolov, T.: Combining heterogeneous models for measuring relational similarity. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1000–1009 (2013)