Power Micro-Blog Text Classification Based on Domain Dictionary and LSTM-RNN

Meng-yao Shen, Jing-sheng Lei, Fei-ye Du, and Zhong-qin Bi(✉)

College of Computer Science and Technology,
Shanghai University of Electric Power, Shanghai, China
zqbi@shiep.edu.cn

Abstract. The micro-blog texts of the national grid provinces and cities will be analyzed as the main data, including the micro-blogs and corresponding comments, which will help us understand the events of power industry and people's attitudes towards these events. In this work, the data set is composed of 420,000 micro-blog texts. Firstly, the professional vocabulary of electric power is extracted, and these vocabulary are manually labeled, thus proposing a new field dictionary closely related to the power industry. Secondly, using the new power domain dictionary to classify the 2018 electric micro-blogs, and we can find that classification accuracy increased from 88.7% to 95.2%. Finally, a classification model based on LSTM (Long Short-Term Memory) and RNN (Recurrent Neural Network) is used to deal with the comments under the micro-blog. The experimental result shows that the classification of the LSTM-RNN is more accurate. The rate was 83.1%, which was significantly better than the traditional LSTM and RNN text classification models of 78.4% and 73.1%.

Keywords: Text classification · Power micro-blog · Domain dictionary · Word vector · Classification accuracy · LSTM-RNN

1 Introduction

The so-called micro-blog emotional analysis is to identify personal emotions [1, 2] from the micro-blog published by users, so as to judge the emotional tendencies of micro-blog texts [3–5], or to get the views expressed by users are "agree", "neutral" or "oppose". Aiming at the problem of feature selection in emotional analysis of micro-blog, Ning, Yang and Zhao [6] proposed the construction method of emotional dictionary based on Synonym Words Forest and micro-blog retrieval system. Cherishing [7] proposed a multi-strategy approach based on emoticons and emotional dictionaries to calculate the emotional tendency of micro-blog texts by counting the number of emoticons and emotional words. However, this approach has some limitations in dealing with micro-blogs that do not obviously contain emotional features.

The language model can be divided into word level and character level according to the prediction results of the output terminal. In the existing studies, most of the language models are at the lexical level [8–11], but a small number of studies focus on the character level. For example, Karpathy, Johnson and Fei-Fei [12] demonstrated the learning ability of LSTM by using character-level model, while Ballesteros, Dyer and

H. Gao et al. (Eds.): TridentCom 2019, LNICST 309, pp. 36–45, 2020.
https://doi.org/10.1007/978-3-030-43215-7_3

Smith [13] constructed LSTM by replacing pronouns with characters, which improved the accuracy of dependency analysis.

In building language model, Socher, Lin and Ng [14] uses RNN to parse syntax. Irsoy and Cardie [15] builds RNN into a deep structure and becomes a typical three-tier deep learning model. However, RNN has the problems of gradient explosion and disappearance [16, 17], and is not suitable for long text [18]. So later researchers put forward LSTM (Long Short-Term Memory) [19], which is a time recursive neural network, and is suitable for processing or predicting time series. Important events with relatively long intervals. Today, LSTM has been applied in many fields, such as emotional classification [20, 21], machine translation [22, 23], semantic recognition [24], intelligent question and answer [25]. In short, LSTM-based natural language processing has become the mainstream research direction.

2 Introduction of LSTM-RNN

In this section, the structure and modeling method of LSTM-RNN neural network model will be described.

2.1 LSTM Text Classification Model

LSTM is widely used in NLP, and has many mature applications in machine translation and text classification. LSTM neural network model is specially designed to deal with the problem of long-term dependence absence. It differs from traditional RNN network model in that LSTM has different cyclic unit module structure, which is different from traditional RNN network model. Some module structures are stored in four interacting layers of neural networks. Specifically, LSTM can effectively control the historical information by improving the structure of RNN, adding memory unit and three gated units, instead of removing the hidden layer state of the previous moment every time as RNN does. These improvements enhance the ability of LSTM to process long text sequences and solve the problem of gradient disappearance.

The cell model structure of LSTM is shown in Fig. 1.

- Forget Gate

In text categorization, sometimes it is necessary to label words as singular or plural. At this time, LSTM can achieve the desired effect.

$$\Gamma_f^{(t)} = \sigma\left(W_f\left[a^{(t-1)}, x^{\langle t\rangle}\right] + b_f\right) \tag{1}$$

W_f is the parameter that controls the forget gate. The value of $\Gamma_{\mathrm{f}}^{(t)}$ is [0, 1]. The forget gate vector multiplies the previous cell $\mathrm{c}^{(\mathrm{t}-1)}$. Therefore, if the value of $\Gamma_{\mathrm{f}}^{(t)}$ is 0 (or close to 0), this means that LSTM will remove the information previously stored in the cell; if the value of $\Gamma_{\mathrm{f}}^{(t)}$ is 1, it Represents retaining information in a cell.

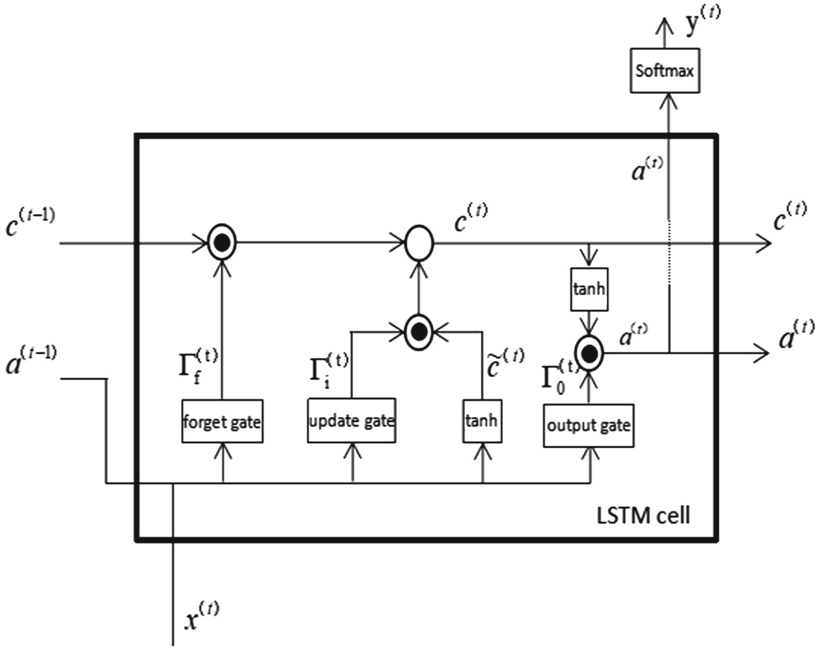

Fig. 1. Cell model structure of LSTM.

- Update Gate

If in forget gate the nature of the object is not required Singular (i.e. forgetting singular information) requires update gate to update the state of the object to be complex. The formula is as follows:

$$\Gamma_u^{(t)} = \sigma\left(W_u\left[a^{\langle t-1\rangle}, x^{\{t\}}\right] + b_u\right) \tag{2}$$

The value of $\Gamma_u^{(t)}$ is [0, 1]. At the same time, the product of $\tilde{c}^{(t)}$ and the element is used to calculate $c^{\langle t\rangle}$.

- Update Cell Status

First, a vector is used to save the state of the previous cell:

$$\tilde{c}^{\langle t\rangle} = \tanh\left(W_c\left[a^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_c\right) \tag{3}$$

The state of the new cell is:

$$c^{(t)} = \Gamma_f^{(t)} * c^{(t-1)} + \Gamma_u^{(t)} * \tilde{c}^{(t)} \tag{4}$$

- Out Gate

The value used to determine the final output:

$$\Gamma_o^{(t)} = \sigma\left(W_o\left[a^{\langle t-1)}, x^{(t)}\right] + b_o\right) \tag{5}$$

$$a^{(t)} = \Gamma_o^{(t)} * \tanh\left(c^{(t)}\right) \tag{6}$$

2.2 LSTM-RNN Model

The structure of LSTM-RNN model is shown in Fig. 2.

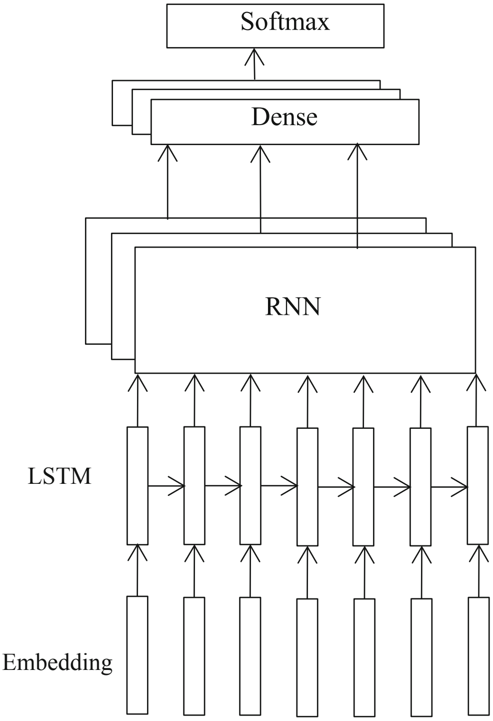

Fig. 2. Structure of LSTM-RNN model.

Firstly, the trained word vectors are input into Embedding layer, and then the output vectors of Embedding layer are input into an initial LSTM layer for semantic feature extraction. Because the original corpus is processed by Padding, the LSTM output needs to be multiplied by Mask matrix to reduce the impact of Padding. Since the RNN can process the data directly, the output of the LSTM will then be used directly as input to the RNN for further feature extraction. Finally, the output of RNN convolution layer will be aggregated to a smaller latitude, and the output will be positive tag 1, neutral tag 0 and negative tag −1, so as to obtain the required text classification results.

3 Experiment

3.1 Experimental Preparation

The experiment in this paper is carried out under Windows 10 system. The CPU used is Inter Core i5-2450M 2.5 GHz, and the memory size is 6 GB. The experimental programming language is Python 3.5, the development tool is Pycharm, and the depth learning framework used is Tensorflow 1.0.1.

3.2 Experimental Design

This section mainly based on the severity of the power accident, to analyze the micro-blog text corpus and get three types.

Data Set for Micro-Blog Text. Using selenium-based micro-blog crawlers, this paper collects all relevant texts and letters published by Sina micro-blog in 2018 from the State Grid and other provinces, include Jiangsu, Jiangxi, Henan, Zhejiang, Hunan, Shanghai, Beijing, Xiamen, Shaanxi, Chongqing. Interest, a total of 420,000 data.

Establishment of Emotional Dictionary in Electric Power Field. Domain dictionary is mainly used to store real words with clear distinction. Because of the particularity of power industry, it is difficult to process them directly by using existing dictionaries. Therefore, we construct a new emotional dictionary in power field. The main flow chart is shown in Fig. 3.

By focusing on the official micro-blog accounts of 11 different provinces and municipalities, the micro-blog published in 2018 is regarded as the corpus content, and the special vocabulary of the power industry is analyzed and annotated artificially. For example, the vocabulary of precision, order, efficiency, environmental protection, automation, power saving and operation is added into the active category, while the vocabulary of damage and operation is damaged. Words such as sudden drop, fall, blackout, electricity theft, lightning strike, emergency repair, arrears and so on were added to the negative category. At the same time, it combines NTUSD Emotional Dictionary of Taiwan University, CNKI Emotional Dictionary and Tsinghua University Chinese Commendatory and Degradation Dictionary to get an Emotional Dictionary for the field of electricity. Using the emotional dictionary in the field of power, this paper carries out emotional analysis on power-related micro-blogs. The idea is as follows: to segment each micro-blog text document, find out the emotional words, negative words and degree adverbs, and then judge whether there are negative words and degree adverbs in front of each emotional word. In each group, if there is a negative word, the emotional weight of the emotional word is multiplied by −1. If there is a degree adverb, the emotional weight of the emotional word is multiplied by the degree value of the degree adverb. Finally, the scores of all groups are added up, that is, the micro-blog. Emotional score of text. Experiments show that the accuracy rate of text categorization by using emotion dictionary in power field is 95.2%, which is significantly higher than 88.7% when using ordinary emotion dictionary for text categorization.

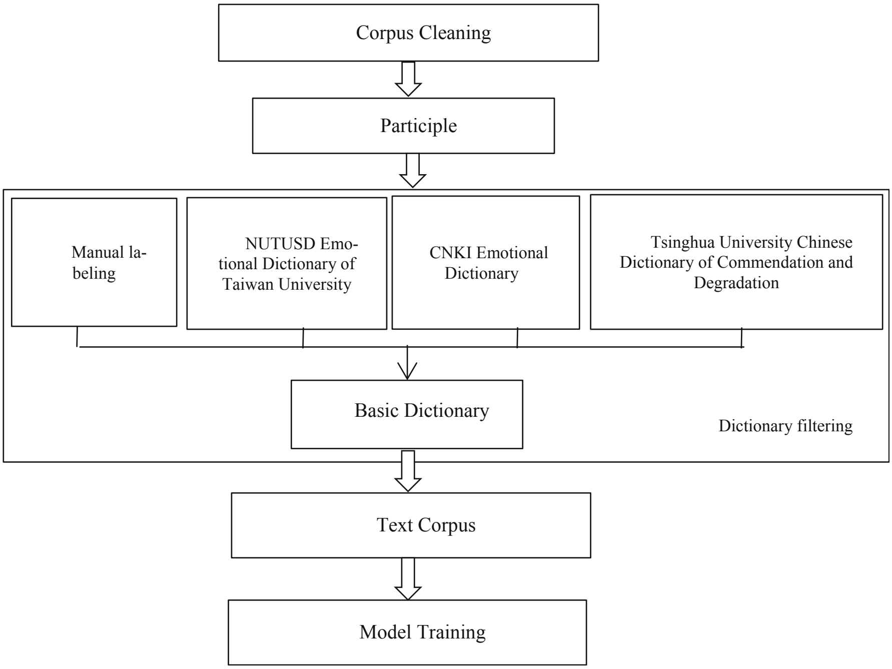
Corpus Cleaning
Participle
Manual labeling
NUTUSD Emotional Dictionary of Taiwan University
CNKI Emotional Dictionary
Tsinghua University Chinese Dictionary of Commendation and Degradation
Basic Dictionary
Dictionary filtering
Text Corpus
Model Training

Fig. 3. Processing flow of micro-blog text.

Processing Comments. Firstly, the text of power micro-blog is divided into three categories: first, there are casualties; second, there are no casualties but financial losses; third, there are no casualties or financial losses. Correspondingly, all comments on micro-blog are classified into the same three categories, thus three sets of power micro-blog comment data sets are obtained. Next, we can use LSTM-RNN model to classify the comment text of power micro-blog.

Due to the limitation of comment length on micro-blog, most of the comments on power-related micro-blog are relatively short. At the same time, there are also many special emoticons. In view of these characteristics of power micro-blog text, this paper considers using word vector to process micro-blog text.

Firstly, the trained word vectors are input into Embedding layer, and then the output vectors of Embedding layer are input into an initial LSTM layer for semantic feature extraction. Because the original corpus is processed by Padding, the LSTM output needs to be multiplied by Mask matrix to reduce the impact of Padding. Since the RNN can process the data directly, the output of the LSTM will then be used directly as input to the RNN for further feature extraction. Finally, the output of RNN convolution layer will be aggregated to a smaller latitude, and the output will be positive tag 1, neutral tag 0 and negative tag −1, so as to obtain the required text classification results.

3.3 Experimental Result

In order to make a more intuitive comparison between LSTM-RNN and LSTM and RNN, this paper uses Matplotlib, a third-party library of Python, to draw graphics.

As can be seen from Fig. 4, when LSTM model and RNN model are trained by gradient descent method, the loss value of function decreases gradually, and finally tends to stable convergence state. Compared with the original LSTM-RNN model, the initial loss value of LSTM-RNN model increases with the increase of model complexity, but the convergence rate increases significantly.

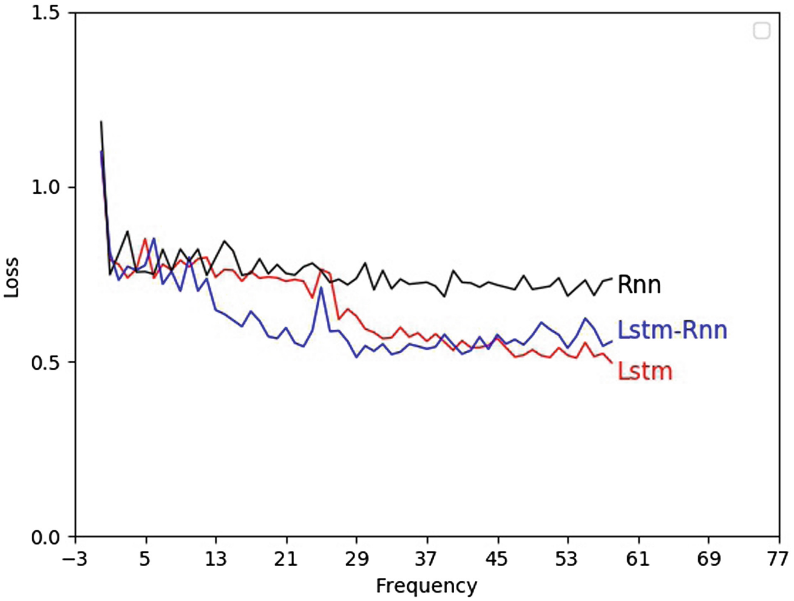

Fig. 4. Function loss value contrast graph.

As can be seen from Fig. 5, the convergence speed of the classification accuracy of LSTM-RNN micro-blog comment model based on word vector is faster than that of traditional LSTM model and RNN model, and the final classification accuracy is also higher. The classification accuracy of LSTM-RNN model is 83.1% after 460 iterations, while that of LSTM model and RNN model is 78.4% and 73.1% respectively.

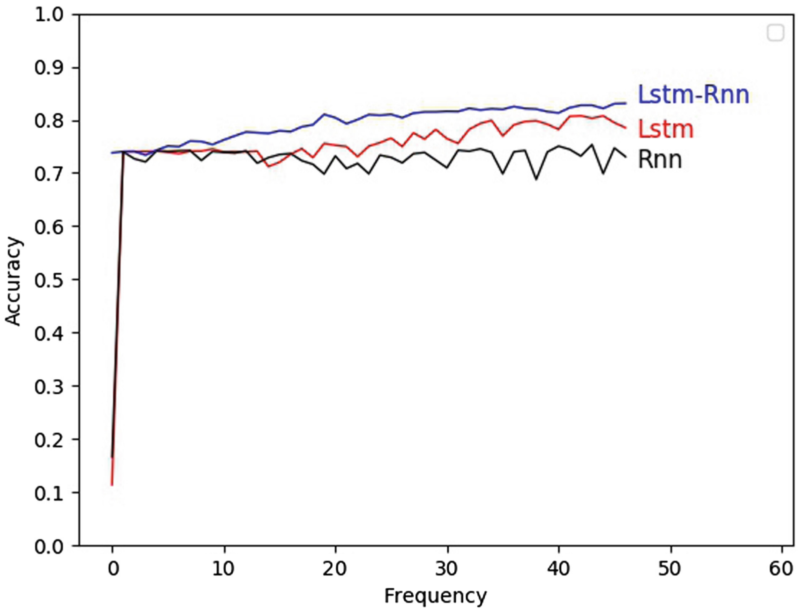

Fig. 5. Model accuracy contrast diagram.

4 Conclusion

The method based on emotional dictionary can make full use of existing emotional resources and has a good effect on emotional classification of normative texts, but this method depends largely on the quality and coverage of emotional dictionary. In this paper, we propose an emotional dictionary in the field of electric power, so as to get a higher accuracy of text categorization of electric power micro-blog. In view of the many characteristics of micro-blog comment text, this paper proposes a text classification model based on word vector LSTM-RNN. Compared with the traditional LSTM model and RNN model, the results show that this method can significantly improve the classification accuracy of micro-blog comment text, thus effectively improve the quality of micro-blog sentiment analysis.

References

1. Ding, Y., Jia, Y., Zhou, B.: Survey of data mining for Microblogs. J. Comput. Res. Dev. **51** (4), 691–706 (2014)
2. Hou, M., Teng, Y., Li, X., et al.: Research on the language characteristics and emotional analysis strategies of topic-based weibo. Lang. Charact. Appl. **2**, 135–143 (2013)
3. Song, S., Li, Q., Lu, D.: A sentiment analysis method for hot events in microblogging. Comput. Sci. **6A**, 226–228 (2012)
4. Zhang, Y., Zheng, J., Huang, G., et al.: Microblog sentiment analysis method based on a double attention model. J. Tsinghua Univ. **58**(2), 122–130 (2018)

5. Qing, F., Wang, H.C.X., Wang, X.: Microblog sentiment analysis based on linguistic context. Comput. Eng. **43**(3), 241–252 (2017)
6. Ning, H., Yang, S., Zhao, Y., et al.: Study of microblog sentiment analysis based on semantic feature. Appl. Sci. Technol. **43**(3), 70–74 (2016)
7. Xie, L., Zhou, M., Sun, M.: Hierarchical structure based hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction. J. Chin. Inf. Proc. **26**(1), 73–84 (2012)
8. Kombrink, S., Mikolov, T., Karafiát, M., et al.: Recurrent neural network based language modeling in meeting recognition. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association, Florence, Italy, pp. 2877–2880 (2011)
9. Mikolov, T., Kombrink, S., Burget, L., et al.: Extensions of recurrent neural network language model. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, pp. 5528–5531 (2011)
10. Shi, Y.Z., Zhang, W.Q., Liu, J., et al.: RNN language model with word clustering and class-based output layer. EURASIP J. Audio Speech Music Process. **2013**, 22 (2013)
11. Zhao, M., Du, H., Dong, C., et al.: Dietary health text classification based on word2vec and LSTM. Agric. Mach. **48**(10), 202–208 (2017)
12. Karpathy, A., Johnson, J., Fei-Fei, L.: Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078 (2015)
13. Ballesteros, M., Dyer, C., Smith, N.A.: Improved transition-based parsing by modeling characters instead of words with lstms. Comput. Sci. **8**(9), e74515 (2015)
14. Socher, R., Lin, C.Y., Ng, A.Y., et al.: Parsing natural scenes and natural language with recursive neural networks. In: Jonny, P., Rob, B. (eds.) International Conference on International Conference on Machine Learning, pp. 129–136. Omni Press, Haifa (2011)
15. Irsoy, O., Cardie, C.: Deep recursive neural networks for compositionality in language. Adv. Neural. Inf. Process. Syst. **3**(5), 2096–2104 (2014)
16. Hochreiter, S., Bengio, Y., Frasconi, P., et al.: Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies, pp. 237–243. Wiley/IEEE Press (2001)
17. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: Sanjoy, D.D. (ed.) Proceedings of the 30th International Conference on Machine Learning, vol. 28, pp. 1310–1318. JMLR Org, Atlanta (2013)
18. Arisoy, E., Sethy, A., Ramabhadran, B., et al.: Bidirectional recurrent neural network language models for automatic speech recognition. In: Proceedings of the 2015 Annual Conference of International Speech Communication Association, pp. 5421–5425 (2015)
19. Liang, J., Chai, Y., Yuan, H., et al.: Emotional analysis based on polarity transfer and LSTM recursive network. J. Chin. Inf. Sci. **29**(5), 152–159 (2015)
20. Liu, P., Qiu, X., Chen, X., et al.: Multrtimescale long short-term memory neural network for modelling sentences and documents. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 2326–2335 (2015)
21. Wang, X., Liu, Y., Sun, C., et al.: Predicting polarities of tweets by composing word embeddings with long short-term memory. In: Proceedings of Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural language Processing, pp. 1343–1353 (2015)
22. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the 20th NIPS, pp. 3104–3112 (2014)
23. Liu, W., Su, Y., Wu, N., et al.: Research on mongolian-chinese machine translation based on LSTM. Comput. Eng. Sci. **40**(10), 1890–1896 (2018)

24. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Proceedings of IEEE International Conference on Acoustics, vol. 38, pp. 6645–6649 (2013)
25. Wang, D., Nyberg, E.: A long short-term memory model for answer sentence selection in question answering. In: Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, pp. 707–712 (2015)