Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Evaluating the Performance of Chinese Multi-Label Grammatical Error Detection Using Deep Neural Networks

Tzu-Mi LIN[a], Chao-Yi CHEN[a], Lung-Hao LEE[a] & Yuen-Hsien TSENG[b*]

[a]*Department of Electrical Engineering, National Central University, Taiwan*[b]
Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan
*samtseng@ntnu.edu.tw

Abstract: In this paper, we describe the process of building a benchmark data set for Chinese multi-label grammatical error detection tasks, comparing the performance of 10 representative neural network models. Experimental results reveal that no matter which deep learning model is used, the performance is still limited which confirms the difficulty of the multi-label detection task. Our constructed datasets and evaluation results will be publicly released on the GitHub repository (https://github.com/NCUEE-NLPLab/CMLGED) to promote further research to facilitate technology-enhanced Chinese learning.

Keywords: Grammatical error detection, multi-label classification, deep learning

1. Introduction

Chinese as foreign language learners make various kinds of grammatical errors during their language acquisition process. Coarse-grained grammatical error types, such as missing words, redundant words, incorrect word selection, or word ordering errors, originate from target modification differences from comparing sentence surface structure with the correct sentence. An effectively automated error detection system would facilitate Chinese learning. Previous Chinese grammatical error detection approaches were based on linguistic rules (Lee et al., 2013), machine learning classifiers (Liu et al., 2016), or their hybrid methods (Lee et al., 2014). Deep learning approaches had also been applied to detect Chinese grammatical errors (Lee et al, 2017; 2020; 2021). These studies assume that a learner's written sentence has only one grammatical error, when they may in fact contain multiple errors of different types. This motivates us to explore Chinese multi-label grammatical error detection.

This study describes our process of building a multi-label grammatical error dataset from the TOCFL learner corpus (Lee et al., 2018). Recent deep-learning models have achieved excellent results in many natural language processing tasks, so we evaluate the performance of 10 representative neural networks on this dataset. The dataset and results will be made publicly available to promote the research development for technology-enhanced language learning.

2. Benchmark Data Construction and Evaluation

The benchmark data was taken from the TOCFL learner corpus (Lee et al., 2018), including grammatical error annotation of 2,837 essays written by Chinese language learners with 46 different mother-tongue languages. In each essay, we divided each paragraph into individual sentences, splitting based on the newline symbol and punctuations including an exclamation mark, question marks, or full stops. We selected sentences containing at least one of four error labels: **M**issing words (denoted as M), **R**edundant words (R), incorrect word **S**election (S), and **W**ord ordering error (W). This filtering process left us with 19,546 sentences with a total of 27,189 labels. Each sentence contains average 27.9 characters and 1.39 labels. About two-third all sentences (13,265/67.8%) were annotated with only one label, followed by two labels (5,024/25.7%), three labels (1,152/5.9%), and four labels (105/0.005%).

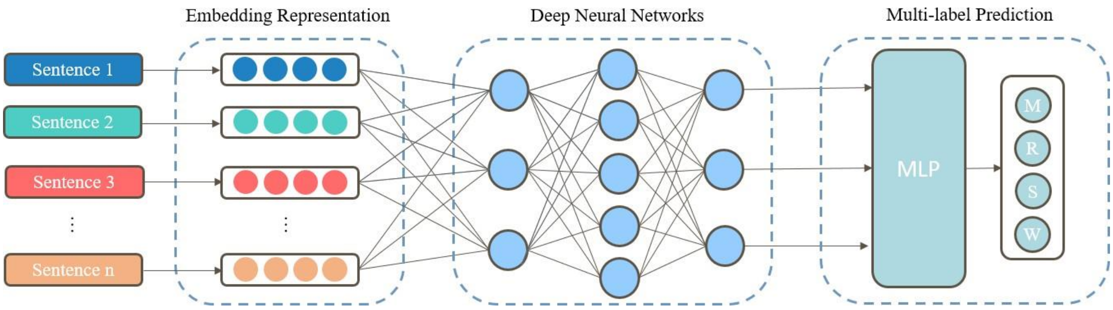

Figure 1. Performance evaluation workflow.

Figure 1 shows our experimental architecture for Chinese multi-label grammatical error detection. For word embedding representation, we dumped Chinese Wikipedia (from March 4, 2021) and automatically segmented words. Words occurring in Wikipedia at least 5 times were retained for embedding training. We pretrained embedding vectors using the GloVe technique (Pennington et al. 2014), obtaining 1,126,163 distinct word vectors, where the dimension of vectors is 300. These word vectors in sentences were then regarded as initial representations to feed into deep neural networks.

Our experimental neural networks can be further divided into three types: 1) Stacked-based: conventional neural network architectures, including traditional BiLSTM (Graves et al., 2013), CNN (Kim, 2014), and fastText (Bojanowski et al., 2017); 2) Graph-based: a class of neural networks used for processing data represented by graph data structures. We used the Graph-CNN (Defferrard et al., 2016), TextGCN (Yao et al., 2019), and HyperGAT (Ding et al., 2020); 3) Transformer-based: a type of pre-trained model that uses the encoder and self-attention mechanism. We compared the BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020). All the sentences with their labeled classes are used to train our deep neural networks to automatically learn all the corresponding parameters. To predict the error classes of a sentence during the testing phase, the sentence unseen in the training phase goes through the neural network architecture to yield a probability value corresponding to each class (the activation is a sigmoid function). The class with a probability exceeding 0.5 will be returned as the prediction results.

3. Experimental Results

Table 1. *Evaluation on Chinese Grammatical Error Detection*

Neural Network Models		**Micro F1**	**Macro F1**	**Weighted F1**	**Subset Accuracy**
Stack-based	BiLSTM	0.4827	0.2993	0.4261	0.2258
	CNN	0.5035	0.3021	0.4338	0.2313
	fastText	0.436	0.2355	0.3457	0.1844
	Ave.	*0.4741*	*0.2790*	*0.4019*	*0.2138*
Graph-based	Graph-CNN	0.4810	0.3972	0.4801	0.1260
	TextGCN	0.4242	0.3436	0.4228	0.1767
	HyperGAT	0.5078	0.2972	0.4246	0.2461
	Ave.	*0.4710*	*0.3460*	*0.4425*	*0.1829*
Transformer-based	BERT	0.5698	0.4973	0.5645	0.3374
	RoBERTa	0.5670	0.4954	0.5612	0.3287
	XLNet	0.5362	0.4495	0.5278	0.3030
	ELECTRA	**0.6147**	**0.5059**	**0.5978**	**0.3739**
	Ave.	*0.5719*	*0.4870*	*0.5628*	*0.3358*

Table 1 shows the experimental results of 5-fold cross-validation. Similar to other natural language processing findings, transformer-based models significantly outperformed graph-based or stack-based neural networks, regardless of the evaluation metrics used. Although the ELECTRA model achieved the best performance among our 10 evaluated neural networks, final results still need considerable improvement.

4. Conclusions

This study describes the process of building a benchmark data set for the Chinese multi-label grammatical error detection task, comparing the performance of 10 representative neural network models. Experimental results indicate considerable room for improvement regardless of which model is used. Our constructed datasets and evaluation results are publicly available at https://github.com/NCUEE-NLPLab/CMLGED to promote further research for technology-enhanced Chinese language learning.

In addition to developing data-intensive computing models such as neural networks, future work will focus on extracting discriminatively linguistic features for the detection task.

Acknowledgements

This study was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3, MOST 109-2410-H-003-123-MY3, MOST 111-2628-E-008-005-MY3.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov T. (2017). Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5, 135-146.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *Proceedings of ICLR'20*, https://arxiv.org/abs/2003.10555

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Proceedings of NIPS'16* (pp. 3844-3852), Barcelona, Spain.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint, https://arxiv.org/abs/1810.04805

Ding, K., Wang, J., Li, J., Li, D., & Liu, H. (2020). Be more with less: hypergraph attention networks for inductive text classification. *Proceedings of EMNLP'20* (pp. 4927-4936), Online: ACL Anthology.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *Proceedings of ICASSP'13* (pp. 6645-6649), Vancouver, Canada: IEEE Digital Library.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of EMNLP'14* (pp. 1746-1751), Dohar, Qatar: ACL Anthology.

Lee, L.-H., Chang, L.-P., Lee, K.-C., Tseng, Y.-H. & Chen, H.-H. (2013). Linguistic rules based Chinese error detection for second language learning. *Proceedings of ICCE'13* (pp. 27-29), Bail, Indonesia: APSCE.

Lee, L.-H., Hung, M.-C., Chen, C.-Y., Chen, R.-A., & Tseng, Y.-H. (2021). Chinese grammatical error detecting using adversarial ELECTRA transformers. *Proceedings of ICCE'21* (pp. 111-113), Online: APSCE.

Lee, L.-H., Lin, B.-L., Yu, L.-C., & Tseng, Y.-H. (2017). Chinese grammatical error detection using a CNN-LSTM model. *Proceedings of ICCE'17* (pp. 919-921), Christchurch, New Zealand: APSCE.

Lee, L.-H., Tseng, Y.-H., & Chang, L.-P. (2018). Building a TOCFL learner corpus for Chinese grammatical error diagnosis. *Proceedings of LREC'18* (pp. 2298-2304), Miyazaki, Japan: ACL Anthology.

Lee, L.-H., Wang, Y.-S., Lin, P.-C., Hung, C.-T., & Tseng, Y.-H. (2020). Multi-channel CNN-BiLSTM for Chinese grammatical error detection. *Proceedings of ICCE'20* (pp. 558-560), Online: APSCE.

Lee, L.-H., Yu, L.-C., Lee, K.-C., Tseng, Y.-H., Chang, L.-P., & Chen, H.-H. (2014). A sentence judgment system for grammatical error detection. *Proceedings of COLING'14* (pp. 67-70), Dublin, Ireland: ACL Anthology.

Liu, Y., Han, Y., Zhou, L., & Zan, H. (2016). Automatic grammatical error detection for Chinese based on conditional random field. *Proceedings of NLPTEA'16* (pp. 57-62), Osaka, Japan: ACL Anthology.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint, https://arxiv.org/abs/1907.11692.

Pennington, J., Socher, R., & Manning C. (2014). GloVe: global vectors for word representation. *Proceedings of EMNLP'14* (pp. 1532-1543), Doha, Qatar: ACL Anthology.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: generalized autoregressive pretraining for language understanding. arXiv preprint, https://arxiv.org/abs/1906.08237

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI'19* (pp. 7370-7377), Honolulu, Hawaii, USA.