Automatic Chinese Spelling Checking and Correction based on Character-based Pre-trained Contextual Representations

Haihua Xie[1], Aolin Li[1], Yabo Li[1], Jing Cheng[1], Zhiyou Chen[1], Xiaoqing Lyu[2], and Zhi Tang[2]

[1] State Key Laboratory of Digital Publishing Technology (Peking University Founder Group Co. LTD.), Beijing, China
{xiehh, lial, liyabo, cheng.jing, czy}@founder.com.cn
[2] Institute of Computer Science and Technology of Peking University, Peking University, Beijing, China
{lvxiaoqing, tangzhi}@pku.edu.cn

Abstract. Automatic Chinese spelling checking and correction (CSC) is currently a challenging task especially when the sentence is complex in semantics and expressions. Meanwhile, a CSC model normally requires a huge amount of training corpus which is usually unavailable. To capture the semantic information of sentences, this paper proposes an approach (named as DPL-Corr) based on character-based pre-trained contextual representations, which helps to significantly improve the performance of CSC. In DPL-Corr, the module of spelling checking is a sequence-labeling model enhanced by deep contextual semantics analysis, and the module of spelling correction is a masked language model integrated with multilayer filtering to obtain the final corrections. Based on experiments on SIGHAN 2015 dataset, DPL-Corr achieves a significantly better performance of CSC than conventional models.

Keywords: Chinese Spelling Checking and Correction · Character-based Model · Pre-trained Contextual Representations · Language Model.

1 Introduction

The target of Chinese spelling checking and correction (CSC) is to detect and correct the misuse of characters or words in Chinese sentences. Similar to spelling checking in English, there are two types of spelling errors in Chinese, real-word spelling errors and non-word spelling errors. Real-word spelling errors refer to the misuse of two real words (i.e., words in the lexicon), such as 'word' and 'world', '中止' and '终止'. Non-word spelling errors refer to misspelling a real word to be one not in the lexicon, e.g., 'word' is misspelled to be 'wrod', '偶尔' is misspelled to be '偶而'.

CSC can be normally divided into two phases. The first phase is spelling errors detection, which is to detect and locate the spelling errors in a sentence.

The second phase is spelling errors correction, which is to give correction suggestions for the detected errors. The candidate corrections are normally selected from a large-size confusion set of similar-pronunciation and similar-shape characters. The final correction result is determined based on the confidence of each candidate correction which is evaluated based on language modeling [1].

Due to the variability and complexity of Chinese semantic expressions, automatic CSC is currently a challenging task. Firstly, spelling errors often occur among single-character words, of which the use is flexible and complicated. Secondly, non-words in a sentence are not necessarily errors. For instance, '偶而' is usually a non-word misspelled for '偶尔', whereas it is correct in the sentence '人生应该随偶而安'. On the other hand, non-words produced by spelling errors usually cannot be correctly segmented by word segmentation, thus word-based matching cannot be applied to detect them. Furthermore, the target of CSC is to detect misused characters instead of words. All in all, word-based models that take words as the basic processing objects are not suitable for the CSC task.

In the current research community of CSC, correction of non-word spelling errors is the main problem, whereas the SOTA performance is about 70% [2]. The accuracy of real-word spelling correction is even lower, because such errors are usually occur among words with similar meanings. Deep semantic analysis of the context is required for correcting real-word spelling errors. Besides the low accuracy of correction, CSC also encounters a problem of high false positive rates. Words that appear rarely in the training data or the lexicon, such as technical terms and people names, are often mistakenly considered spelling errors.

This paper proposes a model for the CSC task, named as DPL-Corr, which aims to handle the problems mentioned above. Instead of building a set of spelling error instances that are infinite, DPL-Corr is mainly designed based on pre-trained contextual representations, so that the knowledge of grammars and concepts learned from a large-size corpus can be readily utilized to reduce the size of training data. Due to the inherent shortcomings of word-based CSC, DPL-Corr is designed based on character-based pre-trained contextual representations. A masked language model is designed to give candidate corrections of detected spelling errors, and select the final correction. According to our experiments conducted on the dataset of SIGHAN 2015, DPL-Corr achieves a F1 score of 0.6528 which beats the previous SOTA models on this task by a wide margin.

2 Background & Related Works

In the existing approaches for the CSC task, language modeling, word segmentation, confusion sets, and spelling errors set are widely used resources and techniques. The procedure of CSC normally consists of two steps, spelling errors detection and spelling errors correction. The target of spelling errors detection is to detect and locate the spelling errors in a sentence, and spelling errors correction is to give correction suggestions for those errors.

Spelling errors detection is the primary part in CSC. Heish et al. [3] recognizes spelling errors based on language model verification and the set of spelling

errors generated from a confusion set, which is a dictionary containing frequently used characters as keys and corresponding characters of similar shapes or pronunciations as values. Zhao et al. [4] constructs a directed acyclic graph for each sentence and adopts the single-source shortest path algorithm to recognize common spelling errors. Wang et al. [5] constructs a mass of corpus with automatically generated spelling errors, and then implements a supervised sequence tagging model for spelling error detection.

For spelling error correction, Heish et al. [3] employs a confusion set to replace all suspect spelling errors and applies a n-gram model to choose the optimal correction. Such mechanism is efficient, whereas the false negative rate is high. Yang et al. [6] applies the ePMI matrix to count the co-occurrences of words and characters, and selects the candidate corrections based on the ePMI matrix. Li et al. [7] translates sentences with spelling errors into grammatically correct sentences, and selects the corrections from the translation results. Wang et al. [8] utilizes a seq2seq model to copy correct characters through a pointer network or generate characters from the confusion set to correct spelling errors.

Most of the above methods are low in efficiency or performance, because they normally require construction of a large-size confusion set and a spelling errors set, and training of a classifier based on a large-scale corpus. Besides, the errors in word segmentation can strongly affect the performance of the CSC task.

3 Workflow of DPL-Corr for CSC

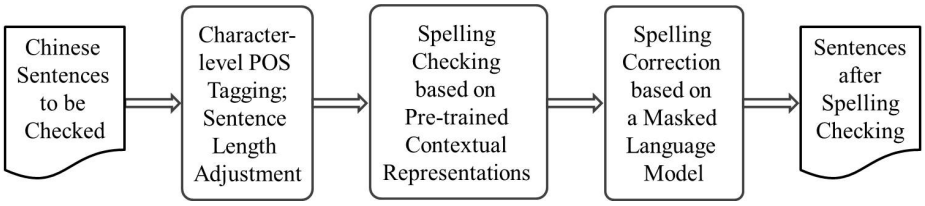

Fig. 1. The workflow of Chinese spelling checking and correction in DPL-Corr.

As shown in Fig. 1, the input of DPL-Corr is a Chinese passage consists of several sentences, and the output is the passage after spelling checking and correction. Below is a brief introduction of the middle three steps in the workflow.

1. Character-level POS Tagging and Sentence Length Adjustment
   During training and inference, the following preprocessing operations are performed on the raw input Chinese sentences.
   (a) Character-level POS Tagging
   POS tags are helpful for detecting certain spelling errors. For instance, if the word after '的' in a sentence is a verb, '的' is very likely to be misused (because '的' is usually followed by a noun). To give a POS tag for each character, character-level POS tagging is performed as follows.

i. Segmenting words for the input sentences and assigning a POS tag for each word.
ii. Based on the POS tags of words, each character is given a character-level POS tag using the labeling schema 'BIES'. For instance, the POS tag of '乌鲁木齐' is 'ns', and the POS tag of each character is: 'ns-B' for '乌', 'ns-I' for '鲁', 'ns-I' for '木', 'ns-E' for '齐'.

(b) Sentence length adjustment
The length of sentences may affect the performance of CSC. Lengthy sentences increase difficulty of computing contextual embeddings, and they might contain multiple spelling errors which break the semantic integrity of the sentence. On the other hand, sentences being too short are disadvantageous in spelling checking because language models are less keen on capturing contextual meaning of short texts. Thus, DPL-Corr has a limit of 15 to 30 characters for the input sentences. Lengthy sentences will be truncated to the first 30 characters, and short sentences will be concatenated with sentences before or after the sentence.

2. Spelling checking based on pre-trained contextual representations
The module of spelling checking in DPL-Corr is designed based on character-based pre-trained contextual representations. Character-based models, such as BERT and XLNet, take characters as the basic processing objects. Because the pre-trained contextual representations are normally trained based on a large-scale corpus, the module of spelling checking in DPL-Corr requires only a small amount of training data to fine-tune the model.
3. Spelling correction based on a masked language model
The module of spelling correction is designed based on a masked language model to give the candidate corrections. In the input of the correction module, the detected errors are masked, and the model gives suggested characters to fill in each vacancy (i.e., a masked character) in the sentence. To reduce false positives, DPL-Corr adopts several means including confidence filtering, ranking filtering, and confusion set filtering to eliminate the inappropriate corrections and select the final optimal result.

4 Detailed Steps of DPL-Corr for CSC

4.1 Pre-trained Context Representations based Spelling Checking

The module of spelling checking in DPL-Corr is illustrated in Fig. 2. The input to the module is a sequence of Chinese characters and their POS tags. Let x denote a character, c_pos(x) denote its character-level POS tag, PCRMod represent a pre-trained contextual representations model, and POSMat represent a POS encoding matrix. The output of encoding x by PCRMod is shown below.

$$r(x) = PCRMod(x) \tag{1}$$

r(x) is the contextual embedding of x which is dynamically computed given both past and future textual information of the entire input sequence, and its dimension is 1*768.

c_pos(x) is the POS tag of x represented by a one-hot vector, with a dimension of 1*144. 144 is the total number of character-level POS tags, because there are 36 POS tags and 4 character-level labels ('B', 'I', 'E', 'S'). After being encoded by POSMat, the POS of x is projected into a lower dimensional space.

$$e\_pos(x) = POSMat(c\_pos(x)) \tag{2}$$

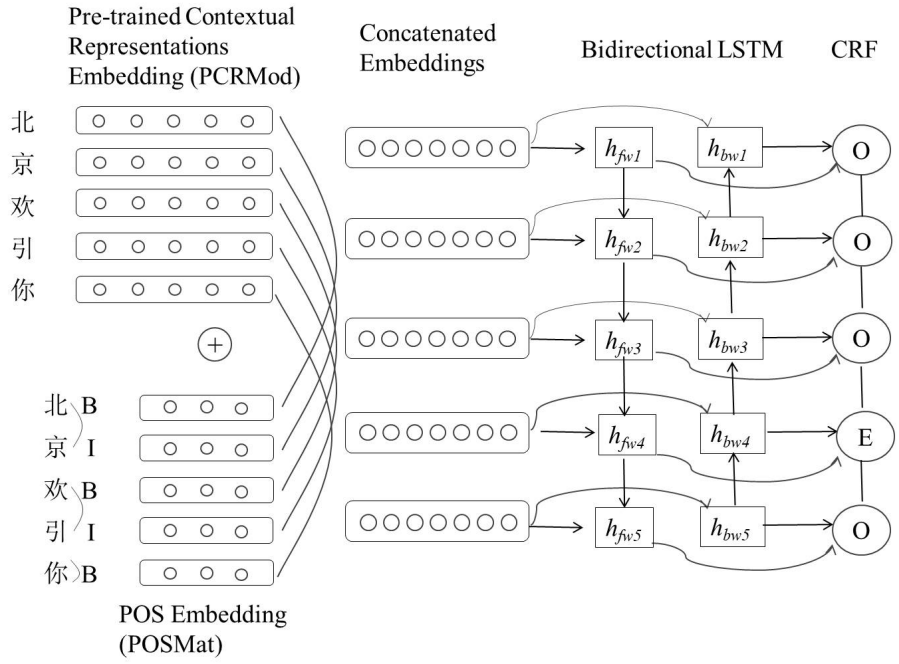

Fig. 2. The framework of Chinese spelling checking in DPL-Corr.

POSMat is a matrix of dimension 144*72. POSMat is randomly initialized and updated during training. e_pos(x) is a vector of dimension 1*72. The character embedding and the POS embedding are concatenated and fed into the Bi-LSTM network.

$$w(x) = catenation(r(x), e\_pos(x)) \tag{3}$$

w(x) is a vector of dimension 1*840, which is input to the Bi-LSTM network that produces u(x) as illustrated in formula (4).

$$u(x) = Bi\_LSTM(w(x)) \tag{4}$$

u(x) is a vector of dimension 1*256, which is input to the CRF layer which outputs the most likely label for x. Label 'E' indicates that x is a spelling error

and label 'O' otherwise. Let X^i=$\{x_1^i,...,x_K^i\}$ denote the ith input character sequence to DPL-Corr, and K is the length of X^i. U^i=$\{u_1^i,...,u_K^i\}$ is the input to the CRF layer and u_k^i=u(x). L^i=$\{l_1^i,...,l_K^i\}$ is the labeling sequence, and l_k^i is the label of x_k^i. The confidence of labeling U^i with L^i is shown in formula (5).

$$P(L^i, U^i) = \sum_{k=1}^{K} (H_{(l_{k-1}^i, l_k^i)} + \varphi(l_k^i, u_k^i)) \tag{5}$$

H is the probability transition matrix of labels. H is 2*2 matrix because there are only two labels ('E' and 'O'). $H_{(l_{k-1}^i, l_k^i)}$ is the transition probability from label l_{k-1}^i to label l_k^i. Additionally, $\varphi(l_k^i, u_k^i)$ is the score given to u_k^i being labeled as l_k^i. φ is a 2^*V matrix where V is the size of the vocabulary and 2 is the size of the label set. Both H and φ are randomly initialized and updated during the training process.

The characters labeled as 'E' are supposed to be suspect spelling errors, and they will be masked and given correction suggestions in the following step.

4.2 Spelling Correction based on Masked Language Modeling

DPL-Corr adopts a masked language modeling (MLM) approach for spelling correction. The correction module is shown in Fig. 3.

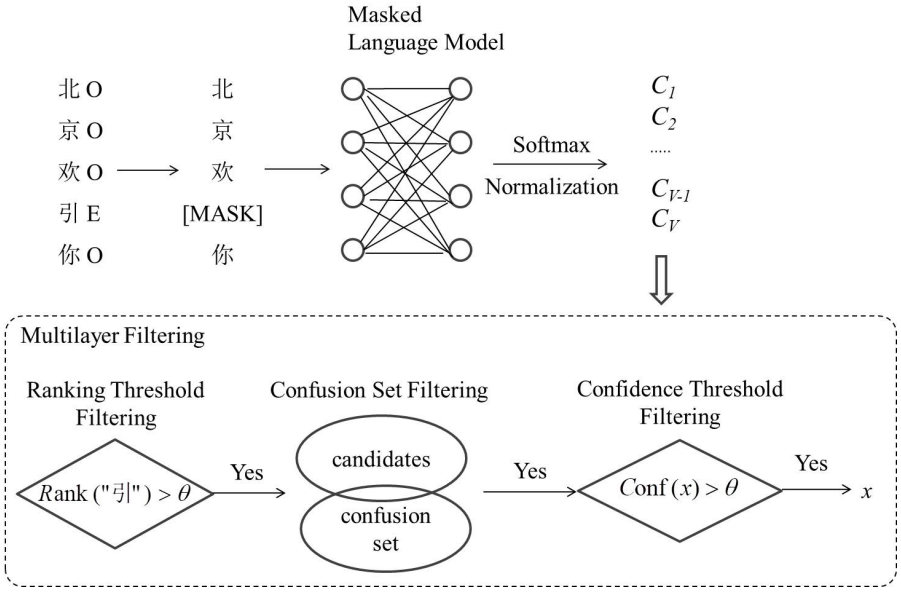

Fig. 3. The framework of Chinese spelling correction in DPL-Corr.

As introduced in the previous section, characters labeled as 'E' are replaced with a special token '[MASK]'. Then the partially masked sequence of characters is fed into the MLM layer to produce a $1 * V$ vector $\{C_1, C_2, ..., C_{V-1}, C_V\}$ at each position with token '[MASK]'. C_i is the confidence value of fitting the ith character in the vocabulary to the targeted position. To enhance the accuracy of selecting the correction result, the following three steps (called as multilayer filtering) are proposed.

1. Ranking threshold filtering
   If the masked suspect character (e.g., '긍' in Fig. 3) is ranked within, for instance, top 100 candidate corrections given by the MLM layer, the suspect is very likely to be a correct use in the sentence. As a result, the suspect will be removed from the list of spelling errors.
2. Confusion set filtering
   Most spelling errors are similar to the correct characters either morphologically or phonetically. Based on a publicly available Chinese confusion set, the candidates that are dissimilar in shape or pronunciation to the suspect are filtered out. Such confusion set filtering is performed to candidates in descending order of the confidence values until a qualified character appears.
3. Confidence threshold filtering
   After the above steps, the candidate with the highest confidence value will be checked to see if its confidence is higher than a certain threshold so it is a credible choice as a correction. Otherwise the suspect at the current position will not be treated as an error.

There is no parameter to be updated in the correction module during training. The MLM model is pre-trained and directly used in this module. The values of the confidence threshold and ranking threshold are manually selected and adjusted according to the correction performance.

5 Experiments and Analysis

5.1 Experimental Datasets

SIGHAN 2015 dataset [2] is used in our experiments. The dataset contains spelling errors made by people learning Chinese and annotations made by Chinese native speakers. Real-word spelling errors and non-word spelling errors are both contained in the dataset. The size of our experimental data is shown below.

- Training set: 970 passages, 3143 spelling errors.
- Test set: 1100 passages, half of which contain at least one spelling errors, and the other half contain no errors.

SIGHAN competitions allow participants to use extra textual and computing resources. Most participants in the SIGHAN 2015 bake-off used datasets from previous SIGHAN competitions for model training. To guarantee fairness, our experiment incorporated both SIGHAN 2013 and 2014 datasets into our training set. The performance of a CSC model is evaluated from the following two aspects.

– Error detection: whether a spelling error is correctly detected.
– Error correction: whether a spelling error is correctly detected and corrected.

Accuracy (Acc.), precision (Pre.), false positive rate (FPR) and recall (Rec.) are used as the evaluation metrics. Besides adding SIGHAN datasets of previous years to our training corpus, we conducted data augmentation targeting at certain types of spelling errors, mostly misuse of single-character words. For instance, we obtained some amount of published articles and randomly replaced the correct character with its erroneous counterparts and vice versa, such as '地' and '的', '在' and '再', etc.

5.2 Performance Analysis

Table 1. Performance of various models for Chinese spelling checking and correction.

Model	FPR	Error Detection				Error Correction			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
CAS-Run1	**0.1164**	0.6891	**0.8095**	0.4945	0.614	0.68	**0.8037**	0.4764	0.5982
CAS-Run2	0.1309	0.7009	0.8027	0.5327	0.6404	0.6918	0.7972	0.5145	0.6254
DPL-Corr	0.1818	**0.7091**	0.7674	**0.6**	**0.6735**	**0.6955**	0.759	**0.5727**	**0.6528**
w/o PCR	0.2182	0.6636	0.7143	0.5455	0.6186	0.6482	0.7022	0.5145	0.5939
w/o POS	0.2091	0.6727	0.7262	0.5545	0.6289	0.6573	0.7146	0.5236	0.6044
w/o filters[1]	0.1818	0.7091	0.7674	0.6	0.6735	0.6645	0.7375	0.5109	0.6037

[1] Because the multilayer filters are performed during error correction, the performance of error detection of the last model is same to that of the final model.

Table 1 shows the performance of DPL-Corr and its comparisons with SOTA models of SIGHAN 2015 bake-off. DPL-Corr adopts BERT as its pre-trained contextual representations with ranking threshold set to 150 and confidence threshold set to 0.5. In Table 1, CAS-Run1 and CAS-Run2 are the two models with the best overall performances in SIGHAN 2015 bake-off. Additionally, in order to thoroughly analyze the results, we evaluated DPL-Corr under various configurations and thresholds as detailed in Table 1.

We conducted in-depth analysis on the experimental results. The performance of error detection and error correction for various types of spelling errors is calculated. Regarding a certain type of errors, we first counted the number of passages that contain such type of errors, and then we applied the metrics introduced in section 5.1 to evaluate the performance of CSC on these passages. It is worth noticing that some passages contain multiple types of errors, therefore repeated calculations were inevitable during evaluation on different types of errors. The results are shown in Table 2.

Table 2. Performance of DPL-Corr for different spelling error types.

Error Type	FPR	Error Detection				Error Correction			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
Non-word spelling errors	0.1594	0.7789	0.8182	0.7171	0.7643	0.7669	0.8131	0.6932	0.7484
Real-word spelling errors	0.2007	0.6505	0.7143	0.5017	0.5894	0.6355	0.7015	0.4716	0.564
Misuse of single-char words[1]	0.1741	0.699	0.7667	0.5721	0.6553	0.6617	0.7407	0.4975	0.5952
Misuse of '的地得'	0.1485	0.7723	0.8235	0.6931	0.7527	0.7525	0.8148	0.6535	0.7253

[1] Misuse of single-character words is a type of real-word spelling errors, and misuse of '的地得' is a type of misuse of single-character words.

By observing Table 2, DPL-Corr is better at correcting non-word spelling errors than real-word spelling errors. Furthermore, data augmentation is verified to be effective because the correction performance of misuse of '的地得', in which data augmentation is performed, is superior to that of misuse of single-character words. In summary, the experimental results indicate that: (1) correction of real-word spelling errors remains challenging; (2) the quality of the training data has a significant impact on spelling errors correction.

6 Conclusions

The framework proposed in this paper for the CSC task is able to utilize the semantic and grammatical knowledge learned from a large-scale corpus, which helps to improve the performance of CSC. Meanwhile, because of the application of character-based models, the impact of incorrect word segmentation on CSC can be avoided. Besides, a novel structure of error correction is designed based on a masked language model, which utilizes the contextual information to give correction suggestions. According to our experiments conducted on the dataset of SIGHAN 2015, the proposed model achieves a CSC performance of F1 value 0.6528, which is better than the previous SOTA model.

Our model performs well in correcting non-word spelling errors and misuse of single-character words. However, our model has an unsatisfactory performance in corrections of real-word spelling errors, which are also the main difficulty of CSC. Besides, the rates of false negatives and false positives are still high, especially for those sentences with many technical terms in specific domains.

Several suggestions to improve our model are presented below.

1. For real-word spelling errors, especially the misuse of words with similar meanings, deep semantic analysis of the sentence is necessary for determining which word is more appropriate. A model with the ability of long-distance context representations is helpful for providing such deep semantic analysis.

2. Selection of candidate corrections based on a confusion set is inefficient, because characters with similar shapes or pronunciations contained in the confusion set are artificially collected and limited. An algorithm of evaluating the similarity of the shapes and pronunciations of two characters can be helpful for improving the effect of selecting candidate corrections.
3. Spelling errors are varied. For each specific type of spelling errors, using a corresponding specific set of training set and even designing a single model for it can significantly enhance the performance of CSC.
4. There are fixed grammatical rules for the use of certain words. Linguistic knowledge is helpful for determining if such words are used correctly. Thus, linguistic knowledge based automatic CSC is a promising direction.

7 Acknowledgement

This work is supported by the projects of National Natural Science Foundation of China (No. 61472014, No. 61573028 and No. 61432020), the Natural Science Foundation of Beijing (No. 4142023) and the Beijing Nova Program (XX2015B010). We also thank all the anonymous reviewers for their valuable comments.

References

1. Chen, K.-Y., Lee, H.-S., Lee, C.-H., et al.: A study of language modeling for Chinese spelling check. In: Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, pp. 79-83. (2013)
2. Tseng, Y.-H., Lee, L.-H., Chang, L.-P., et al.: Introduction to SIGHAN 2015 Bakeoff for Chinese Spelling Check. In: Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, pp. 32-37. (2015)
3. Hsieh, Y.-M., Bai, M.-H., Chen, K.-J.: Introduction to CKIP Chinese spelling check system for SIGHAN Bakeoff 2013 evaluation. In: Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, pp. 59-63. (2013)
4. Zhao, H., Cai, D., Xin, Y., et al.: A Hybrid Model for Chinese Spelling Check. ACM Transactions on Asian and Low-Resource Language Information Processing **16**(3), 1-22 (2017)
5. Wang, D.-M., Yan, S., Li, J., et al.: A Hybrid Approach to Automatic Corpus Generation for Chinese Spelling Check. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2517-2527. (2018)
6. Yang, Y., Xie, P.-J., Tao, J., et al.: Embedding Grammatical Features into LSTMs for Chinese Grammatical Error Diagnosis Task. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, Shared Tasks, pp. 41-46. (2017)
7. Li, C.-W., Chen, J.-J., Chang, J.-S.: Chinese spelling check based on neural machine translation. In: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. (2018)
8. Wang D.-M., Tay, Y., Zhong L.: Confusionset-guided Pointer Networks for Chinese Spelling Check. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, pp. 5780-5785. (2019)