7th International Conference on Information, System and Convergence Applications
12 – 15 Feb. 2020, Ho Chi Minh, Vietnam

Self - Supervised Korean Spelling Correction via Denoising Transformer

Chanjun Park, Sungjin Park, Heuiseok Lim
Korea University Dept. Computer Science, Seoul , Korea
bcj1210@naver.com, genom1324@gmail.com, limhseok@korea.ac.kr

Abstract—**Spelling correction refers to the task of converting misspelled words in each sequence into correct words. Thus, most spelling correction models require parallel corpora, including misspelled sequences as inputs and ground truth as outputs. To alleviate this limitation, we propose two noise generation methods and Transformer based denoising architecture for the semi supervised Korean spelling correction task. Our model receives 78.53 points of GLEU on the task, which shows that it currently outperforms two representative systems for Korean spelling correction.**

Keywords-component: Korea Spelling Correction, Transformer, Denoising Transformer, Self-Supervised Learning

I. INTRODUCTION

Spelling correction can be applied to various fields such as a post-processing module for speech recognition and a pre-processing module for machine translation in real-time interpretation systems. Currently, in Korea, Pusan National University and Naver have successfully operated spelling corrector services. These services consist of large rule-based systems. The advantage of the rule-based system is that it fixes incorrect parts without altering the structure of input sentences; however, it has the disadvantages of not correcting sections that deviate from the rules and being difficult to construct large-capacity rules.

In this paper, we look at the spelling correction system through the perspective of machine translation. Machine translation refers to a system that translates a source language into a target language. When applied to a spelling correction system, the source has an error in the spelling of a sentence, whereas, a target language can be viewed as the correct sentence. In this paper, we utilize a monolingual corpus to construct a Korean spelling parallel corpus and create a Transformer[1] - based Korean spelling corrector system that resulted in the highest GLEU[2] score of 78.53 points, which is superior to that of the existing spelling correction system.

Our main contributions are as follows:

- We present a semi supervised approach to conduct the spelling correction task without employing any parallel corpora. Instead, random character replacement (RCR) and predefined error list driven noising (ELN) are incorporated into input sequences.
- We apply a novel translation model, Transformer, to the Korean spelling correction task. Misspelled tokens can be more attended in the self-attention layers of the Transformer.
- Our model obtains new state-of-the-art results on the spelling correction task with scoring 78.53 for GLEU (17.69 points improvement) and 71.13 for F1 (18.70 points improvement).

II. RELATED WORK

There has been active research for Korean spelling correction at Pusan National University (PNU) and commercialized versions of Korean spelling correction has been implemented in Naver and PNU. In the past, there have been various studies on the methods of the spelling correction system, such as the rule-based spelling correction system[3,4], the statistical based spelling correction method[5,6], spelling correction using machine learning and recently, deep learning-based correction systems[7,8].

In addition, one disadvantage with using machine learning is that there is a potentially incorrect assumption that the surrounding context of the detected word is correct. When looking at the spelling correction system from the perspective of machine translation, it is advantageous to fix various aspects of spelling mistakes without constructing rules only when a high-quality parallel corpus is built well.

However, it is not easy to construct a parallel corpus, and building a high-quality parallel corpus presents an even more difficult challenge. In this paper, we propose constructing a parallel corpus with only a monolingual corpus.

We propose constructing this parallel corpus using error lists, which are word pairings of one incorrect word and its incorrect counterpart.

III. Model Description

We applied Transformer, a novel generative deep learning model, for the first time in the Korean spelling correction task.

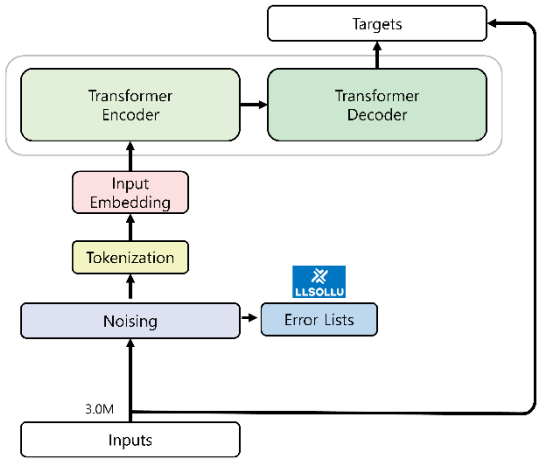

Figure 1. Overview of our system. A parallel corpus is created using an error list with a Mono corpus.

A. Error Lists

An error list is a pair of one incorrect word and its correct counterpart, such as a "Hollo" and "Hello". In the case of error lists, they can be automatically acquired using any existing real-time interpreter services. These error lists are sourced from an interpreter secretary mobile application system called ezTalky, which is a commercialized service by LLsoLLu.[1] These error lists are a reliable source of data because they are sourced from a currently active service. A total of 45,711 error lists were finally constructed.

B. Denoising Transformer

The training is based on the Transformer model introduced from "Attention is All You Need". [1] Transformer is a sequence-to-sequence model that only uses the attention technique without using convolution and recurrence. The system implementation uses OpenNMT Pytorch. [9] We construct the Pretrained Embedding vector through FastText and use it for learning. In the FastText model, because the unregistered word is viewed as a form of the composed n-gram of the word, a similar word can be estimated using the partial n-gram of which FastText makes it robust against errors. [10]

IV. Experiments

In this paper, we constructed the parallel corpus using the error lists that we proposed and applied it to the Transformer model and compared our results with other existing, well known Korean spelling correction systems. The proposed system outperforms traditional commercialization systems and demonstrated state-of-the-art level performance. Because it was approached from a machine translation perspective, GLEU scoring was used as the scoring metrics and under these metrics, the proposed model performed better than previous commercialized spelling correction systems.

A. Dataset

The data is crawled from newspaper articles. The data and vocab size used for learning are shown in the table below. The following numbers are also the result of parallel Corpus filtering:

TABLE I. Data Size

Dataset	Size
Training	3.0M
Validation	5,000

B. Experimental Setting

The learning hyperparameters of Transformer are as follows:

TABLE II. Hyper Parameter:

Hyper-parameter	Setting
Source Vocabulary	32,004
Target Vocabulary	32,002
Batch Size	4,096
Word Vector Size	512
Attention Head	8
Transformer FF	2,048
Dropout	0.1
Optimizer	Adam
Decay Method	Noam

There were four GPUs used and trained for 80,000 steps.

C. Metrics

GLEU [2] is a performance evaluation index. GLEU is similar to BLEU but it considers source information while BLEU doesn't. It is a performance evaluation index specialized in Grammar Error Correction system.

$$\mathrm{GLEU}\,(C,R,S) = \mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n \log p'_n\right) \quad (1)$$

C is correction Sentence, S is source S and R is reference. GLEU(C,R,S) is shown in Equation 1. In our experiments, we used N = 4, w_n = 1/N, and the same brevity penalty as BLEU.

[1] http://llsollu.com/

D. *Delete symbol in source sentence*

When taking the input data, we will delete all symbols from the source sentences. Through this data variation, we will see the effect of putting symbols "?" or "," into the sentences, according to the context.

V. Experimental Results

The experiments were carried out by extracting the test set from the training data and then randomly replacing or deleting alphabetical units from the test set. The reason for this noise generation is that extracting test sets without any modification to the data wouldn't fairly demonstrate the performance of the model. We then compare the performance with the spelling correction system that is currently commercially available in Korea.

TABLE III. Evaluation Table:

Model	GLEU
Commercial01	58.19
Commercial02	60.84
Ours	76.28
Ours (+ FastText)	**78.53**

The precision, recall, and f-1 scores are shown below:

TABLE IV. Evaluation Table:

Model	Precision	Recall	F1-score
Commercial01	59.11	29.64	39.49
Commercial02	51.75	53.12	52.43
Ours(+FastText)	**82.11**	**62.74**	**71.13**

Precision, recall, and f-1 scores all demonstrated better results when compared to existing commercialization systems.

There are two additional effects on this system:

1. Sentence separation and spacing
2. Context-based symbol insertion

1 is as follows:

TABLE V. Example of Sentence Seperation

Input	죄송합니다 모든 좌석이 매진됐습니다 (Sorry all seats are sold out)
Output (Translation)	죄송합니다. 모든 좌석이 매진됐습니다. (Sorry. All seats are sold out.)

Example of Sentence separation and Spacing effect. It has an automatic sentence separation effect and automatic spacing effect that is useful for Speech recognition systems

2 is as follows:

TABLE VI. Example of Symbol Attachment Effect According to Context.

Input (Translation)	여기 가까운 식당이 어디있습니까 (Whereis the nearest restaurant here)
Output (Translation)	여기 가까운 식당이 어디 있습니까? (Where is the nearest restaurant here?)

1 and 2 can be combined to be used as a post-processing module for speech recognition. STT (Speech to Text) results usually do not include symbols and the spacing is sometimes incorrect. Furthermore, there are some instances when the results of the STT process produce an unnatural sentence or a sentence that the user feels does not match the flow of the input sentence. In this case, the proposed spelling correction system can be used to solve the problem mentioned above.

VI. Conclusion

This paper is the first attempt to apply the Transformer model to a Korean spelling correction system. It also demonstrated performance exceeding that of existing commercialized systems. In addition, we will be studying how to strengthen the error lists later.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & communications Technology Promotion) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2017M3C4A7068189).

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[2] Napoles, Courtney, et al. Ground truth for grammatical error correction metrics. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Vol. 2. pp. 588-593. 2015.

[3] Kwon, Hyuk-Chul & Kang, Mi-young & Choi, Sung-Ja. (2004). Stochastic Korean Word-Spacing with Smoothing Using Korean Spelling Checker. Int. J. Comput. Proc. Oriental Lang.. 17. 239-252. 10.1142/S0219427904001103

[4] Kim, Minho & Choi, Sung-Ki & Kwon, Hyuk-Chul. (2014). Context-Sensitive Spelling Error Correction Using Inter-Word Semantic Relation Analysis. ICISA 2014 - 2014 5th International Conference on Information Science and Applications. 1-4. 10.1109/ICISA.2014.6847379.

[5] Lee, Jung-Hun & Kim, Minho & Kwon, Hyuk-Chul. (2017). The Utilization of Local Document Information to Improve Statistical

Context-Sensitive Spelling Error Correction. KIISE Transactions on Computing Practices. 23. 446-451. 10.5626/KTCP.2017.23.7.446.

[6] Lee, Jung-Hun & Kim, Minho & Kwon, Hyuk-Chul. (2017). Improved Statistical Language Model for Context-sensitive Spelling Error Candidates. Journal of Korea Multimedia Society. 20. 371-381. 10.9717/kmms.2017.20.2.371.

[7] Xie, Ziang, et al. Neural language correction with character-based attention. arXiv preprint arXiv:1603.09727. 2016.

[8] Woo Cho, Seung & Kwon, Hong-seok & Jung, Hun-young & Lee, Jong-Hyeok. (2018). Adoption of a Neural Language Model in an Encoder for Encoder-Decoder based Korean Grammatical Error Correction. KIISE Transactions on Computing Practices. 24. 301-306. 10.5626/KTCP.2018.24.6.301.

[9] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. CoRR, abs/1701.02810.

[10] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, Tomas Mikolov, FastText.zip: Compressing text classification models, arXiv:1612.03651