Generating Diverse Corrections with Local Beam Search for Grammatical Error Correction

Kengo Hotate, Masahiro Kaneko, and **Mamoru Komachi**
Tokyo Metropolitan University, Tokyo, Japan
{hotate-kengo, kaneko-masahiro}@ed.tmu.ac.jp
komachi@tmu.ac.jp

Abstract

In this study, we propose a beam search method to obtain diverse outputs in a local sequence transduction task where most of the tokens in the source and target sentences overlap, such as in grammatical error correction (GEC). In GEC, it is advisable to rewrite only the local sequences that must be rewritten while leaving the correct sequences unchanged. However, existing methods of acquiring various outputs focus on revising all tokens of a sentence. Therefore, existing methods may either generate ungrammatical sentences because they force the entire sentence to be changed or produce non-diversified sentences by weakening the constraints to avoid generating ungrammatical sentences. Considering these issues, we propose a method that does not rewrite all the tokens in a text, but only rewrites those parts that need to be diversely corrected. Our beam search method adjusts the search token in the beam according to the probability that the prediction is copied from the source sentence. The experimental results show that our proposed method generates more diverse corrections than existing methods without losing accuracy in the GEC task.

1 Introduction

Grammatical error correction (GEC) is a task that corrects grammatical errors in an input text. Depending on the input, there are multiple ways to correct such text. For example, 10 annotators can produce 10 different valid correction results for the same grammatically incorrect text (Bryant and Ng, 2015). If a GEC model presents multiple candidates for correction, it helps the user decide whether to utilize the correction results such that the user can select a favorite correct expression from among the candidates.

However, currently existing GEC models do not consider the generation of multiple correction candidates. Generally, in GEC, the method for obtaining multiple corrections involves the use of a *plain beam search* to generate the n-best candidates (Grundkiewicz et al., 2019; Kaneko et al., 2020). However, it has been shown that a plain beam search does not provide a great enough variety of candidates and produces lists of nearly identical sequences (Vijayakumar et al., 2018). Therefore, the n-best candidates generated by a beam search without diversity control are not expected to provide useful additional information. Considering this problem, several beam search methods have been proposed to generate diverse candidates (Li et al., 2016; Vijayakumar et al., 2018; Kulikov et al., 2019). These diverse beam search methods encourage diversity by globally rewriting all tokens in a sentence. We will refer to such methods as *diverse global beam search* methods. Conversely, considering a local sequence transduction task in GEC, wherein most of the tokens in the source and target sentences overlap, excessive correction of the input sentence is not preferred because unnecessary rewriting damages the grammatically correct parts of the input sentence. Furthermore, encouraging more corrections than necessary decreases the performance of the GEC itself (Hotate et al., 2019). We hypothesize that both plain beam search and diverse global beam search methods may not be suitable for GEC tasks, and a GEC model must correct the grammatical errors of the input sentence in diverse ways while preserving the correct portions of the sentence.

Proceedings of the 28th International Conference on Computational Linguistics, pages 2132–2137
Barcelona, Spain (Online), December 8-13, 2020

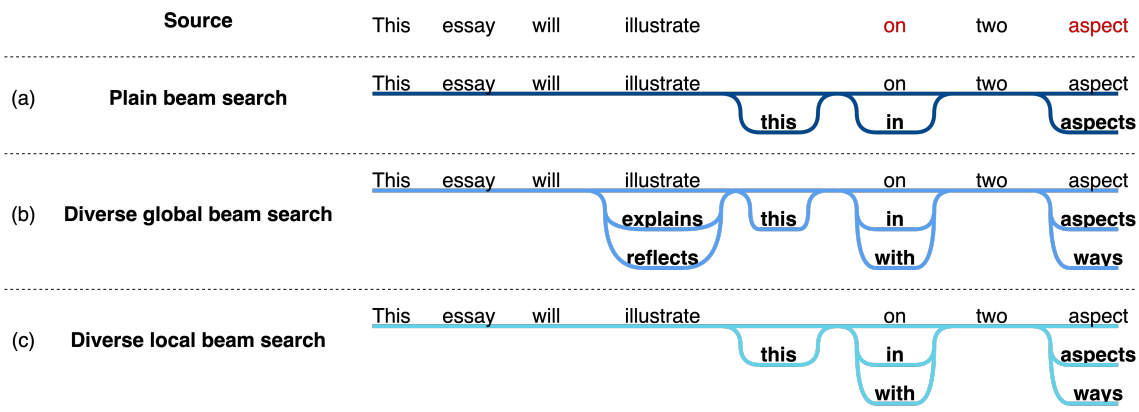

Figure 1: Illustration of error corrections obtained with previous and proposed beam search methods.

In this study, we propose a *diverse local beam search* method to obtain diverse outputs considering whether a token should be corrected during the beam search. Note that our method can be used for any local sequence transduction task. Figure 1 shows a comparison of existing and proposed methods. The proposed beam search method considers the following: (a) In plain beam search, the correction is concentrated on a specific path. Therefore, this method generates sentences with similar token combinations and a small number of word types. (b) The diverse global beam search method explores many different paths. Therefore, unlike plain beam search, this method generates sentences with various token combinations and a large number of word types. However, it also generates correction candidates for tokens that are not corrections. (c) The proposed diverse local beam search expands different paths only for tokens that require corrections. Therefore, our diverse local beam search generates sentences with combinations of more diverse tokens than plain beam search only at points that need correction. It should be noted that all the above methods have the same n but different paths. The experimental results show that our diverse local beam search could generate more diverse and accurate n-best candidates than the existing methods. The performance of the general evaluation datasets in the GEC task demonstrated almost no deterioration.

2 Related work

Several studies have been proposed to obtain diverse outputs using beam search. Li et al. (2016) modified standard beam search to penalize the scores of searches with the same parent node. Their algorithm recommends only those hypotheses that come from different parent beams. Vijayakumar et al. (2018) proposed a method to divide the beam into several groups and perform beam search for each group. Additionally, they added a constraint to make it harder to select tokens selected by other groups in the same time step. Kulikov et al. (2019) proposed an iterative beam search, which produces a more diverse set of candidate responses in neural dialogue modeling. However, these studies do not distinguish between the parts that do not need to be rewritten and those parts that require diverse corrections.

3 Diverse local beam search

Diverse local beam search encourages candidates of diverse corrections for parts of the input sentence that must be corrected and discourages candidates for already correct parts of the input sentence. Consequently, it runs the computation for fewer parts to generate diverse candidates. For this purpose, a penalty score $s_{b,t}$ is assigned to each beam b at each time step t, indicating whether a correction should be made. Although different methods can be used to calculate a penalty score, in this study, we use a copy probability from the copy-augmented model (Zhao et al., 2019) as a penalty score $s_{b,t}$. We explain the copy-augmented model in greater detail in Section 4.1. Using the penalty score, we penalize the beam search score k as follows:

$$k_{b,t} = (\lambda s_{b,t} + \beta) \log p_{b,t} \quad (1)$$

where p is the output distribution of the GEC model. β and λ are hyperparameters, where β prevents the penalty from falling to zero and λ determines the strength of the penalty.

4 Experiments

4.1 Model

We used the copy-augmented model (Zhao et al., 2019) as the GEC model. This model controls the balance between the copy distribution p^{copy} and generation distribution p^{gen} via the balancing factor α^{copy}. p^{copy} is the probability distribution of the tokens to be copied from the source sentence, and p^{gen} is the generation probability distribution that predicts the output tokens. The output distribution $p_{b,t}$ is calculated as follows:

$$p_{b,t} = \left(1 - \alpha_{b,t}^{\text{copy}}\right) p_{b,t}^{\text{gen}} + \alpha_{b,t}^{\text{copy}} p_{b,t}^{\text{copy}} \tag{2}$$

Copying from a source sentence or generating a token can be considered as a choice between correcting or not. Therefore, we used $\alpha_{b,t}^{\text{copy}}$ as the penalty score $s_{b,t}$ in Equation 1.

As a diverse global beam search, we used the diverse beam search (Vijayakumar et al., 2018) approach, wherein the number of groups is defined from the number of desired diverse sentences, and a diversity strength for the beam search is selected such that the output tokens at each time step in each group differs. We set the number of groups to n of n-best and the diversity strength to 0.7 for diverse global beam search. For diverse local beam search, we used $\beta = 1.0$ and $\lambda = 4.0$ for diverse local beam search[1].

4.2 Datasets

For models that have been pre-trained with publicly available pseudo-data, we fine-tuned them using published training data[2]. We used the public NUCLE (Dahlmeier et al., 2013), Lang-8 (Mizumoto et al., 2011), and FCE (Yannakoudakis et al., 2011) corpora as our training data. We used the JFLEG test set and dev set corrected by four different annotators (Napoles et al., 2017) as the development set. We also used the CoNLL-2014 dataset as the test set. The original CoNLL-2014 dataset was corrected by two different annotators (Ng et al., 2014). However, in this work, eight corrections made by (Bryant and Ng, 2015) and four corrections with minimal corrections made by (Sakaguchi et al., 2016) were also used as references.

4.3 Evaluation

Performance of GEC (G-score). We evaluated each decoding method using the GLEU score (Napoles et al., 2017) for the JFLEG and the $\text{F}_{0.5}$ score by using the MaxMatch scorer (Dahlmeier and Ng, 2012) for the CoNLL-2014 test set as general evaluation metrics for the GEC.

Diversity of corrections (C-score). To evaluate the diversity of corrections, we calculated the coverage score between the n-best candidates and references. We used the weighted recall, which was used as the evaluation metric in the 2020 Duolingo Shared Task[3] as a coverage score. In this work, the number of duplicated corrections divided by the total number of corrections is used for weighting. This gives higher weight to corrections with more duplicates.

Correctness of corrections (DF score). To evaluate the correctness of the corrections, we distinguished between the correct and incorrect parts of the input sentences and considered the correction outputs acceptable if they only change tokens that require corrections. In this study, we used document frequency (DF) to assess the correctness of the corrections.

First, we calculated the number of n-grams that are present in both the source and predicted sentences[4]. Then, we calculated the DF score by dividing it by the number of all n-grams in both the source and

[1] We searched for β in increments of 0.1 with a range of 0.0 to 1.0 and λ in increments of 1 with a range of 1 to 10.

[2] `https://github.com/zhawe01/fairseq-gec`

[3] `https://sharedtask.duolingo.com/`

[4] In this case, the number of n-grams in the predicted sentences is limited to the number of occurrences in the source sentences.

	Method	$n = 10$			$n = 15$			$n = 20$		
		G-score	C-score	DF score	G-score	C-score	DF score	G-score	C-score	DF score
JFLEG	PBS	50.92	29.75	84.92	50.92	31.80	83.24	50.92	33.18	82.08
	DGBS	**51.24**	19.99	83.08	**51.27**	20.65	78.97	**51.27**	21.30	76.53
	DLBS	50.97	**30.35**	**85.50**	50.94	**32.64**	**84.26**	50.92	**33.74**	**83.35**
CoNLL	PBS	**59.67**	32.61	87.04	**59.64**	34.29	84.99	**59.61**	35.10	83.49
	DGBS	57.88	24.26	86.83	57.94	25.13	81.98	57.89	25.61	79.67
	DLBS	59.47	**33.79**	**87.51**	59.44	**35.10**	**85.67**	59.40	**36.38**	**84.65**

Table 1: Evaluation results for 10-, 15-, and 20-best candidates. The G-score denotes performance of the GEC task, which is GLEU for JFLEG and $F_{0.5}$ for CoNLL-2014. The coverage score (C-score) and DF score denote the diversity and correctness of the corrections, respectively.

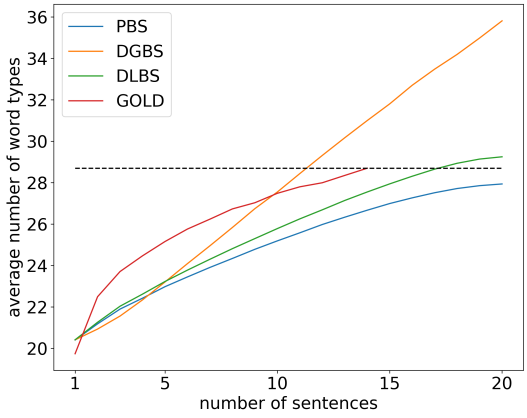

Figure 2: Illustration of the average number of word types per sentence for each method.

reference sentences. Finally, the average score of multiple outputs per source sentence was calculated over the entire test set.

4.4 Results

Table 1 shows the evaluation results of our two baselines of plain beam search (PBS) and diverse global beam search (DGBS), as well as those of our proposed diverse local beam search (DLBS). We performed experiments for 10-, 15-, and 20-best candidates. Thus, DGBS has lower coverage scores compared to PBS. Moreover, the DF scores of DGBS are lower compared to PBS. These results show that DGBS cannot produce more diverse outputs than PBS in GEC. We hypothesized that this is because DGBS attempts to rewrite and diversify all tokens regardless of their correctness in the source sentence.

By contrast, our proposed approach's coverage scores and DF scores are higher than those of our baselines. Therefore, we can conclude that our proposed method diversifies only the tokens that must be corrected in the source sentence.

4.5 Analysis

Diversity has two aspects: the diversity of the correction points and that of the words to be corrected. The DF score can measure the diversity of the corrected parts, but not the diversity of the corrected words. Here, we analyze the number of word types for the output as the diversity of the words to be corrected. Figure 2 shows the average number of word types per sentence for each method. The horizontal axis represents n of the n-best or the number of reference sentences, and the vertical axis represents the average number of word types per sentence. The GOLD value indicates a reference for CoNLL-2014,

Source		You can share photos , videos and *every* meaningful *experiences of* your life .
PBS	✓	You can share photos , videos and every meaningful **experience** of your life .
	✓	You can share photos , videos and every meaningful **experience in** your life .
		You can share photos , videos **,** and every meaningful **experience** of your life .
DGBS	✓	You can share photos , videos and every meaningful **experience** of your life .
		You can share photos , videos and experiences of your life **with everyone else** .
		You can share photos , videos and every meaningful **experience** of life .
DLBS	✓	You can share photos , videos and every meaningful **experience** of your life .
	✓	You can share photos , videos and every meaningful **experience in** your life .
	✓	You can share photos , videos and **all the** meaningful experiences of your life .

Table 2: Examples of outputs from CoNLL-2014. The *italic* tokens represent the tokens that are corrected in reference, and the **bold** tokens represent tokens generated by the model that are different from the source sentence. The ✓represents the output sentences that matched one of the references.

and the dashed line indicates the maximum value of GOLD. In all methods, the number of word types increases as the number of sentences increases. However, in GOLD, the increase is particularly small as the number of references increases. This implies that, although there are several references, not all of them use different words. PBS has fewer word types for the output than GOLD. The number of word types in DGBS increases linearly with the number of n, indicating that DGBS generates more word types than necessary when compared to GOLD. Conversely, DLBS can diversify the output more appropriately than PBS and DGBS.

4.6 Examples of corrections

Table 2 shows the output examples of each decoding method. We can observe that DGBS has unnecessarily changed parts of the input sentence (e.g., "meaningful" and "your"). Conversely, the proposed method generates diverse outputs only for the tokens that should be rewritten (e.g., "every," "experiences," and "of").

5 Conclusion

Existing methods of acquiring diverse outputs rewrite all tokens of the input sentence. In this work, we proposed a method to produce diverse candidates for only the parts that require corrections via GEC. It was shown that the proposed method can diversify the outputs more properly in GEC than the existing methods. In the future, we would like to apply this method to other model architectures and extend it to tasks other than GEC.

Acknowledgements

We would like to thank Aizhan Imankulova for useful discussions. We gratefully thank Yangyang Xi and Lang-8 contributors for sharing their data.

References

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proc. of ACL*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proc. of NAACL-HLT*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The nus corpus of learner English. In *Proc. of BEA*.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proc. of BEA*.

Kengo Hotate, Masahiro Kaneko, Satoru Katsumata, and Mamoru Komachi. 2019. Controlling grammatical error correction using word edit rate. In *Proc. of ACL SRW*.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proc. of ACL*.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proc. of INLG*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *ArXiv*, abs/1611.08562.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proc. of IJCNLP*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proc. of EACL*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proc. of CoNLL*.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Proc. of TACL*.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proc. of AAAI*.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proc. of ACL*.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proc. of NAACL-HLT*.