Data Augmentation for SentRev using Back-Translation of Lexical Bundles

Zhendong Du
Waseda University
Kitakyushu, Japan
zhendong@fuji.waseda.jp

Kenji Hashimoto
Waseda University
Kitakyushu, Japan
kenji.hashimoto@waseda.jp

Abstract

Sentence-level Revision (SentRev) is dedicated to enhancing the English writing fluency of non-native English speakers. However, due to the lack of high-quality training data, outstanding results have not been achieved in the past. The synthetic training data generation method employed in the baseline work did not truly address the fundamental pain point of non-native English speakers′ inability to produce fluent English writing. In this study, we propose a novel method for synthetic data generation by utilizing the technique of Back-Translation of lexical bundles to disrupt sentences in academic papers and obtain parallel corpora. We evaluated our data on three Grammatical Error Correction (GEC) models using multiple metrics, and significant improvements are observed in comparison to the baseline.

1 Introduction

The limited English proficiency among many non-native English-speaking researchers has emerged as a prominent issue, impeding their ability to effectively disseminate their research findings in the English language. Therefore, an increasing amount of attention has been devoted to the provision of writing support for non-native English speakers.

The field of Grammatical Error Correction (GEC) has made tremendous bounds in the past few years, especially with the introduction of Transformer (Vaswani et al., 2017), GEC system has been able to deal with most of the English grammar errors and can give good corrective results. Nowadays, non-native English speakers can easily solve grammatical errors very well, however, for academic writing, academic style English expression is also essential.

For non-native English speakers, learning English faces negative transfer problems (Smith Jr, 1958), making it difficult for most of them to write authentic academic English texts. In addition, there is a lack of a clear definition of what is considered academic style English in academia.

This is a complex linguistic issue, and many researchers have worked on the composition of English in depth, and have come to many constructive and helpful conclusions: (Wray, 2000) proposed formulaic sequences in second language teaching, the author consider the mastery of idiomatic forms of expression very important, and refers to (Willis, 1990) , (Nattinger and DeCarrico, 1992) , (Lewis, 1993) to stress that: larger units can, and should, be perceived by the learner and teacher in terms of their component parts.

As a larger unit, lexical bundles have been proven by many researchers (e.g. (Biber and Conrad, 1999), (Hyland, 2008)) to be very important for fluent English expressions. We believe that academic style English sentences should have sufficient standard lexical bundles, and the authors have taken this into account in (Goh and Lepage, 2019) .They extract and publicly release more than 18,000 lexical bundles from the ACL Anthology Reference Corpus (ACL-ARC) (Bird et al., 2008). ACL-ARC is a collection 10, 920 academic papers from the ACL Anthology, so the English sentences of this corpus conform to the academic style in common sense, and the lexical bundles extracted from them are representative.

We believe that lexical bundle can be very helpful for non-native English speakers in writing English. Since there is no clear definition of academic writing style, we propose Lexical Bundle as a consideration of academic writing style. We would like to have a corresponding parallel corpus, i.e., non-academic style sentences and academic style sentences. Considering the characteristics of this task, we found (Ito et al., 2019)'s work to be similar to our idea. The authors of this work proposed Sentence-level Revision (SentRev) as a new academic writing support task, which takes non-academic style sentences as Drafts and academic style sentences as References, and then converts Drafts to References through several operations. The problem, however, is that such a parallel corpus is equally difficult to obtain. The authors' method is to select a number of sentences from the ACL Anthology Sentence Corpus (AASC)[1] as References, translate these sentences into Japanese by machine translation, and then have native Japanese speakers who are not very good at English translate them back into English manually, and the re-obtained sentences are used as Drafts, thus obtaining a parallel corpus. Although the quality of the parallel corpus obtained in this way is high enough, the high labor cost makes it difficult to achieve large-scale corpus generation, so the parallel corpus generated in this way is used by the authors as a evaluation dataset, rather than a training set. The authors named this evaluation dataset the SMITH dataset. As an alternative, the authors have used a series of methods to generate a synthetic dataset as a training set.

The authors used three strategies (Heuristic noising and denoising model, Enc-Dec noising and denoising model and GEC model (Zhao et al., 2019)) to establish baseline scores, and the model performance was evaluated with multiple metric and then the final evaluation score was very low. We believe this is mainly due to the low quality of the authors' training data, which cannot simulate the manual annotation as well as the SMITH dataset.

Our goal is to propose a new synthetic data generation method: to obtain new sentences (Drafts) by back-translating and destroying the lexical bundles in academic style English sentences (References). We employed this approach to generate a substantial amount of synthetic data, which was subsequently used to fine-tune (Howard and Ruder, 2018) three GEC models. The results demonstrate a significant improvement of our method over the baseline across multiple metrics, thus confirming the effectiveness of our approach.

2 Related work

2.1 Gramatical Error Correction (GEC)

GEC is the task of detecting and correcting grammatical errors in texts written by non-native English writers. The goal is to convert a sentence containing a grammatical error into a correct sentence without the grammatical error. Since both incorrect and correct sentences are sequences, one of the past research approaches was to treat it as a Machine Translation (MT) task, i.e., MT models learn the mapping from the source sentence to the target sentence, to translate incorrect sentences into correct sentences. (Felice et al., 2014) (Junczys-Dowmunt and Grundkiewicz, 2014).

With the development of deep learning, the approach of Neural Machine Translation (NMT) has also been applied to GEC and has become one of the mainstream approaches for GEC in recent years (Kalchbrenner and Blunsom, 2013) (Bahdanau et al., 2015). Recent work (Omelianchuk et al., 2020) has used tagging sequences as an alternative to the mainstream Seq2Seq model, which has excellent inference speed and requires less training data. In addition, (Stahlberg and Kumar, 2020) et al. proposed Seq2Edit, which does not perform well for GEC tasks that require large scale corrections, but works well for local corrections. The characteristics of the GEC task were similar to our goal, except that we wanted to convert ″bad″ English expressions to ″good″ English expressions in the presence of syntactic errors.

2.2 Back-Translation

Back-Translation is a technique used to improve the quality of machine-translated text. It involves translating text from one language to another, and then translating the translated text back to the original language.

The purpose of Back-Translation is to identify errors and inconsistencies in the machine translation process. By comparing the back-translated text with the original text, one can identify areas where the initial translation system may have made errors or failed to capture the intended meaning of the text. This information can then be used to improve the translation system and increase the accuracy and quality of machine-translated text. Back translation was first applied to parallel corpus generation for Neural Machine Translation by (Sennrich et al., 2016), and (Xie et al., 2018) was inspired to use it in parallel corpus generation for GEC, achieving excellent performance.

2.3 Sentence-level Revision (SentRev)

A recent work presents a new task: SentRev. The authors regarded enhancing the fluency of academic English writing as a sentence-level rewriting process. They generated a substantial amount of synthetic training data using several common Data Augmentation methods employed in other tasks.

[1] https://github.com/KMCS-NII/AASC

Three models were trained using this data, and baseline scores were reported on the SMITH dataset using multiple metrics. However, due to the authors' training data generation methods not taking English fluency into account, the baseline scores were not satisfactory. In contrast to traditional NLP tasks like machine translation, the parallel corpus for SentRev needs to have sufficiently high quality to benefit the model's performance. Therefore, despite the synthesis of over two million pairs in the baseline, it still fell short in adequately addressing this task.

2.4 SMITH dataset

The SMITH dataset (Ito et al., 2019) is a parallel corpus designed for evaluating the effectiveness of academic style transfer, and its generation method is illustrated in Fig. 1

The main approach of this method is to extract and screen sentences (references) from a number of published papers, translate them into another language (Japanese) using machine translation, and then have several native Japanese speakers manually translate them back into English. In this way, a non-academic style English sentence (draft) is obtained, which constitutes a parallel corpus.

The authors commissioned language experts to evaluate the generated corpus and found that only 5 percent of the translations were noticeably inadequate, leading to the conclusion that the corpus generated through this approach is effective. However, the problem is that the authors only obtained 10,804 sentence pairs at a cost of 4,200 dollars, making this manual approach unsuitable for generating large-scale corpora. As a result, the authors only used it as an evaluation dataset.

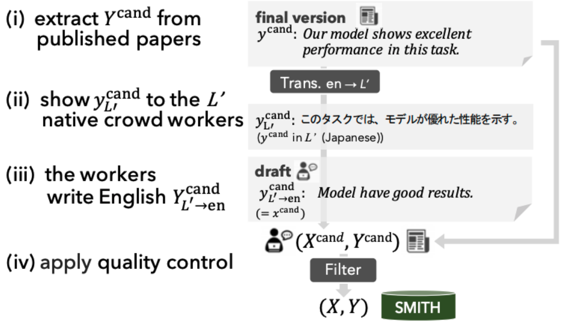
(i) extract Y^{cand} from published papers
(ii) show $y^{\text{cand}}_{L'}$ to the L' native crowd workers
(iii) the workers write English $Y^{\text{cand}}_{L'\to\text{en}}$
(iv) apply quality control
final version
y^{cand}: Our model shows excellent performance in this task.
Trans. en → L'
$y^{\text{cand}}_{L'}$: このタスクでは、モデルが優れた性能を示す。
(y^{cand} in L' (Japanese))
draft
$y^{\text{cand}}_{L'\to\text{en}}$: Model have good results.
($= x^{\text{cand}}$)
$(X^{\text{cand}}, Y^{\text{cand}})$
Filter
(X, Y)
SMITH

Figure 1: The Generation Method of the SMITH Dataset. Figure copied from (Ito et al., 2019)

2.5 ACL Anthology Sentence Corpus (AASC)

AASC is a corpus of natural language text extracted from ACL Anthology, a comprehensive repository of scientific papers on computational linguistics and natural language processing, containing 2,339,195 sentences from PDF papers.

2.6 ACL Anthology Reference Corpus (ACL-ARC)

ACL-ARC (Bird et al., 2008) is an enhanced and standardized reference corpus extracted from ACL Anthology. The goal of ACL-ARC is to become a widely available standard testbed that encourages other researchers to use it for bibliometric research and reference.

2.7 Lexical Bundle

Lexical bundle is a linguistic term referring to a group of common words or phrases that frequently occur and combine in a fixed way in language to convey a particular meaning or express a particular concept. They can be considered as idiomatic expressions that are typically used in specific contexts and situations, and are among the most common lexical units in everyday English. For instance, in scientific papers, common academic lexical bundles include phrases such as ″in terms of,″ ″as shown in,″ ″it is worth noting that,″ and so on. These phrases are composed of several words, with fixed syntax and usage, conveying the same meaning across different contexts.

Lexical bundles are considered a useful unit of analysis in language research as they provide insight into the patterns of language use and can aid in the development of natural language processing systems.

A recent work (Goh and Lepage, 2019) has demonstrated that the use of lexical bundles is essential for fluent academic writing, Non-native speakers of English usually lack the capability of using bundles in their writing. The authors extracted 18,000 publicly available lexical bundles from ACL-ARC, we hope that this outcome can be efficiently utilized to provide English writing support to non-native English speakers.

3 Method

We filtered several sentences from AASC and from the corpus we generated following the same strategy as AASC, and lexical bundles previously extracted from (Goh and Lepage, 2019) were used as a dictionary. These bundles were then extracted in order from all of the selected sentences to serve as the input text for back-translation. After obtaining the Back-Translation results, we conducted comparative experiments using (Ito et al., 2019)'s synthetic data as a baseline and validated the results on the SMITH dataset. Fig. 2 shows that the basic process of data generation.

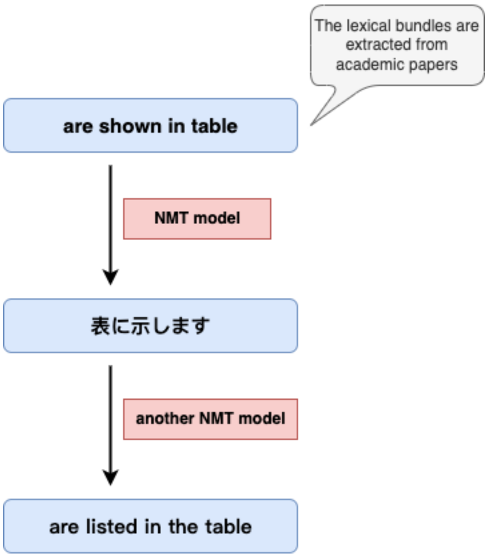

Figure 2: The basic process of data generation.

3.1 Data

Our objective is to develop a better artificial data synthesis method relative to the baseline. The baseline approach is similar to the data generation method used for the SMITH dataset, where a number of sentences were selected from AASC (excluding those already used in the SMITH dataset) based on certain filtering strategies. The baseline method then used four techniques to synthesize training data: Heuristic Noising and Denoising, Grammatical Error Generation, Style Removal, and Entailed Sentence Generation. Our approach is to generate a new corpus of sentences according to AASC, and combine the remaining unused sentences in AASC, and then follow the same filtering strategy sentences as baseline, and use the lexical bundle dictionary to generate synthetic drafts according to our Back-Translation strategy by using the sentences that meet the requirements as candidates.

We evaluated the performance of our model using the SMITH dataset, which contains a development set of 500 sentence pairs and a test set of 10, 034 sentence pairs.

3.2 Back-Translation strategy

Previous works have typically employed either Beam Search (Sennrich et al., 2016) or solely relied on Greedy Search (Imamura et al., 2018) for generating synthetic data. However, the utilization of the Beam Search and Greedy Search primarily concentrates on the uppermost region of the model's distribution, leading to the generation of synthetic source sentences that exhibit a high degree of regularity and fail to adequately capture the underlying data distribution. As alternative, (Edunov et al., 2018) consider sampling from the model distribution as well as adding noise to beam search outputs.

We evaluated the above three methods and found that they generate a lot of noise which causes the semantics of the back-translated sentence to change significantly compared to the original sentence, which is contrary to our needs. Our work is distinct from previous works in that our corpus consists of lexical bundles, while prior research has focused on sentence-level analysis. Additionally, we do not intend to introduce artificial noise into the generated translations, as our aim is solely to create parallel corpora from the perspective of writing style, rather than random errors such as grammatical or spelling mistakes. Therefore, we have opted to utilize large pre-trained language models for the purpose of conducting Back-Translation.

3.3 Model

3.3.1 Back-Translation model

We employed T5-base (Raffel et al., 2020) to translate English lexical bundles into Japanese. T5-base is a pre-trained language model based on the Transformer architecture. Due to its extensive pretraining on a large corpus of resources, it does not tend to generate academic text during Back-Translation, as some pre-trained models focused primarily on academic corpora might. Furthermore, T5-base provides reliable translation accuracy.

When translating the obtained results back to English, we utilized M2M-100 (Fan et al., 2020). The M2M-100 model was trained with a large-scale automated data generation technique, including the automatic extraction of sentence pairs from open-source datasets such as Wikipedia and CCAligned (El-Kishky et al., 2020), and the use of natural language generation techniques to create new sentence pairs. Furthermore, the model employed both word-level and subword-level encoders to better handle the complex grammar structures of Japanese. Additionally, due to the utilization of extensive non-academic corpora in training, the model produced many results in which the lexical bundles had synonymous but different forms from the source.

3.3.2 Data validation model

We have referred to the baseline idea and reviewed our task and concluded that our task is better suited to be implemented with a GEC model, i.e., the process of converting non-academic English expressions into Lexical bundle is treated as grammatical error correction. A good GEC model not only handles this process precisely, but also corrects grammatical errors in the synthetic corpus together at the same time.

Considering that the replacement for lexical bundles involves only a small part of the sentence modification, we believe that the recent Tagging model and the Seq2Edit model might work better than the NMT-based model. To better validate the performance of our data, we chose the same GEC model using a copy-augmented architecture (Zhao et al., 2019) as the baseline, the Tagging GEC model GECToR, and Seq2Edit, which we will later call **GEC-1**, **GEC-2** and **GEC-3**, respectively.

3.4 Evaluation metrics

3.4.1 Semantic evaluation metrics

We employed cosine similarity to evaluate the generated corpus, which is formulated as follows:

$$similarity = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| ||\mathbf{B}||}$$

Here, $\mathbf{A}$ and $\mathbf{B}$ represent two vectors, θ represents the angle between them, and $||\mathbf{A}||$ and $||\mathbf{B}||$ represent their magnitudes.

After obtaining the cosine similarity of each set of parallel corpora, the following formula is used to calculate their average value:

$$average\_similarity = \frac{\sum_{i=1}^{n} similarity_i}{n}$$

The symbol $similarity_i$ denotes the similarity score of the i-th pair of parallel corpus, while n represents the total number of parallel corpus pairs. The formula expresses the averaging of similarity scores for all parallel corpus pairs, which involves summing the similarity scores of each pair and dividing the sum by the total number of pairs.

3.4.2 Model performance evaluation metrics

To fully evaluate this work, we refer to the baseline work with some modifications and evaluated the model performance from multiple perspectives using BLEU (Papineni et al., 2002), ROUGE-L, BERT-score (Zhang et al., 2020), grammaticality score (Napoles et al., 2016), PPL and $F_{0.5}$.

In particular, we used LanguageTools[2] to calculate the number of syntactic errors in the sentences, KenLM[3] to calculate the PPL and ERRANT (Bryant et al., 2017) to calculate $F_{0.5}$.

4 Experiments

4.1 Data Generation

Since the data of AASC is limited and a large part of it is already used by the baseline work, we generated a large corpus of new sentences following the AASC generation method and used them as a supplement to the AASC.

We referred to the data selection strategy of the SMITH dataset, selecting sentences with lengths between 70 and 120 characters and containing no special symbols. We then compared these sentences with all the references in the SMITH dataset and filtered out the duplicate ones. We processed the candidate bundles text from (Goh and Lepage, 2019) by removing all extraneous fields except for lexical bundles, and used it as a dictionary. We used this dictionary as a reference to traverse the previously extracted sentences one by one, extracting all discovered lexical bundles and rearranging them in their original positions in a new document. We generated synthetic drafts using our translation strategy. Some examples of generated results are shown in Table 1

4.2 Data Analysis

Initially, we analyzed the results by counting all the parallel corpora where the source lexical bundle was the same as the Back-Translation result. The statistical analysis indicated that this subset of data accounted for approximately **35.5%** of the corpus, this part of the result is considered invalid data and they are removed. Due to the characteristics of Back-Translation, the presence of ineffective results in this portion aligns with our initial expectations, and the proportion of such results is relatively low, thereby substantiating the effectiveness of our

[2] https://github.com/languagetool-org/languagetool/releases/tag/v3.2

[3] https://github.com/kpu/kenlm

Source lexical bundle	Result translated into Japanese	Result translated back into English
are expressed by	によって表現される	are conveyed through
approach could be	アプローチは	strategy may be
an explanation for	説明	a cause of
a more detailed description of	より詳細な説明	a comprehensive account of
we present a method of	方法を提示する	we introduce an approach for
the work presented in this paper	この論文で紹介されている作品	the research outlined in this article

Table 1: Some examples of generated results

Back-Translation strategy. Subsequently, we employed Sent2Vec (Moghadasi and Zhuang, 2020) to transform all remaining parallel corpora into vectors, and then computed the average similarity score, which was found to be **0.79**. This result provides evidence that the majority of the parallel corpus obtained in this method ensures semantic similarity.

It is worth noting that in the generated results, many identical lexical bundles were translated into different results. While this may be considered undesirable for other tasks, it is an advantage for our specific task. This is because, typically, different individuals exhibit distinct stylistic preferences when writing in English, and these diverse translations correspond to those preferences. Moreover, they form a many-to-one mapping relationship with the lexical bundles, which means that various non-academic English writing styles can be modified to conform to a standard academic English writing style.

We inserted the Back-Translated results back into the original sentences in the position and order in which they were extracted, and removed those corpus that did not have any changes. The percentage of the corpus with no alterations constituted approximately **41%** of the total data. This relatively high percentage can be attributed, in part, to a number of invalid Back-Translation results, as well as the fact that our dictionary was unable to cover all lexical bundles, particularly those that are discontinuous or have changed tense. Ultimately, we obtained 2,600,355 pairs of parallel sentences (The baseline was 2,269,216 pairs, and in the experiment we performed a control variable to control for their numbers), and then subjected to a sampling analysis, found that many sentences thus produced some grammatical errors, especially the missing or incorrect use of prepositions. Table 2 shows some examples of the parallel corpus obtained. Considering that we will later use a grammar error correction model to validate the data, these grammatical errors will theoretically be well modified.

4.3 Data Performance Validation

To reasonably validate the validity of our data, we set up three sets of control trials for each of the three GEC models. (The synthetic data of the four methods for the baseline will be referred to as **Data-base**, and our data will be referred to as **Data-our** later). The experiment settings is shown in Table 3:

We use the performance of Data-base in the three GEC models as the control group, and then set up three experimental groups. Experimental group 1 is used to verify the performance of the three models after completely replacing Data-base with Data-our; experimental group 2 is used to verify the performance of the three models when Data-base and Data-our are used together, and experimental group 3 is used to verify the performance of the three models when Data-our is partially replaced by Data-base. Incidentally, In experimental group 1, we reduced the number of data to the same as the baseline in order to control for variables, and in experimental group 2, we used the full baseline data and our synthetic data, in experimental group 3, the percentage of data settings were set empirically. The results of all experimental evaluations are shown in Table 4:

4.4 Experimental results analysis

The results of experimental group 3 show that our data as expanded data can effectively improve the model performance. The results of experimental group 1 scored lower than the baseline because the baseline data generation methods had a large lexical-level transformation of the sentences, which allowed the fine-tuned model to transform the whole sentences, while our method could only transform a small part of the sentences, so it was not as effective for the whole sentences. As an ablation experiment, experimental group 3 validated partial replacement of baseline data for the data

Source sentence 1	But anyway, a system like this **will be a contribution** to the development of intelligent systems.
Result 1	But anyway, a system like this **will be beneficial** to the development of intelligent systems.
Source sentence 2	The nouns **play a key role** in the understanding part as they constitute the class or type hierarchy.
Result 2	The nouns **be essential** in the understanding part as they constitute the class or type hierarchy.
Source sentence 3	**It was decided that** a text based information system should be built, **regardless of** the status of the speech rocgnition and speech synthesis effort, **which proved to** lag behind after a while.
Result 3	**The decision was made** a text based information system should be built, **in spite of** the status of the speech rocgnition and speech synthesis effort, **proven** lag behind after a while.

Table 2: Some examples of the parallel corpus obtained

Group	**Model**	**Data**
Control group (baseline)	GEC-1	100% Data-base
	GEC-2	100% Data-base
	GEC-3	100% Data-base
Experimental group 1	GEC-1	100% Data-our (The quantity is equal to the baseline)
	GEC-2	100% Data-our (The quantity is equal to the baseline)
	GEC-3	100% Data-our (The quantity is equal to the baseline)
Experimental group 2	GEC-1	100% Data-base + 100% Data-our
	GEC-2	100% Data-base + 100% Data-our
	GEC-3	100% Data-base + 100% Data-our
Experimental group 3	GEC-1	75% Data-base + 25% replaced by Data-our
	GEC-2	75% Data-base + 25% replaced by Data-our
	GEC-3	75% Data-base + 25% replaced by Data-our

Table 3: Experiment settings

Group	Model	P	R	$F_{0.5}$	BLEU	ROUGE-L	BERT-P	BERT-R	BERT-F	Gramm.	PPL
Control group (baseline)	GEC 1	22.2	6.2	14.6	11.9	49.0	80.8	79.1	79.9	96.7	414
	GEC 2	24.2	5.8	14.8	17.5	51.3	81.2	80.3	80.7	**97.3**	401
	GEC 3	19.5	5.5	12.9	15.0	45.2	79.9	79.3	79.6	95.8	419
Experimental group 1	GEC 1	11.6	4.9	9.1	11.9	31.7	53.6	49.3	51.4	96.4	570
	GEC 2	14.4	3.2	8.4	8.2	28.3	63.5	62.1	62.8	95.5	530
	GEC 3	14.5	8.5	12.7	12.0	37.6	65.2	65.0	65.1	96.9	433
Experimental group 2	GEC 1	27.3	12.1	21.8	27.5	62.0	83.9	82.1	83.0	97.1	299
	GEC 2	**34.5**	**20.0**	**30.1**	**31.5**	**63.4**	**84.0**	**83.5**	**83.7**	96.9	**273**
	GEC 3	30.1	18.8	26.9	26.2	58.1	83.8	83.0	83.4	97.2	283
Experimental group 3	GEC 1	25.3	11.8	18.1	16.7	55.1	81.3	80.1	80.7	**97.3**	302
	GEC 2	27.5	14.3	21.2	23.0	55.6	83.1	82.3	82.7	97.1	343
	GEC 3	25.6	8.3	18.1	18.8	53.3	82.4	82.2	82.3	94.2	347

Table 4: Results of quantitative evaluation. Gramm. denotes the grammaticality score.

produced by our method in order to control for variables, which was equally more effective compared to the baseline method, proving that the data produced by our method is well suited for the task SentRev.

In addition, experiment Group 1 also shows that the grammaticality score achieves competitive results even though we do not use the baseline data to fine-tune the GEC models. This is partly due to the fact that our chosen GEC models already handle grammatical errors well, and partly due to the fact that the data generated by our method also tends to generate a good number of pairs of grammatical errors, which helps to improve the GEC performance of the models, confirming our previous conjecture.

Finally, due to the limited proportion of Lexical Bundles in sentences, our method generates drafts and references with comparatively minor modifications. In contrast, the baseline method allows for extensive modifications of the reference, but deviates significantly from the authentic English expressions made by non-native English speakers. Hence, experimental results demonstrate that the synthetic data generated by our method, in combination with the baseline approach, can complement each other's shortcomings, leading to superior outcomes. Additionally, the experimental results robustly validate the effectiveness of Lexical Bundles in enhancing English fluency.

5 Conclusion

In this study, we propose a Data Augmentation method for SentRev using Back-Translation of lexical bundles. We demonstrate the effectiveness of our method in conjunction with a baseline approach through fine-tuning three GEC models, resulting in significant improvements in multiple metrics for the SentRev task. Our findings confirm the importance of lexical bundles for academic writing. On the other hand, our experiments have also demonstrated that the utilization of lexical bundles, as a constituent of enhancing English fluency, yields optimal outcomes when combined with other methods.

6 Future work

As a larger unit, lexical bundles, although effective in enhancing fluency in English, often exhibit discontinuity. Consequently, the approach employed in this study is unable to handle those discontinuous lexical bundles. Thus, future research will prioritize the investigation of deeper syntactic dependency relationships. Furthermore, our Lexical bundle dictionary still has limited coverage, and the extraction of high-quality, discontinuous lexical bundles poses a challenging task for future endeavors.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Douglas Biber and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. *Language and Computers*, 26:181–190.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014.

Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24, Baltimore, Maryland. Association for Computational Linguistics.

Chooi Ling Goh and Yves Lepage. 2019. Extraction of lexical bundles used in natural language processing articles. In *2019 International Conference on Advanced Computer Science and information Systems (ICACSIS)*, pages 223–228.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1):4–21.

Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. 2018. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63, Melbourne, Australia. Association for Computational Linguistics.

Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the rough: Generating fluent sentences from early-stage drafts for academic writing assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 40–53, Tokyo, Japan. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33, Baltimore, Maryland. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

M Lewis. 1993. The lexical approach: The state of elt and the way forward. england.

Mahdi Naser Moghadasi and Yu Zhuang. 2020. Sent2vec: A new sentence embedding representation with sentimental semantic. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4672–4680.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

James R Nattinger and Jeanette S DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford University Press.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

Henry Lee Smith Jr. 1958. Linguistics across cultures: Applied linguistics for language teachers.

Felix Stahlberg and Shankar Kumar. 2020. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dave Willis. 1990. *The lexical syllabus*, volume 30. London: Collins.

Alison Wray. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied linguistics*, 21(4):463–489.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 619–628, New Orleans, Louisiana. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota. Association for Computational Linguistics.