ADVERB IS THE KEY: SIMPLE TEXT DATA AUGMENTATION WITH ADVERB DELETION

Juhwan Choi and Youngbin Kim
Chung-Ang University, Seoul, Republic of Korea
{gold5230, ybkim85}@cau.ac.kr

ABSTRACT

In the field of text data augmentation, rule-based methods are widely adopted for real-world applications owing to their cost-efficiency. However, conventional rule-based approaches suffer from the possibility of losing the original semantics of the given text. We propose a novel text data augmentation strategy that avoids such phenomena through a straightforward deletion of adverbs, which play a subsidiary role in the sentence. Our comprehensive experiments demonstrate the efficiency and effectiveness of our proposed approach for not just single text classification, but also natural language inference that requires semantic preservation. We publicly released our source code for reproducibility.

1 INTRODUCTION

Text data augmentation is an important regularization technique to alleviate overfitting and improve the robustness of an NLP model. Even though various strategies have been proposed, text data augmentation methods often involve a trade-off between complexity and potential performance gain (Feng et al., 2021). While rule-based methods (Zhang et al., 2015; Belinkov & Bisk, 2018; Wei & Zou, 2019; Karimi et al., 2021; Choi et al., 2023) are simple and easy to implement, they often introduce small diversity, which suffers to enhance performance significantly (Zhang & Ma, 2022). Whereas, relatively complex methods that utilize deep learning models (Sennrich et al., 2016; Wu et al., 2019; Anaby-Tavor et al., 2020) may acquire different expressions, which leads to potentially large performance improvement. However, deploying them for real-world applications can be expensive, as they often necessitate additional deep learning models. This becomes particularly restrictive when hardware resources are limited, as is often the case for projects in developing countries.

Consequently, rule-based approaches are widely employed for real-world applications owing to their low cost. However, unlike rule-based image augmentation such as flipping or cropping, which maintains the original contents, rule-based text data augmentation methods have another challenge of semantic preservation. The conventional approach of rule-based text augmentation methods relies on introducing a perturbation through predefined rules, such as synonym replacement, random insertion, random swap, and random deletion (Wei & Zou, 2019). While previous researchers have recognized this drawback and proposed several refinements, they tend to introduce less variation (Karimi et al., 2021) or not fully maintain the semantics of the original sentence (Choi et al., 2023).

In light of these previous studies, we propose a straightforward yet novel rule-based text augmentation method by deleting adverbs from the given dataset. Adverbs mostly focus on adjusting another word, by maximizing or diminishing the meaning of other words (Delfitto, 2006; Ruppenhofer et al., 2015). Through the explicit removal of these adverbs and keeping more important words such as nouns and verbs, we can attain new sentences while maintaining the core semantics of the original sentence.

To the best of our knowledge, this work is the first study to actively focus on the role of adverbs, which has often taken less interest (Nikolaev et al., 2023), for text data augmentation. We evaluated our approach on various text classification tasks, as well as natural language inference (NLI) tasks, which are more complex compared to single text classification. Our empirical experimental result showed the effectiveness and efficiency of the proposed method, especially on NLI tasks, which require further semantic preservation compared to single text classification.

2 METHOD

We aim to synthesize the given data $\boldsymbol{x} = \{w_1, w_2, ..., w_n\}$ that comprises the dataset $\mathcal{D}$ and generate $\hat{\boldsymbol{x}}$, the modified version of $\boldsymbol{x}$ following the predefined rule. Previous methods such as easy data augmentation (EDA) and an easier data augmentation (AEDA) have applied this rule on a random word, regardless of the importance of each word. However, the proposed method aims to remove only adverbs from the given sentence. This process can be formulated as follows:

$$\hat{\boldsymbol{x}} = \boldsymbol{x} \setminus \{W_{adv}\},\ where\ \{W_{adv}\} = \texttt{POS_Tagger}(\boldsymbol{x}, \text{ADV})$$

In this formulation, $\{W_{adv}\}$ denotes a set of words that are tagged as adverbs by a part-of-speech (POS) tagger. We skip $\boldsymbol{x}$ that does not have at least one adverb from the augmentation process.

3 EXPERIMENT

Table 1: Accuracy (%) and performance gain (%p) across eight datasets. The best values for each dataset are boldfaced. Results that reported a lower value than the baseline are colored in gray.

BERT	SST2	SST5	CoLA	TREC	RTE	MNLI-M	MNLI-MM	QNLI
No Aug	88.74	50.58	79.24	95.08	64.72	75.35	77.17	85.81
EDA	89.18	50.36	76.80	95.27	62.43	74.51	75.85	83.17
AEDA	89.41	50.63	79.45	95.39	65.92	74.60	77.35	86.55
softEDA	89.24	50.89	76.32	95.86	65.41	74.06	75.93	84.97
Ours	**89.73**	51.25	80.28	96.48	65.58	76.11	77.68	87.07
Ours w/ Curr.	89.52	**52.37**	**83.15**	**96.95**	**68.72**	**77.97**	**78.14**	**87.21**

To evaluate the effectiveness of the proposed method, we conducted the experiment on various text classification and NLI datasets, in comparison with previous rule-based text augmentation methods. We built each model using BERT (Devlin et al., 2019) as the baseline model. For comparison with previous methods, we adopted EDA (Wei & Zou, 2019), AEDA (Karimi et al., 2021), and softEDA (Choi et al., 2023), a method that compensates semantic damage through soft label. More details about implementation details and datasets used for experiment can be found in Appendix A and C.

Table 1 demonstrates the experimental result. The results indicate that the proposed method is able to enhance the performance of each model in various tasks, including NLI, where previous methods mostly suffered from performance degradation. This is due to the complexity of NLI tasks, which require a proper connection between two sentences after the augmentation, prohibiting semantic damage after the augmentation. These results clearly showcase the superiority of the proposed method in terms of preventing semantic damage while providing simple and low-cost augmentation technique. Moreover, our additional experiment that combines our method with curriculum data augmentation (Wei et al., 2021; Ye et al., 2021; Lu & Lam, 2023) reveals further performance gain.

4 CONCLUSION

We proposed a novel rule-based text data augmentation method focusing on adverbs. Through this straightforward approach, it is able to easily preserve the original meaning of the given sentence compared to previous rule-based text augmentation methods. This advantage led to performance gain on not just text classification tasks but also NLI tasks, where other rule-based methods suffer to enhance the performance. Future research could expand our method to other tasks, including text summarization. Additionally, we are planning to extend this approach to other languages, based on the universality of adverbs (Delfitto, 2006).

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1C1C1008534), and Institute for Information & communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program, Chung-Ang University).

URM STATEMENT

First author Juhwan Choi meets the URM criteria of ICLR 2024 Tiny Papers Track. He is outside the range of 30-50 years, non-white researcher.

REFERENCES

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7383–7390, 2020.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018.

Juhwan Choi, Kyohoon Jin, Junho Lee, Sangmin Song, and YoungBin Kim. Softeda: Rethinking rule-based data augmentation with soft labels. In *ICLR 2023 Tiny Papers*, 2023.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190, 2005.

Denis Delfitto. Adverb classes and adverb placement. *The Blackwell companion to syntax*, pp. 83–120, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pp. 4171–4186, 2019.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988, 2021.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. AEDA: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2748–2754, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, pp. 1–7, 2002.

Hongyuan Lu and Wai Lam. Pcc: Paraphrasing with bottom-k sampling and cyclic learning for curriculum data augmentation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 68–82, 2023.

Dmitry Nikolaev, Collin F Baker, Miriam RL Petruck, and Sebastian Padó. Adverbs, surprisingly. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pp. 512–526, 2023.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Josef Ruppenhofer, Jasper Brandes, Petra Steiner, and Michael Wiegand. Ordering adverbs by their scaling effect on adjective intensity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 545–554, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, 2016.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6382–6388, 2019.

Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5493–5500, 2021.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional bert contextual augmentation. In *ICCS 2019: 19th International Conference*, pp. 84–95, 2019.

Seonghyeon Ye, Jiseon Kim, and Alice Oh. Efficient contrastive learning via novel data augmentation and curriculum learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1832–1838, 2021.

Linfeng Zhang and Kaisheng Ma. A good data augmentation policy is not all you need: A multi-task learning perspective. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, volume 28, 2015.

A IMPLEMENTATION DETAILS

This section describes experimental setups and implementation details for reproduction. We have mainly followed the source code provided by softEDA (Choi et al., 2023). Please refer to our source code for further investigation.[1]

Augmentation Technique. We used a POS tagger offered by spaCy (Honnibal et al., 2020) library, with `en_core_web_sm` model. We removed all words that are tagged as `ADV` from each sentence of a dataset. We skipped a sentence with no adverb from the augmentation process. For softEDA baseline, we used $\alpha = 0.2$ for every experiment.

For curriculum data augmentation, we utilized two-stage curriculum learning suggested by previous researchers (Wei et al., 2021). Specifically, we first trained the initial model using only original data for 2 epochs and embraced augmented data for the rest 3 epochs of the training procedure.

Model and Dataset. The model and datasets were mainly implemented with Hugging Face Transformers (Wolf et al., 2020) and Datasets (Lhoest et al., 2021) libraries. The BERT model has used `bert-base-uncased`. We note that the NLI datasets used in our experiments are drawn from the GLUE (Wang et al., 2019) benchmark and the test set is not publicly available. Instead, we used the validation set as the test set and N_{Test} denotes the amount of validation data. Similar to datasets that do not have a predefined validation set, we used randomly selected 20% of the training set as the validation set for these datasets.

Hyperparameters. Adam (Kingma & Ba, 2015) has been deployed as the optimizer, with a learning rate of 1e-4. We trained each model for 5 epochs with a batch size of 32.

Further Details. In cases where no predefined validation set exists, we randomly selected 20% of the training data as the validation set. The training procedure was performed with a single NVIDIA RTX 3090 GPU.

B CASE ANALYSIS AND DISCUSSION

Table 2 displays examples of augmentation results using the traditional EDA technique and our proposed method. The examples show that our method generates more syntactically acceptable, and semantically consistent sentences compared to EDA. This superiority comes from the randomness of EDA techniques, such as deleting important words or swapping the order within the sentence. Whereas, our proposed method could generate better samples through targeted deletion of adverbs, which play supplementary roles in the sentence.

However, while our method has a distinct advantage, we acknowledge an existing limitation. As our method is solely composed of the deletion of adverbs, the scope of variation is relatively small. This can be interpreted as a trade-off between semantic consistency and diverse augmentation. Nonetheless, the variation is larger than that of AEDA, which only injects punctuation marks, as our method aims to delete several words from the original sentence. Furthermore, in a real-world scenario, the engineer could combine our method with previous methods to acquire further performance boost. Future work could extend our approach to ensure more variability of augmented samples.

[1] `https://github.com/c-juhwan/adverb-deletion-aug`

Table 2: Comparison between EDA and our proposed method on examples from the SST2 dataset.

Original	EDA	Ours
The film is strictly routine.	The is strictly film routine.	The film is routine.
The santa clause 2 proves itself a more streamlined and thought out encounter than the original could ever have hoped to be.	The santa clause 2 itself a streamlined and original thought out encounter than the could have to be.	The santa clause 2 proves itself a streamlined and thought out encounter than the original could have hoped to be.
This is a very ambitious project fora fairly inexperienced filmmaker, but good actors, good poetry and good music help sustain it.	This is a very ambitious for a fairly inexperienced actors, but filmmaker good, good poetryand good music sustain it.	This is a ambitious project for a inexperienced filmmaker, but good actors, good poetry and good music help sustain it.
Perhaps the best sports movie i've ever seen.	Best perhaps movie seen.	the best sports movie i've seen.

C DATASET SPECIFICATIONS

Table 3: Dataset used for the experiment. The datasets with $*$ indicate that these datasets used the validation set as the test set, as stated in Appendix A.

Dataset	Task	N_{Class}	N_{Train}	N_{Test}
SST2 (Socher et al., 2013)	Sentiment	2	6,919	1,820
SST5 (Socher et al., 2013)	Sentiment	5	8,544	2,210
TREC (Li & Roth, 2002)	Question Type	6	5,452	500
CoLA (Warstadt et al., 2019)	Acceptability	2	8,551	527
RTE* (Dagan et al., 2005)	NLI	2	2,490	277
MNLI-matched* (Williams et al., 2018)	NLI	3	392,702	9,815
MNLI-mismatched* (Williams et al., 2018)	NLI	3	392,702	9,832
QNLI* (Rajpurkar et al., 2016)	NLI	2	104,743	5,463