UdS 2024: Final project for Computational Linguistics course "POS-Aware Data Augmentation for NLP Tasks"

Konstantin Chernyshev
kdchernyshev@gmail.com

Abstract

The challenge of achieving high performance of the NLP models in a limited resources environment is a pressing never-ending issue. In this study, we investigate Part-Of-Speech-aware Augmentation methods and compare them to random augmentation techniques on 4 tasks of SUPER GLUE benchmark with ROBERTA-BASE model. Our experiments on BoolQ, CB, WiC, and RTE tasks show that POS-aware augmentation can enhance model performance, and provide a more stable fine-tuning process than random augmentation. Source code available at GitHub[1].
We also introduce the FAST-AUG[2] library for faster data augmentation.

1 Introduction

Natural Language Processing (NLP) aims to enable machines to understand, interpret, and generate human language. Benchmarks like Super GLUE (Wang et al., 2019a) help measure progress in the field by proposing diverse and complex tasks. Despite recent advancements in NLP models, their performance on certain nuanced language tasks remains limited, especially when computational resources are constrained. Data augmentation emerges as a solution to enhance model robustness and performance.
We explore the effects of Part-of-Speech (POS)-aware data augmentation on the roberta-base model's performance across four Super GLUE tasks: BoolQ, CB, WiC, and RTE.

Our research questions and objectives revolve as

- **RQ1:** How does POS-aware data augmentation compare to random data augmentation?
- **RQ2:** How does POS-aware data augmentation affect the model's performance across different Super GLUE tasks?

[1]https://github.com/k4black/uds-2024-coli
[2]https://github.com/k4black/fast-aug

We aim to answer these questions by experimenting with various data augmentation techniques and assessing their impact on the model's performance. First, we introduce the related work in Section 2. Then, we describe the methodology in Section 3. We present the experiments results in Section 4 and discuss the findings in Section 5 and Section 6. Finally, we conclude the study in Section 7.

2 Related Work

Recent advancements in NLP have underscored the importance of robust data augmentation techniques for enhancing model performance, especially in scenarios with limited data availability. The exploration of data augmentation strategies has revealed diverse methodologies aimed at enriching the linguistic variety of datasets, thereby improving the generalization capabilities of NLP models. Among these, POS-aware (Part-of-Speech-aware) data augmentation holds promise due to its focus on preserving syntactic structures while introducing meaningful variations.

The significance of data augmentation goes beyond merely expanding the dataset size. A comprehensive survey by Feng et al. (2021) categorizes augmentation techniques, underlining the importance of syntactic and semantic consistency in transformations, mentioning POS-aware augmentation strategies as doing so.

Close, but not directly connected work of Şahin (2022), dives into syntactical, token-level, and character-level augmentations for part-of-speech tagging, dependency parsing, and role labeling.

A newer empirical survey (the only one widely cited) by Chen et al. (2023) touches upon a wide array of augmentation methods, including POS-aware techniques. However, the study stops short of a deep dive into these methods, signaling an area ripe for further exploration.

In the context of narrow data augmentation techniques, the EDA (Easy Data Augmentation)

method introduced by Wei and Zou (2019) stands out as the most popular one. By employing simple strategies: synonym replacement, random insertion, random swap, and random deletion methods, they demonstrated that using a fraction of the training data could match full-dataset training performance. It underscores the potential efficiency gains achievable through thoughtful augmentation strategies.

Building on EDA's foundations, Karimi et al. (2021) introduced AEDA (An Easier Data Augmentation), a streamlined approach focusing exclusively on the insertion of punctuation marks. Their findings suggest that even minimalistic augmentations can significantly enhance text classification tasks, and show superior performance compared to using the EDA-augmented data in all five datasets.

While POS-aware augmentation techniques appear to be understudied, they represent an area with considerable untapped potential. By working in this direction, we hope to contribute a nuanced understanding of this emergent field of more linguistically informed and robust NLP models.

3 Methodology

In this section, we describe the methodology used to conduct the experiments, including data, models, and augmentation techniques.

3.1 Dataset and Metrics

Dataset	Avg.Words	#Train	#Val.
BoolQ	106.3	9427	3270
CB	62.6	250	56
WiC	14.4	5428	638
RTE	52.4	2490	277

Table 1: Dataset Overview: average word count in the training split and the number of samples in the training and validation sets.

We use the Super GLUE benchmark (Wang et al., 2019a) to evaluate the performance of the roberta-base model. The benchmark consists of eight tasks, each designed to test different aspects of language understanding. This benchmark is a harder version of the GLUE benchmark (Wang et al., 2019b) and is designed to be more challenging for the models.

As we have limited resources, we focus on four Super GLUE tasks - having less than 10k samples, having different tasks, and various dataset sizes. Table 1 shows the statistics of the datasets used in the experiments.

- BoolQ (Clark et al., 2019) task targets the ability of models to answer yes/no questions based on Wikipedia passages, formed as a binary classification task. This task requires text 'understating' and factual correctness on the model.
- CB (De Marneffe et al., 2019) focuses on natural language inference tasks and commitment bank statements. The task objective is to classify a given pair of sentences into entailment, contradiction, or neutral categories. The main challenge is the lack of training data.
- WiC (Pilehvar and Camacho-Collados, 2019) is a benchmark for the evaluation of context-sensitive word embeddings. It is a classification task to identify if the occurrences of a given word in the two contexts correspond to the same meaning or not. This is a complex, semantically-oriented task, that requires deep understanding.
- RTE (Bentivogli et al., 2009) is a dataset that comes from a series of textual entailment challenges. Examples are constructed based on news and Wikipedia. Entailment detection is one of the base tasks to evaluate models' language understanding capacity.

As the Super GLUE benchmark does not provide the test set for offline validation we are report validation set results in all of our experiments.

The dataset splits and corresponding examples were used as is. No additional pre-processing was used, except for model-specific tokenization.

3.2 Model

We use the ROBERTA-BASE model (Liu et al., 2019) as our base model. It is a transformer-based model that has been pre-trained on a large corpus of text data. The model has 12 layers, 768 hidden units, and 12 attention heads. This model was chosen as well-studied and widely used in the NLP community, additionally, the roberta model has Super GLUE results. The base version of the model is relatively small and can be fine-tuned on a single GPU.

The roberta-base model builds upon the BERT model with architectural improvements like increased batch size and longer training. Its pre-training process involves masked language modeling, which facilitates fine-tuning for specific tasks. The impact of data augmentation on the fine-tuning process can lead to improved model performance.

Aug.Type	Before	After
sub	Hello, little cute world!	Hello, **dog** cute **Me**!
swap	Hello, little cute world!	**little**, **Hello** cute world!
pos-sub	Hello$_{UH}$, little$_{JJ}$ cute$_{JJ}$ world$_{NN}$!	Hello$_{UH}$, **big**$_{JJ}$ cute$_{JJ}$ **dog**$_{NN}$!
pos-swap	Hello$_{UH}$, little$_{JJ}$ cute$_{JJ}$ world$_{NN}$!	Hello$_{UH}$, **cute**$_{JJ}$ **little**$_{JJ}$ world$_{NN}$!

Table 2: Augmentation type example of transformation applied. Random Substitution, Random Swap, POS-Aware Substitution, and POS-Aware Swapping.

3.3 Augmentation

In our augmentation process, we utilize a concept known as Part-Of-Speech (POS) tagging. POS tagging is the task of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

We hypothesize that POS-aware data augmentation, which preserves grammatical structure, could enhance model performance. Our study compares traditional augmentation methods with POS-aware techniques, specifically focusing on word substitution and swapping strategies. These methods aim to maintain or improve the semantic integrity of augmented sentences.

- Random Words Substitute: This technique involved randomly selecting words within text and replacing them with words from a predefined list, obtained from the training set. Note, that overall sentence meaning can change.
- Random Words Swap: This method required selecting and swapping words within the text. Similarly, overall sentence meaning can change.
- POS-Aware Substituting: This approach involved selecting replacement words that matched the POS of the original word, leading to more meaningful substitutions.
- POS-Aware Swapping: This technique identified swap candidates with matching POS tags.

Table 2 illustrates the different augmentation types and their corresponding transformations.

To ensure the consistency and effectiveness of our augmentation, we introduce a simplification step, where certain POS tags mapped to their simplified forms. E.g. different forms of adjectives ("JJR", "JJS") are all mapped to "JJ". This simplification allows us to treat different forms of the same part of speech in a uniform manner, increasing the likelihood of successful augmentation (that we have at least 2 words with the same POS tag to swap).

Similarly, we limit set of POS tags, defined as Adjectives, Adverbs, Interjections, Determiners, Pronouns and Modals - to select which words in the text are eligible for augmentation.

4 Experiments

4.1 Setup

For the experiments, we followed a systematic approach to ensure consistency and reliability. Hugging Face's Transformers library (Wolf et al., 2020) and PyTorch were used as main frameworks. For model checkpoints Hugginface Hub ROBERTA-BASE[3] checkpoints were used. Similarly, for Super GLUE datasets, Hugging Face's SUPER-GLUE dataset[4] was used.

For random data augmentation, we developed a custom library FAST-AUG[5] provided Random Words Substitute, Random Words Swap augmenters. For POS tags obtaining we used POS fine-tuned BERT-BASE-MULTILINGUAL[6] (Sajjad et al., 2022) checkpoints from Huggingface Hub, with reported F1 score on Penn Tree Bank of 96.69.

4.2 Finetuning Experiments

The fine-tuning process is crucial for adapting the model to our tasks. First of all, we need to select hyperparameters such as learning rates and batch sizes, as well as the number of epochs.
This selection was guided by both the established practices within the community and a series of preliminary experiments designed to gauge their impact on model performance within our computational limits.

Important note, the Super GLUE benchmark does not provide a test set for offline validation, so we are using and reporting validation set results

[3]https://huggingface.co/roberta-base
[4]https://huggingface.co/datasets/super$_glue$
[5]https://github.com/k4black/fast-aug
[6]https://huggingface.co/QCRI/bert-base-multilingual-cased-pos-english

LR	BoolQ Acc.	CB F1/Acc.	WiC Acc.	RTE Acc.
GLUE	87.1	90.5 / 95.2	69.9	88.2
1e-4	62.2	**77.4 / 85.7**	-	47.3
5e-5	**79.5**	56.2 / 80.4	**66.1**	**71.5**
1e-5	78.8	47.4 / 67.9	65.1	71.1
5e-6	72.1	19.4 / 41.1	65.2	67.9

Validation scores by Learning Rate. Batch size fixed on 32.

BS	BoolQ Acc.	CB F1/Acc.	WiC Acc.	RTE Acc.
GLUE	87.1	90.5 / 95.2	69.9	88.2
8	62.2	53.4 / 76.8	-	-
16	76.1	**93.1 / 94.6**	**66.8**	**72.2**
32	**79.5**	77.4 / 85.7	66.1	71.5
64	OOM	53.5 / 76.8	66.6	71.8

Validation scores by Batch Size. The learning rate is fixed on 1e-4 for CB and 5e-5 for all other tasks.

Table 3: Validation scores for ROBERTA-BASE model, fine-tuned 5 epochs on selected tasks, first varying by Learning Rate (left) than Batch Size (right). OOM stands for Out Of Memory. The best results are highlighted in bold. "GLUE" refers to ROBERTA-LARGE **test** scores from the SUPER-GLUE leaderboard. Dash means the model was not converged resulting prediction of one class (e.g. 50%).

Dataset Name	Learning Rate	Batch Size	Weight Decay	Warmup Ratio	Max Epochs
BoolQ	5e-5	32	0.01	0.1	10
CB	1e-4	16	0.01	0.1	10
WiC	5e-5	16	0.01	0.1	10
RTE	5e-5	16	0.01	0.1	10

Table 4: Resulting parameters of roberta-base model fine-tuning for Augmentation experiments on 4 selected tasks.

in all of our experiments and provide Super GLUE results for rough comparison.

Learning Rate Optimization: Starting with the most important hyperparameter, we investigate values around community-recommended values 1e-5 – 5e-5, while keeping the batch size constant at 32 across 5 epochs. This approach allowed us to identify the learning rate that facilitated the best balance between training speed and model accuracy. The left side of Table 3 shows obtained values, indicating the best learning rate of 5e-5 for all of the tasks except CB, where 1e-4 is shown to be the best one.

Batch Size Variation: After determining the optimal learning rate, we shifted our focus to the batch size. The batch size can significantly affect the model's generalization ability and training dynamics, especially for smaller datasets. The batch size variations are shown on the Right side of Table 3, indicating 16 samples for all tasks except for the BoolQ task being 32.

Epoch Consideration: The number of epochs, or complete passes through the training dataset, was set at 5 for these experiments. This decision was based on community experience recommending a 5-10 range for such model fine-tuning. The train/validation loss charts show saturation with small room for improvement at the 5th epoch. For Augmentation experiments, 10 epochs with Early Stopping are used.

By methodically adjusting these hyperparameters and analyzing their effects, we aimed to optimize the roberta-base model for our specific NLP tasks. The final parameters to be used in further experiments can be found in Table 4.

4.3 Augmentation Experiments

In this section, we present the results of our experiments with different augmentation techniques. We fine-tuned the roberta-base model on the BoolQ, CB, WiC, and RTE tasks using the selected augmentation methods described in Section 3. All methods were applied in online mode, meaning that augmentation was applied during the training process, so each epoch model was trained on different augmented data.

For each augmentation technique, we varied the augmentation probability between 1%, 5%, 10%, and 20% to test the effectiveness of a specific augmentation technique. In each experiment, we fine-tuned the ROBERTA-BASE model on the augmented data using parameters mentioned in Table 4.

The results are presented in Table 5.

Aug.Type	Aug.Prob.	BoolQ Acc.	CB F1/Acc.	WiC Acc.	RTE Acc.
none		79.9	93.1 / 94.6	68.0	74.7
sub	1%	79.2	93.0 / 94.6	-	-
	5%	77.8	**95.0 / 96.4**	-	-
	10%	77.6	64.1 / 82.1	-	-
	20%	75.8	74.3 / 85.7	61.0	-
swap	1%	80.4	90.4 / 91.1	-	74.4
	5%	79.4	83.6 / 87.5	68.0	71.8
	10%	79.1	86.2 / 89.3	-	75.5
	20%	76.0	79.9 / 85.7	69.4	-
pos-sub	1%	79.9	79.7 / 83.9	-	74.7
	5%	80.1	78.9 / 85.7	-	74.0
	10%	80.3	89.2 / 91.1	70.2	**75.8**
	20%	78.8	87.8 / 89.3	**70.5**	74.0
pos-swap	1%	**80.8**	79.7 / 83.9	-	74.7
	5%	80.3	87.8 / 89.3	-	68.6
	10%	79.6	79.6 / 83.9	69.8	-
	20%	78.8	55.0 / 78.6	68.0	-

Table 5: Augmentation type and word selection probability. 10 epochs. The best results are highlighted in bold. "GLUE" refers to ROBERTA-LARGE **test** scores from the SUPER-GLUE leaderboard. Dash means the model was not converged resulting prediction of one class (e.g. 50%).

5 Results and Discussion

Our experiments encompassed a variety of data augmentation techniques, including POS-aware augmentations, applied to four datasets from the Super GLUE benchmark: BoolQ, CB, WiC, and RTE. The performance of the roberta-base model was evaluated under each augmentation method, and the results were compared to those of the model with pure fine-tuning.
The augmentation techniques under consideration included random word substitution, random word swapping, POS-aware substitution, and POS-aware swapping. The latter two methods preserved the POS tags of the original words, ensuring syntactic consistency in the augmented data.

Our findings indicate that data augmentation generally improved the model's performance across all tasks. More specifically, POS-aware augmentation techniques outperformed random augmentation methods for all tasks except CB. This exception could be attributed to the small size of the CB dataset, which might have led to score fluctuations.

Interestingly, the WiC and RTE tasks, which were more challenging and required more careful parameter selection, showed better results with POS-aware augmentation techniques. For random augmentation methods, these tasks tend to fall to not converged - produce single class predictions. In particular, the WiC task, which involves short sentences and relies on the meaning of individual words, 'suffered less' from POS-aware augmentation techniques.

Comparing sub/pos-sub and swap/pos-swap methods side-to-side, we can see that POS-aware augmentation techniques are more effective in maintaining or improving the model's performance. This is likely due to the preservation of syntactic structures and semantic consistency in the augmented data.

However, it is important to note that the effectiveness of data augmentation is task-dependent and requires careful selection of augmentation parameters. While POS-aware augmentation techniques showed promising results in our experiments, further research is needed to validate these findings across a broader range of tasks and datasets.

6 Limitations and Future Work

The primary limitation of this study is the absence of a test set for the Super GLUE tasks. This makes it difficult to evaluate the model's performance

on unseen data. Additionally, computational constraints limited the model size and considered tasks to four Super GLUE tasks. Similarly, the POS-tagging model used for POS-aware augmentation techniques was reported to have a 96.69 F1 score on Penn Tree Bank, which is not the best score and could affect the augmentation quality.

For future research, we suggest investigating other augmentation strategies and benchmarks, as well as further refining existing POS-aware techniques to maximize their impact on model performance. Furthermore, the official roberta-large Super GLUE submission indicates usage of MNLI-fine-tuned (Williams et al., 2018) model for RTE and CB tasks, which should be addressed in experiments. Additionally, exploring the impact of data augmentation on generative models could provide valuable insights into enhancing the robustness and generalization capabilities of NLP models.

7 Conclusion

In this study, we investigated the impact of Part-Of-Speech (POS)-aware data augmentation on the performance of the roberta-base model across four Super GLUE tasks. Our experiments showed that POS-aware augmentation techniques outperformed random augmentation methods, and introduced a more stable training process on challenging tasks like WiC and RTE. These findings suggest that syntactic consistency and semantic integrity in augmented data can improve model performance. However, further research is needed to validate these findings across a wider range of tasks and datasets.

Disclosure

ChatGPT and Grammarly were used to fix the grammar and style of the paper. Overall, ChatGPT was not helpful for factual and structural corrections, as it tends to provide very generic and often incorrect paragraphs. Similarly, most of the grammar and style corrections were done manually or with Grammarly, as ChatGPT provides extremely 'high' and poetic English, which is far from my writing style =)

P.S. 'We' in paper was used as style element. The paper was written alone.

References

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT 2019*.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. To appear in proceedings of Sinn und Bedeutung 23. Data can be found at https://github.com/mcdm/CommitmentBank/.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Aeda: an easier data augmentation technique for text classification. *arXiv preprint arXiv:2108.13230*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*.

Gözde Gül Şahin. 2022. To augment or not to augment? a comparative study on text augmentation techniques for low-resource nlp. *Computational Linguistics*, 48(1):5–42.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Rafae Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. In *North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL)*, NAACL '22, Seattle.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b.

GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.