Improving Readability in Extractive Summarization Procedures

Varun Dashora and Marcus Manos

9 March 2022

1 Abstract

Extractive text summarization can be affected by several preprocessing steps including stripping grammar, casing, tokenization, and data masking. One of the more unexplored factors is data masking; the masking of select tokens forces the model to predict the hidden word and has the potential to enhance the readability of extractive summarization. The State of the Art approach is to randomly mask words in the corpus, which is good for imitating the distribution of words in the original text but not necessarily for creating a summary of the text. By analyzing the summaries from similar pieces of text we were able to randomly mask words by examining their parts of speech and assigning them a probability proportional to their distribution in the summary. We tested our theory out on the "sshleifer/distilbart-xsum-12-6" model developed by Sam Shleifer using the BBC News data set to fine-tune the model and validate our assumptions.

2 Introduction

The goal of extractive text summarization is to accurately summarize the text by splicing together sentences that are the most relevant to the overall message. This can be done via several methods, the current state-of-the-art approach being the use of transformer models and their self-attention layers to find the most important segments of the text to include in the summary. These models take advantage of a technique called masking; hiding select words so the transformer can predict them from the context to the left and right of the masked word, to focus on specific words. The current approach focuses on implementing random masking uniformly across the dataset, meaning that all words will have an equal chance of being focused on by the self-attention mechanism (Song et al., 2019). This means that the distribution of masked words will ultimately follow the distribution of parts of speech in the text; if the text is made up of 60% nouns then around 60% of the masked word will be nouns. This process creates summaries that contain relevant information, at the expense of the quality of writing. Since this technique relies on extracting sentences from the summaries it can be hard to understand, rendering it useless for many business applications. It is for this reason that we wanted to improve the readability of extractive text summarization while retaining its relevant information. It is for this reason that we examined the impact of masking different parts of speech, choosing to emulate the distribution of parts of speech in the summary.

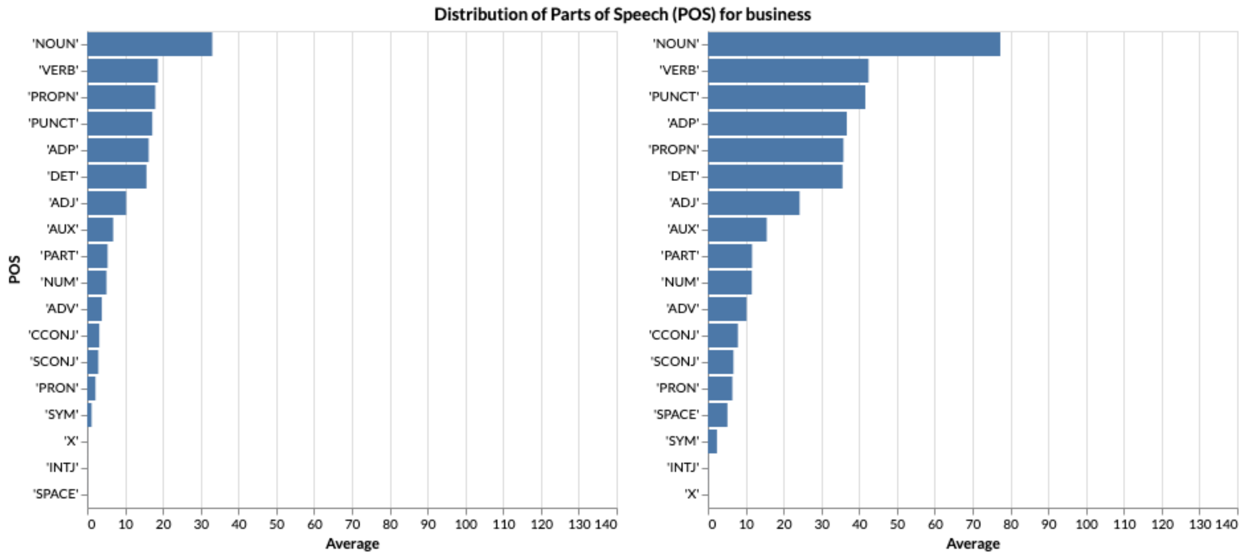
3 Background

There have been several attempts to address extractive summarization with varying model architectures. First, there is a modification of BERT called BERTSum that has been fine-tuned specifically for extractive text summarization [1]. In addition, researchers have been exploring how to use the article's natural organization and hierarchy with an encoder-only model [2]. A third attempt at improving extractive summarization involves not a neural network, but a generic multi-iteration computation that identifies relationships between sentences to avoid redundancy in the final summarization [3]. A fourth attempt evaluated the effect of masking length on masked language models [4]. This study analyzed how different masking lengths affected the ROGUE scores of masked language models using a variety of masking techniques ranging from span to entity-level masking. In addition to model architectures, there is an evaluation metric very useful for gauging readability. This is the GRUEN metric, which assesses a text for grammatical correctness, coherency, and focus [5].

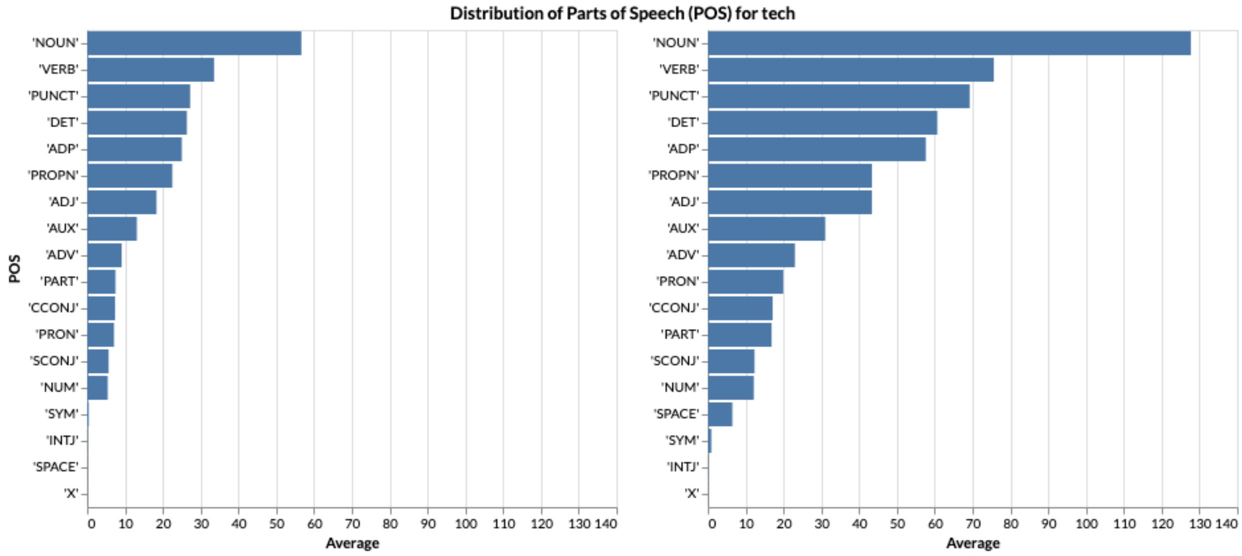
4 Methods

Typically, words are masked randomly across the dataset, which results in roughly the same proportion of words being masked that make up the text. This method may not necessarily be the best as the makeup of a summary is often different from the makeup of the text. We began our process by examining the distribution of the parts of speech for 100 summaries across each of the five categories in the BBC news dataset. We validated our assumptions that the distribution of parts of speech was indeed different in summaries versus articles, and discovered the distributions varied depending on the category of article.

Since we believed the main regarding surrounding readability was the BERT model focusing on the masked words in proportion to the article rather than the summary we implemented masking according to the average word distribution for each article category. To do this we used a spacy model trained on the ”en_core_web_md” corpus and predicted the probability of masking according to the distribution of parts of speech that made up the summary (if adjectives accounted for 8% of the words in the summary, then the probability of masking a word if it was an adjective would be 8%). To examine the difference we fine-tuned an existing BERT model (Distil Bert, pre-trained on the xsum dataset) with a randomly masked dataset with a probability of .20, and one that was masked according to the summary. Since our dataset was limited in size we randomly selected 30 groups of 100 articles for testing and 10 groups of 30 articles for fine-tuning. We examined the variance between the groups and found that it was minimal, validating our results further. To analyze the readability of our results we used a series of ROGUE scores, and the GRUEN metric developed by Zhu and Bhat, which examines the linguistic quality of the generated text. The GRUEN metric evaluates the Grammar, non-Redundancy, focUs structure, and coherENce, and was an important measure as it aims to echo what a human would do if they looked at the generated summaries and tried to score it on readability.



Distribution of Parts of Speech (POS) for business



Distribution of Parts of Speech (POS) for tech

5 Results and Discussion

5.1 Model Results

l

Model	Business	Entertainment	Politics	Sports	Tech
Fine-Tuned DistillBART	0.773	0.859	0.801	0.840	0.838
2	7	78	5415		
3	545	778	7507		
4	545	18744	7560		
5	88	788	6344		

5.2 Part-of-Speech Exploration

This is another piece of placeholder text. This is another piece of placeholder text.

6 Conclusion

Bla Bla [5]

7 Citations

References

[1] Yang Liu. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019.

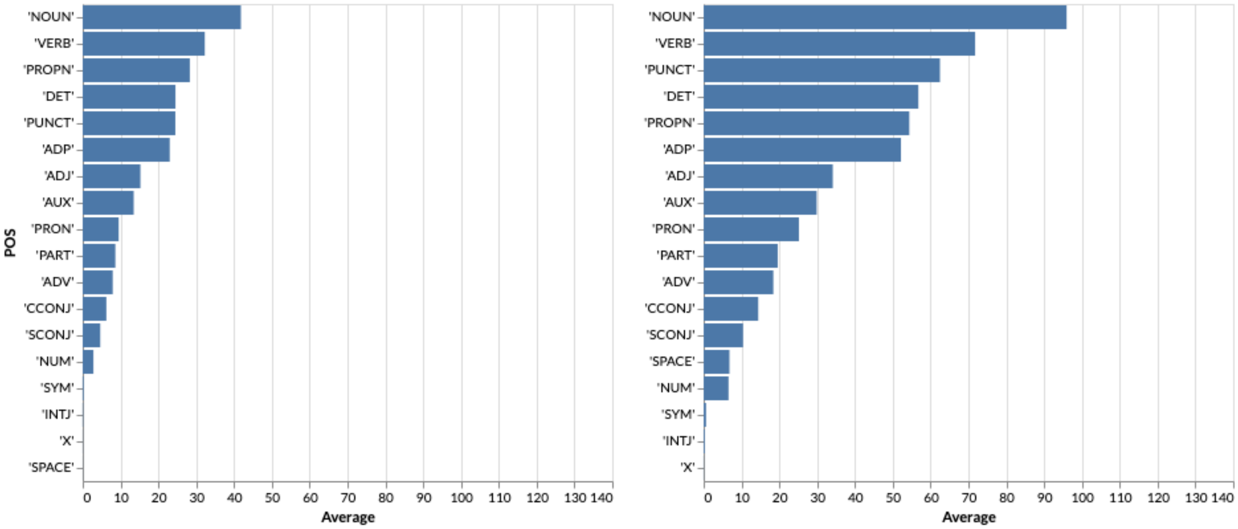
[2] Qian Ruan, Malte Ostendorff, and Georg Rehm. Histruct+: Improving extractive text summarization with hierarchical structure information. *arXiv preprint arXiv:2203.09629*, 2022.

[3] Dehao Tao, Yingzhu Xiong, Jin He, Yongfeng Huang, et al. An unsupervised extractive summarization method based on multi-round computation. *arXiv preprint arXiv:2112.03203*, 2021.

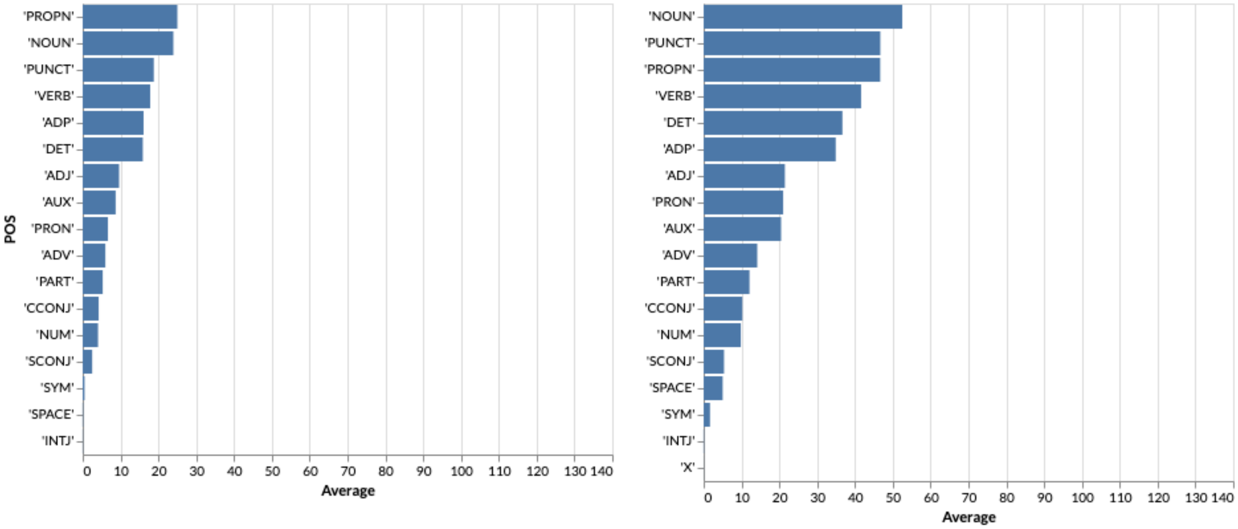
[4] Changchang Zeng and Shaobo Li. Analyzing the effect of masking length distribution of mlm: An evaluation framework and case study on chinese mrc datasets. *Wireless Communications and Mobile Computing*, 2021, 2021.

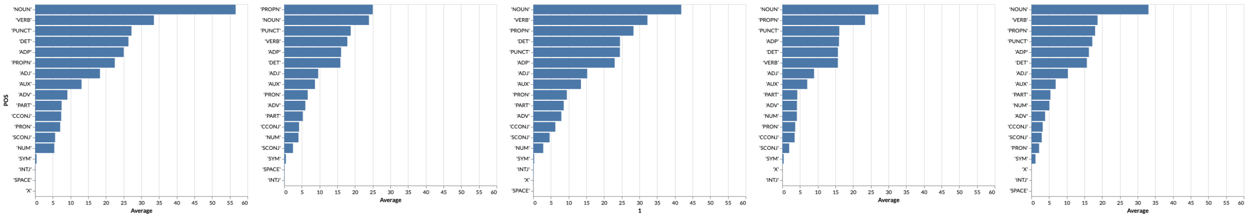
[5] Wanzheng Zhu and Suma Bhat. Gruen for evaluating linguistic quality of generated text. *arXiv preprint arXiv:2010.02498*, 2020.

Distribution of Parts of Speech (POS) for politics



Distribution of Parts of Speech (POS) for sport


POS
'PROPN'
'NOUN'
'PUNCT'
'VERB'
'ADP'
'DET'
'ADJ'
'AUX'
'PRON'
'ADV'
'PART'
'CCONJ'
'NUM'
'SCONJ'
'SYM'
'SPACE'
'INTJ'
Average
'NOUN'
'PUNCT'
'PROPN'
'VERB'
'DET'
'ADP'
'ADJ'
'PRON'
'AUX'
'ADV'
'PART'
'CCONJ'
'NUM'
'SCONJ'
'SPACE'
'SYM'
'INTJ'
'X'
Average


POS
'NOUN'
'VERB'
'PUNCT'
'DET'
'ADP'
'PROPN'
'ADJ'
'AUX'
'ADV'
'PART'
'CCONJ'
'PRON'
'SCONJ'
'NUM'
'SYM'
'INTJ'
'SPACE'
'X'
Average
'PROPN'
'NOUN'
'PUNCT'
'VERB'
'ADP'
'DET'
'ADJ'
'AUX'
'PRON'
'ADV'
'PART'
'CCONJ'
'NUM'
'SCONJ'
'SYM'
'SPACE'
'INTJ'
Average
'NOUN'
'VERB'
'PROPN'
'DET'
'PUNCT'
'ADP'
'ADJ'
'AUX'
'PRON'
'PART'
'ADV'
'CCONJ'
'SCONJ'
'NUM'
'SYM'
'INTJ'
'X'
'SPACE'
'NOUN'
'PROPN'
'PUNCT'
'ADP'
'DET'
'VERB'
'ADJ'
'AUX'
'PART'
'ADV'
'NUM'
'PRON'
'CCONJ'
'SCONJ'
'SYM'
'X'
'INTJ'
Average
'NOUN'
'VERB'
'PROPN'
'PUNCT'
'ADP'
'DET'
'ADJ'
'AUX'
'PART'
'NUM'
'ADV'
'CCONJ'
'SCONJ'
'PRON'
'SYM'
'X'
'INTJ'
'SPACE'
Average