Probing of Quantitative Values in Abstractive Summarization Models

Nathan M. White
James Cook University
nathan.white1@jcu.edu.au

Abstract

Abstractive text summarization has recently become a popular approach, but data hallucination remains a serious problem, including with quantitative data. We propose a set of probing tests to evaluate the efficacy of abstract summarization models' modeling of quantitative values found in the input text. Our results show that in most cases, the encoders of recent SOTA-performing models struggle to provide embeddings that adequately represent quantitative values in the input compared to baselines. Under our assumptions, this suggests that the encoder's performance contributes to the quantity hallucination problem. Furthermore, performance versus standard BERT suggests that typical pre-training and fine-tuning approaches for the summarization task may play a role in underperformance for some encoders.

1 Introduction

Concomitant with the rise of abstractive summarization has been a phenomenon termed hallucination—i.e., the appearance of content in the generated summarization that is not in the original text and is typically erroneous (Nan et al., 2021). This represents a grave problem in the development of abstractive summarization; Maynez et al. (2020), for example, found that abstractive summarization systems produced summaries with inappropriate content more than 63% of the time.

One type of hallucination that plagues abstractive summarization in areas such as finance and economics is quantity hallucinations (Zhao et al., 2020a), where quantitative values appear in the summary output but not in the input.

Approaches have emerged to probing word embeddings to evaluate the effectiveness of quantitative representation (Naik et al., 2019; Wallace et al., 2019). A probing approach can be applied to the abstractive summarization context, given that for a model to adequately summarize text, it must be able to recognize the quantitative values that need to be reproduced in the output.

Our contributions are as follows:

- We propose and explore several probing tasks related to representations of quantitative values in the abstractive summarization context for the first time.
- We found that model architectures with recent state-of-the-art performance on the abstractive text summarization task struggled to adequately represent the quantitative values in their input as compared to baselines, suggesting that this is an important source of quantity hallucinations.

2 Methods

We explore six numerical probing tasks to evaluate the encoder's output representations. The focus of each task is to discern whether the signal of the quantity or unit is recoverable from the summarization encoder's output, where each task focuses on a different type of quantitative representation in text. The tasks are: percent decoding, basis point decoding (i.e., 15 basis points = 0.0015), order decoding (i.e., for orders of magnitude, such as *billion* in 15.3 billion), identifying endpoints of ranges (e.g., 27.3–65.1), addition, and identification of units associated with quantitative values (e.g., m^2 in 16.8 m^2).

Eight Transformer-based models are considered for the probing tasks, each fine-tuned for abstractive summarization; several have achieved state-of-the-art (SOTA) results for the task. These are compared against three baselines:

Randomly initialized vectors, an untrained version of BERT (Devlin et al., 2019), and a trained version. The experiment models were composed of two parts: the fine-tuned Transformer-based model with its parameters frozen and a probing model trained in each experiment.

3 Results

The encoders of all of the abstractive summarization models considered provide embeddings that model the numerical values to some degree as compared to the random vector baseline in most tasks.

Moreover, as represented, for example, by the excerpt of results for the percent decoding task in Table 1, the encoder outputs from several summarization models provide representations for quantities that perform worse than each baseline. Furthermore, the trained BERT baseline outperforms the abstractive summarization models in a number of cases, suggesting that standard pre-training and/or fine-tuning methods may play a role in inferior modeling of quantitative values.

Finally, no model provides consistent superior performance, suggesting that typical methods for text summarization do not provide a decisive advantage to adequately represent quantitative values.

4 Conclusion

Our results show that in most cases, the encoders of recent SOTA-performing models struggle to provide embeddings that adequately represent quantitative values in the input compared to baselines. Under our assumptions, this suggests that the encoder's performance contributes to the quantity hallucination problem.

Furthermore, performance versus standard BERT suggests that typical pre-training and fine-tuning approaches for the abstractive summarization task may play a role in underperformance for some encoders.

	Percent Decoding [0.0, 99.9] (RMSE)
Baselines:	
Random Vectors	9.365
$BERT_{Untrained}$	9.464
$BERT_{Trained}$	5.617
Pegasus-XSum	*9.756*
Pegasus-CDM	*11.349*
T5-CDM	*11.711*
BART-XSum	5.332
BART-CDM	9.419
DistilBART-XSum	6.234
DistilBART-CDM	6.209
ProphetNet-CDM	**4.601**

Table 1: Mean results for the percent decoding task with floats in ranges [0.0, 99.9]; mean results worse than all baselines italicized.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019. 4171–4186.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 1906–1919.

Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3374–3380.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2727–2733.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 5307–5315.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020a. Reducing Quantity Hallucinations in Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2237–2249.