MEDIQA-Chat 2023 Section Summarization

Ionescu Andrei[1]
[1]Artificial Intelligence - 407

June 28^{th}, 2023

Abstract

In the context of the medical field doctors that consult pacients can take notes along their discussions with the people that they are helping, in order to highlight important information that might be indicative of possible symptoms, illnesses, or just tracking the evolution of the health of pacients. Moreover, reviewing the history of many such notes can create fatigue which diminishes the actual service of helping the pacients. By highlighting essential information, using automated procedures, that have become possible due to the increasing research in deeplearning approaches, we can reduce this burden. As such the EACL 2023 Paper[1] was proposed and a couple of shared tasks were created to drive more interest in this area (see ACL-ClinicalNLP 2023). In this paper I inspect the performance of two **encoder-decoder** models namely T5[2] and In-BoXBART[3] that are instruction fine-tuned on the first shared task.

1 Introduction

Machine learning methods have evolved at a very fast pace in the latest years. In the not-so-distant past, the usual approach to text processing was to do manual engineering of features and train well-known models on top of these. While this approach is still widely used in many contexts, more recent developments in the deeplearning subfield have shown to provide solutions that can be learned end-to-end on various **complex tasks** and output reasonable results. Therefore we can leverage these techniques in order to decrease the amount of time spent by a doctor highlighting his own notes in order to take decisions at a faster pace, with less fatigue resulting in higher quality services and greater pacient care. This paper's objective is to tackle the task of summarizing such discussions or "chats" between a doctor and a pacient, and to highlight the type of such a discussion. A shared task was proposed in this direction and has multiple subtasks, that use different datasets. In the context of this paper I tackle the objective of both generating a section header and summarizing the content of a doctor-pacient discussion with the only released, as of 28^{th} June, MTS-Dialog Dataset which also provides some benchmark values. According to the paper the sections can be classified by using one of the twenty categories which include: assessment, allergy, diagnosis, exam, past surgical, medications and so on. The dataset consists of $1.7k$ training paired samples, where each sample has a dialogue of about 9 turns between the doctor and the pacient, a section header and a summary. The dialogue is synthetic and created by experts from the

summaries in order to keep the privacy of the data. Moreover there are also 100 entries for the validation set, 200 samples for the test set and another training set of $3.6k$ examples, that was augmented through the back-translation technique using spanish and french for their closeness to english.

2 Related Work

Summarization tasks are in general quite challenging even for a human as it requires a deep understanding of the text such as the comprehension of the text, the ability to exclude redundant information and the ability to formulate logical ideas that provide the essentials in a short amount of words. However, due to the usefulness of this objective in day-to-day life and the abundant presence of data with popular general datasets such as DailyMail, CNN, SAMSum, Wikihow and others, and with the aid of the Transformer architecture[4], which has gone over the previous performance of RNNs, and has revolutionized the NLP field, many new approaches were developed on top of this and provided new benchmarks which represent the current state-of-the-art. Summarization can also be tackled in two ways:

- **extractive summarization** - which uses existing words and sentences, and techniques that retrieve these relevant elements and concatenate them to offer a summary
- **abstractive summarization** - a more powerful technique, which can rephrase ideas using words different from the initial text, and can also provide the same idea albeit written in a different manner.

Many recent models mostly focus on the abstractive summarization but can also mix in the extractive summarization task in order to enforce the models to tend to sentences that have higher saliency scores.

Works in the field have evolved from pointer-generator networks[5] that used recurrent neural networks and attentions to reduce repetitive tokens by taking into account their occurence while also reusing existing tokens from the text in the favour of rare occurences. Recent methods use the encoder-decoder scheme and leverage the transformer architectures such as the BART[6] model which is trained to reconstruct documents, in an auto-regressive manner, that have had noise introduced by masking out tokens or by having entire ranges permuted. Because BART is an example of a model that is generally suitable for various tasks, it has been explored in many ways such as the PEGASUS[7] which focuses solely on the abstractive summarization task by masking out a big contiguous portions of tokens and trying to reconstruct them. In turn, this enforces the model to better understand context and key ideas that should be extracted. Another example of such an architecture that focuses on summarization is the Curriculum-Guided Abstractive Summarization[8] which extends the Transformer architecture by adding another cross-atention layer in the decoder, in order to pay attention to more salient regions of the text. This is due to the fact that BART has a

shortcoming in content selection. Because of this, the approach is two-staged and optimzes both the objective of saliency regression for each sentence and also enforces the abstractive summarization objective.

3 Methods

According to the EACL 2023 Paper, the **augmentation** done via the **back-translation technique** showed improved results pertaining to abstractive summarization. Therefore, I considered using the augmentation dataset. In addition, it has a greater amount of samples which can reduce the chance that the models might overfit the task.

Considering the recent developments in encoder-decoder architectures and their proved performance I considered using two pretrained instruction-based models: T5[2] and In-BoXBART[3]. I chose these models as they have already been trained on extensive datasets: T5 has been pretrained on the *Collosal Clean Crawled Corpus*, a cleaned web archive and further finetuned on multiple downstream tasks to improve generalization, while In-BoXBART was trained on BoX a collection of 29 widely adopted biomedical NLP datasets. Furthermore, both these models have been trained in a text-to-text manner in order to uniformize the training process. The T5 has also been finetuned, on the CNN/Daily Mail, to include the **summarization** instruction. Another motivation for choosing these models is to be able to compare their performance as they were trained on different data domains (*open vs biomedical*).

The data is preprocessed considering a few aspects:

- The dialogue length is capped at 160 tokens while the summary is capped at 60 tokens. I considered these values to be above the 75^{th} percentile and to accomodate most of the cases. If the samples are shorter than this then the samples are padded to the right.
- A "summarize:" instruction was prepended to every dialogue sample in order to give a hint to the models as this is the desired instruction that should be applied
- A new special token, "<HEADER>" is added to the tokenizer and to the models' token embeddings. Every augmented dialogue text is prepended with this token given a binomial probability draw of p=0.5 (and n=1) for every sample. If this token is added to a dialogue text, then also the ground-truth header is prepended to the ground-truth summary. By training the models across a number of epochs with this approach the model learns that it should classify the section when the token is present.
- The test data is processed similarly as above with only the mention that the "<HEADER>" and the ground-truth header are prepended to all respective samples.

The two models are finetuned using AdamW with a learning rate of $5 * 10^{-5}$ and $(\beta_1, \beta_2) = (0.9, 0.999)$ with $w_{decay} = 10^{-2}$. The training is performed over 10 training epochs on the augmented dataset with batches of 32 samples. At every epoch a validation step is performed

to investigate the performance of the models. The evaluation is done between the predicted tokens and the ground truth tokens by leveraging the following metrics: ROUGE-1, ROUGE-2, ROUGE-L, BLEURT-20, BERTSCORE.

4 Results

After applying the training approach, both models learn about the new token and classify the section if the user prompts it. Furthermore, the models learn the abstractive summarization task and give outputs similar to the ground truth , however at times the information is hallucinated or important details are either left out or are wrong. From a qualitative point of view the approach is lacking but it represents a good start for further improvement. The metrics obtained quantitatively can be observed in Figure 1.

5 Conclusion

The encoder-decored pretrained models both have good generalization capabilities on downstream tasks but due to their lack of robustness they can hallucinate, mislabel the section header and omit or get wrong essential details which might disrupt a doctor's workflow if deployed in practice. As a future direction an idea would be to treat the header prediction as a separate task and perform joint learning by having another output head for this task. Another idea for research would be to take inspiration from the work in Curriculum-Guided Abstractive Summarization[8] and come up with new cross-attention modules in order to enforce the model to pay attention to more salient information that can be found in the dialogue. For example, NER could be performed on the input text and the resulting entities may be forwarded to the models' cross-attentions to include those distinctive keywords that can indicate ilnesses or body parts and so on.

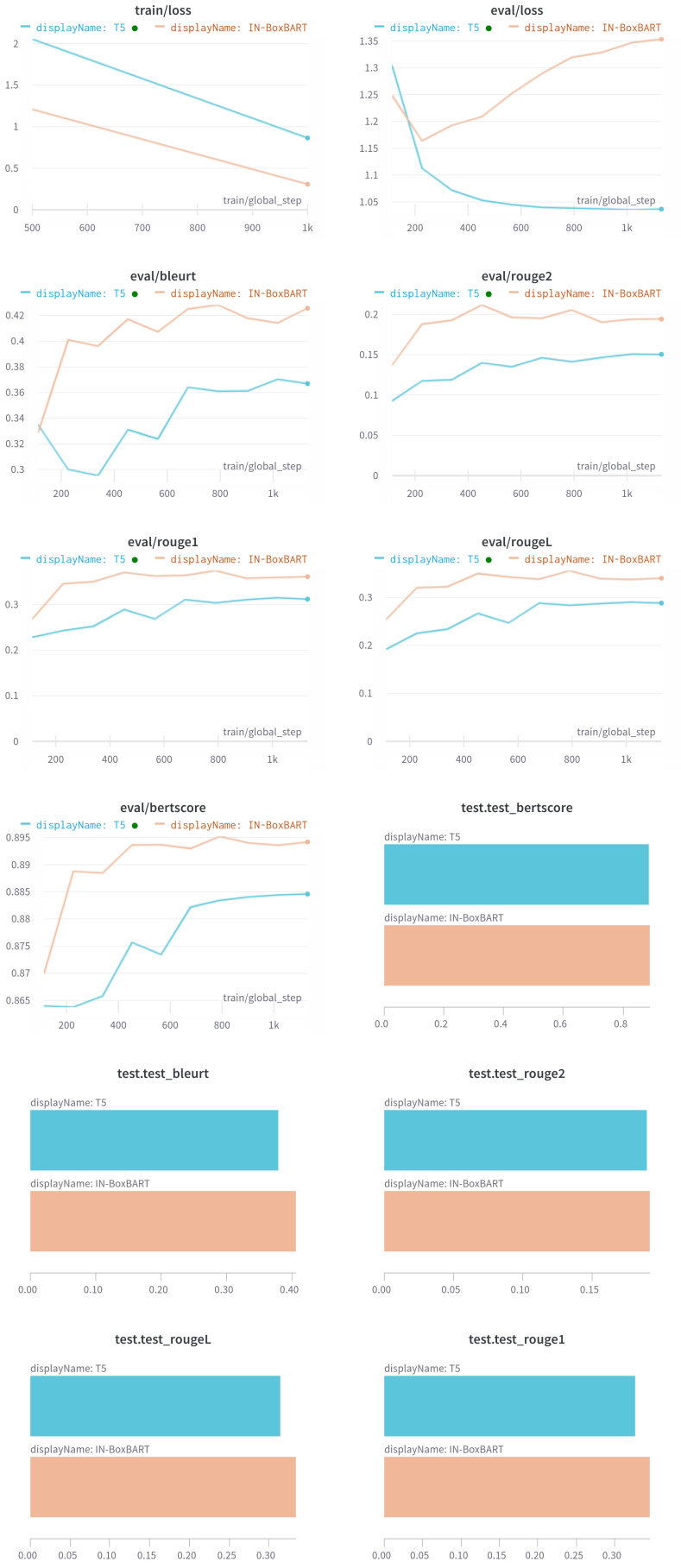

Fig. 1. Results

WANDB.AI[9] was used to track and produce these charts.

6 References

1. Ben Abacha, A., Yim, W.-w., Fan, Y. & Lin, T. *An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters* in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, Dubrovnik, Croatia, May 2023), 2291–2302.
2. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* 2020. arXiv: 1910.10683 [cs.LG].
3. Parmar, M., Mishra, S., Purohit, M., Luo, M., Murad, M. H. & Baral, C. *In-BoXBART: Get Instructions into Biomedical Multi-Task Learning* 2022. arXiv: 2204.07600 [cs.CL].
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. *Attention Is All You Need* 2017. arXiv: 1706.03762 [cs.CL].
5. See, A., Liu, P. J. & Manning, C. D. *Get To The Point: Summarization with Pointer-Generator Networks* 2017. arXiv: 1704.04368 [cs.CL].
6. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* 2019. arXiv: 1910.13461 [cs.CL].
7. Zhang, J., Zhao, Y., Saleh, M. & Liu, P. J. *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization* 2020. arXiv: 1912.08777 [cs.CL].
8. Sotudeh, S., Deilamsalehy, H., Dernoncourt, F. & Goharian, N. *Curriculum-Guided Abstractive Summarization* 2023. arXiv: 2302.01342 [cs.CL].
9. Biewald, L. *Experiment Tracking with Weights and Biases* Software available from wandb.com. 2020.