Alexis Kaldany Individual Report
NLP Fall 2022
December 9th, 2022

Introduction

Our project analyzes the relationship between the summary/text ratio (a variable we constructed) to the various Rouge metrics. This project was conceived by me with the intent of discovering if summarization tasks varied in their effectiveness (as determined by the rouge score) depending upon the relative size of the summary compared to the size of the summarized text.

Description of Individual Work:

1. I tested all of the models I identified as being possible candidates inside the "training_dialogsum.py". I also created this testing suite where I used the autotrainers to rapidly iterate and test.
2. I built the main training/evaluation/testing loop "dialogsum.py" where all statistics about the models effectiveness and the outputted summary predictions are generated and saved. The underlying model has a checkpoint system enabling extended training over multiple days. It is based on the T5 models held on HuggingFace. This file, which is the core of my work, does not include or use any autotrainers. The entire training/evaluation/testing loop was created by me.
3. I formatted and created the data for my teammates so they were able to perform the Exploratory Data Analysis section and the Post Evaluation Analysis.
4. I conceived of this project, identified the dataset, organized the repo and prepared the work for my teammates, as well as directed them in their respective individual projects.
5. I identified the metrics we should use and implemented them.

Results

All of the code I wrote worked precisely as intended and generated all desired outputs.

According to Saagar's work there is no relationship between the summarization ratio and any of the rouge metrics.

Epoch	Loss	Rouge1	Rouge2	Rouge L
1	1.3	0.56	0.27	0.53
2	1.24	0.58	0.3	0.55
3	1.22	0.59	0.31	0.56
4	1.21	0.59	0.31	0.57
5	1.21	0.59	0.31	0.57

The above table shows the results from 5 epochs of training the model.

Summary

1. I built a summarization model using the pre-trained “t5-small” model from huggingface to summarize the DialogSum dataset (also from HuggingFace).
2. We analyzed the relationship between the scoring metrics (rouge-1,rouge-2,rouge-L) and the ratio of the length of the summary divided by the length of the original text.
3. We found there was no significant relationship at all.

Calculation

I write all my own code, otherwise I wouldn’t understand it. So:

0
— ==== 0
514

References

1. Dataset: https://huggingface.co/datasets/knkarthick/dialogsum

2. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140), 1-67.

3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

4. Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19 news translation task submission. arXiv preprint arXiv:1907.06616.