Insights

1. The use of the [BART model](https://huggingface.co/ainize/bart-base-cnn) really helped simplify the task.
2. Using model’s pretrained on [Facebook conversations](https://huggingface.co/facebook/bart-large-cnn) also helped capture the fact that a lot of the conversations in the dataset were informal.
3. Predicted and True summary length appear to have a solid positive correlation.
4. Interesting that the ROUGE scores all appeared to have pretty normal distributions except for R2. Not surprising that there is a pretty consistent linear (positive) relationship between the various scores.
5. I was surprised to see that there did not appear to be any strong relationship between the quantitative metrics I created for conversations and ROUGE score. If anything, there was a slightly negative relationship between conversation complexity and ROUGE scores. I would argue this points to how this metric is an incomplete measure of summarization.

Problems

* **Data Size/ Resource Efficiency**—The time and resources required to build these models (and to generate summaries, if one is not careful with their choice of parameters) is a major drawback that really only applies in this case since I lacked the resources. Nevertheless, it made for true exploration of the different parameters very difficult. Even trying to use the bigger FB-based model increased training time by a substantial multiple. It also limited the training process as my computer overloaded several times. Ultimately, I had to figure out workarounds to address execution-time errors like ‘pin-memory’.
* **Ability to Assess**—I believe that automatic assessment is just not a good idea for dialogue/conversation summarization. There is too much nuance in summarizing a conversation that a since quantitative metric cannot capture. See future steps. If the business problem requires some sort of quantitative metric, I think the best data will be provided (directly or indirectly) by users. Things like engagement and screen time; also user satisfaction surveys are going to be the best way to assess a model and understand how to improve it.

Future Steps

* **Translating emoji’s and other chat abbreviations**
* **Sample Dialogues for Testing**—I think a great way to analyze these models is to devise test case dialogues and summaries that capture various nuances based on the business needs. New models can be tested and understood by the summaries they generate for these special cases as well as a broader analysis on what is set aside for testing in the original dataset.
* **Improving human-made summaries**—as discussed in the project proposal, I noticed the human-made summaries tended to be an extraction of information revealed in the conversation rather than be a summary of the conversation itself. For example:

|  |  |
| --- | --- |
| * **Amanda**: I baked cookies. Do you want some? * **Jerry**: Sure! * **Amanda**: I'll bring you tomorrow :-) | **Amanda** baked cookies and will bring **Jerry** some tomorrow. |

The human summary here reports the results/consequences (what happened and what will happen) of the conversation but, in my opinion, an important aspect of this conversation is Jerry’s affirmation. I might re-write this summary as follows: “Amanda baked cookies and will bring them to Jerry, at his request.” (This is not a perfect summary either, but the final clause is my point.). Ultimately, the summaries need to be designed based on the specific business problems the model is meant to address.

* **Further Research**—I wonder if these models can be expanded to analyze entire chat threads with multiple conversations being had simultaneously (and it not being clear who was speaking with whom). Or at least handle conversations that have multiple topics that need to be summarized. I thinking designing such a dataset would be a very straight-forward matter with the dialogue dataset(s) readily available.
* **Practical Application**—Working with this project and seeing what was possible with these models has led me to believe that using this agent would ideally be supplemented with a classification tool so that the user can filter chat’s based on topic covered (& participants) before optionally accessing the generated summary(ies). This is much more practical, especially since the generated summaries are not necessarily going to be easier to consume.