Scott Belarmino Raw Notes from Overleaf

approach: The primary task is to test if our best customized fine-tuned pre trained model can accurately categorize each element of the SOAP framework relative to our dataset's test summarizations. Prior to this we've built and tested five models to determine which of these models summarize well overall and of those top models, how do they perform in categorizing SOAP elements.

The "off the shelf" DistilBART model without FT was used as a baseline against the following DistilBART versions: DistilBART w/ FT, DistilBART w/ FT + Annotations and DistilBART w/FT + KeyBERT. The DistilBART model is a distilled version of the BART (Bidirectional and Auto-Regressive Transformer) model. The specific model (sshleifer/distilbart-cnn-12-6) from HuggingFace was used for all the DistilBART experiments. This was chosen for its ability to excel in summarization tasks while being lightweight compared to other BART models such as bart-large-cnn (300 MM parameters vs. 406 parameters - cite huggingface). This version of DistilBART was also chosen as it was pre-trained with the CNN/Daily Mail dataset with the specific task of generating detailed summarizations, ideal for inferring doctor and patient dialogue. This was chosen over other models that were pre-trained on XSum (Extreme Summarization) which generates abstractive single sentence (i.e., headlines).

During the generation process, the baseline model was given a simple prompt (see fig. 3) to explicitly summarize the dialogue using the SOAP format. This prompt was excluded in the following approaches as the input dataset targets/references are in the SOAP format.

This approach takes the DistilBART baseline model and fine-tune with the Bilal-Mamji/Medical-summary dataset from HuggingFace. Data pre-processing was simplified using HuggingFace's BartTokenizer and adjusting the input max token length to 900 to match the longest doctor-patient dialogue example in the dataset. This was to ensure all aspects of the dialogue were capture so the model would not lose any context. The target max token length was set to 600, however, most of the ground truth SOAP summaries were well below this number. Training arguments were established by keeping memory resources in mind. Key arguments to save on memory and proper training include: setting smaller batch size per gpu (4), limiting to three epochs, setting the learning rate low to 5e-5 to avoid overshooting, and using mixed-precision (16-bit floats) to speed up training and reduce memory. These arguments will be consistent for each fine-tuned model in the experiment. Similarly, the model generation/inference arguments will be constant across all fine-tuned models. This includes output summaries with max length set to 500 tokens with number of beams set to 4 and temperature set to 1.0 to ensure a balanced extractive and abstractive summarization.

Building upon the DistilBART w/ FT model, this approach includes medical annotations to the input text. The hypothesis was to improve SOAP generation by including annotations from the Universal Medical Language System (UMLS) developed by the National Library of Medicine (cite). The UMLS provides a standard bio-medical library which was used to identify common medical entities in the dialogue and apply medical annotations prior to pre-processing and tokenization. The test was to see if the added annotations to the input text will provide a richer input for the model and remove any disambiguity with the entities and their relationships. Additionally, user SOAP based rules were introduced during the annotation process which consists of manual SOAP categorization of UMLS entity groups. There were over 50 UMLS entity groups with multiple subgroups that cover a large breadth of medical terminology. There were over 50 UMLS entity groups with multiple subgroups that cover a large breadth of medical terminology. Take a few entities from of these subgroups, the annotation of "short of breath" would be categorized as Sympton (S) and the annotation of "albuterol" would be categorized as Observation (O). Naturally, the addition of annotations to the input text increases the token length larger than the max token length for our DistilBART model. To overcome this obstacle, a sliding window chunking function was used to break these larger inputs down into chunks (max length=900 tokens) with an overlap (stride=256 tokens) per example to maintain context and consistency with the chunks. This process is done during the pre-processing phase prior to training (see fig 5) with the goal of providing improved SOAP summarization over baseline.  The DistilBART with FT + KeyBERT is another model built with the goal of enriching the input text for improved SOAP summarization. The KeyBERT model is a keyword/key phrase extraction BERT model which was trained to pull the most relevant information from the input text. In this process the input text is distilled to the most salient information for the model to interpret. The hypothesis is if we can denoise the dialogue prior to DistilBART's denoising process, then we will get a more informative summarization.

The "out of the box" DistilBART model (baseline) with a simple prompt of "Summarize the following dialogue using the SOAP framework:" which included descriptions of each element to guide the pre-trained model in the summary generation. We observed the results of the baseline were underwhelming with low evaluation scores across all metrics (Fig. 2). We hypothesize this may be due to the way the DistilBART was trained specifically for summarization tasks and not very strong at taking specific instructions. Additionally, the long prompt may have overwhelmed the model, yet tested simpler prompts also yielded the same results. The baseline's overall summarization did not follow the SOAP framework and summarized the the dialogue with three to four sentences.

The best performing models were DistilBART FT and DistilBART FT with Annotation with evaluation scores nearly three times the baseline model. With each generated summary, both models matched 0.6741 (Rouge-1) of the words in the ground truth summary, which suggest the models capture important information yet provides some room for diversity in the summarization. Beyond evaluating the matching between words, the F1 BERTScore of 0.9203 (FT) and 0.9211 (FT + Annotations) were also an improvement over baseline meaning the models summaries are semantically aligned to the ground truth.

As both models had a near identical performance, the added time and effort to introduce our own SOAP based rules during the annotation process did not yield the results we were expecting. While the annotations may have helped the model understand entity relationships and mitigate disambiguity, the model may had more focus on the reference data during training or the annotations may have been added noise it ignored during training. Given this, the DistilBART w/ FT only is a better option for implementation.

While FT greatly improves DistilBART's performance not all FT performs as well as our best FT models. The DistilBART FT w/ KeyBERT performed better than baseline, but did not match as may words as the ground truth (R1=0.4728) or semantically align (F1=0.8716) to the ground truth as the other FT models. We hypothesize this may be due to the over distillation of the input dialogue using the KeyBERT model, which reduces the dialogue to the most relevant components. Our thought process was to denoise and add structure to the input dialogue prior to fine-tuning the DisitlBART model. However, we believe denoising the input prior to it being denoised again by the model drove the drop in performance. This led to the model having less of an understanding of the true dialogue during training.

Focusing on the individual elements of SOAP, we found a few interesting observations when comparing each model. Note the DistilBART baseline model did not produce a SOAP summary and will be excluded in the discussion. All three FT models performed similarly to their overall summary evaluation, however, the FT and FT w/ Annotations had different results in Rouge-1 and F1 score between Subjective/Objective and Assessment/Plan. For subjective and objective both were capturing 68\% to 70\% of the ground truth words in subjective and objective categories and semantically aligned with 93\% F1-score. However, the models did not perform as well with assessment and plan categories, 48\% word matching for assessment, 53\% word matching for plan, and 90\% F1-score. The lower Rouge-1 scores may be due to the ambiguity between assessment and plan. For medical SOAP notes, the assessment's task is to summarize the subjective and objective notes to arrive a diagnosis (citation). The plan's task is to provide tests and treatments associated with the diagnosis. As both involve context of the diagnosis there may be miscategorization between the two.

In addition to the NLP evaluation metrics, we were also able to have three medical experts examine the generated summaries of our FT models. They were asked to evaluate a sample from the test set of 10 doctor/patient dialogues and their generated summaries. They were given a Likert rating scale from 1 to 10 to rate the following; Accuracy: Does the prediction correctly summarize the doctor/patient dialogue?, Completeness: Does the prediction include all key details from the doctor/patient dialogue?, and 3. Consistency: Does the prediction follow the SOAP format (Subjective, Objective, Assessment, Plan)? The model overall averaged high scores in Accuracy (8.8/10), Completeness (8.5/10), and Consistency (8/10). The medical evaluators noted some issues on the model mislabelling elements from the dialogue. One evaluator noted, "The prediction model did not discriminate between objective data (the model inserted the procedural details into the Assessment, again assessment is the section of the SOAP note in which the provider generates a clinical impression (ie how sick or severe is the patient and the associated condition as well as potential causes of the patient’s current condition)." Overall, the evaluators were impressed by the model and all noted how useful SOAP summarization would benefit their practice.

The pursuit of text summarization goes back as 1958 (Luhn, 1958) with IMB scanning machine-readable text and calculating the most relevant aspects of the document by measuring word distribution and word position within a sentence. Over the past ten years, the pursuit of not only extractive, but also abstractive summarization has grown. In 2016, nearly 60 years later, another IBM team set their research on absrtactive text summarization using an attentional Recurrent Nerual Network encoder-decoder architecture (Nallapati et al., 2016).

(Goo and Chen, 2018) were one of the early researches to review dialogue summarization by reviewing dialogue acts or interactive cues between individuals during a meeting. The process of dialogue summarization has evolved to other domains such as medicine. (Joshi et al., 2020) approached medical dialogue summarization through pointer generator networks finding a balance between copying their source text and generating novel content.

Inspired by these past studies, especially in the medical domain, we wanted to experiment with developing a SOAP summarization model using existing pre-trained models and apply novel techniques. While many of these models have been trained very well with many large corpusus, we set out to improve them prior to fine-tuning by influencing their input dialogue by discovering entity relationships (NER, KeyBERT) and supplemenatl annotation.