Summarizing Medical Documents using Transfer Learning

Problem Statement

Having a great deal of first and second-hand emergency care experience, I recognize the importance of personal advocacy in these situations. A key component of individual advocacy in healthcare comes from understanding our past procedures and how these affected our care outcomes; however, for the layperson, understanding the complex language used in medical documents can inhibit our ability to understand the nature of the procedures and tests done. Beyond the information physician provides us, what else can we learn from the plethora of documents our care creates? Is there other relevant information? How can it be identified and thereby reduce the amount of learning needed to instill a sense of ownership over our continued care? Can we fine-tune a model using transfer learning to summarize medical documents?

Data Sets

- MTSamples.com data of over 5000 transcribed medical documents from a wide range of medical specialties.

The data for this project was accessed in csv format through Kaggle.com here.

Data Wrangling

For this project, the data wrangling and normalization process focused on extracting entire sections of text from within each document such that all relevant information was retained. The sections of primary interest are: Indications, Findings, and Impressions, as they generally held the most relevant information about a visit and or a procedure. To extract the sections I first explored a regex splitting pattern after noticing that a period followed by a comma (.,) typically separated sections within a document. However, this would also break list-like findings into individual components. I ultimately ended up splitting documents into sections by using a forward search regex pattern –”(.*?)(?=˙[A-Z][A-Z]+—)”– to find all words before a section header. Using this, I was able to pull out specific sections and then remove the headers, as this would retain the structure of each sentence in each section. See below for an example:

> PROCEDURE: , Sigmoidoscopy.,INDICATIONS:, Performed for evaluation of anemia, gastrointestinal Bleeding.,MEDICATIONS: , Fentanyl (Sublazine) 0.1 mg IV Versed (midazolam) 1 mg IV,BIOPSIES: , No BRUSHINGS:,PROCEDURE:, A history and physical examination were performed. The procedure, indications, potential complications (bleeding, perforation, infection, adverse medication reaction), and alternative available were explained to the patient who appeared to understand and indicated this. Opportunity for questions was provided and informed consent obtained. After placing the patient in the left lateral decubitus position, the sigmoidoscope was inserted into the rectum and under direct visualization advanced to 25 cm. Careful inspection was made as the sigmoidoscope

was withdrawn. The quality of the prep was good. The procedure was stopped due to patient discomfort. The patient otherwise tolerated the procedure well. There were no complications.,FINDINGS: , Was unable to pass scope beyond 25 cm because of stricture vs very short bends secondary to multiple previous surgeries. Retroflexed examination of the rectum revealed small hemorrhoids. External hemorrhoids were found. Other than the findings noted above, the visualized colonic segments were normal.,IMPRESSION: , Internal hemorrhoids External hemorrhoids Unable to pass scope beyond 25 cm due either to stricture or very sharp bend secondary to multiple surgeries. Unsuccessful Sigmoidoscopy. Otherwise Normal Sigmoidoscopy to 25 cm. External hemorrhoids were found.

Would become:

FINDINGS: Was unable to pass scope beyond 25 cm because of stricture vs very short bends secondary to multiple previous surgeries. Retroflexed examination of the rectum revealed small hemorrhoids. External hemorrhoids were found. Other than the findings noted above, the visualized colonic segments were normal. INDICATIONS: Performed for evaluation of anemia, gastrointestinal Bleeding.

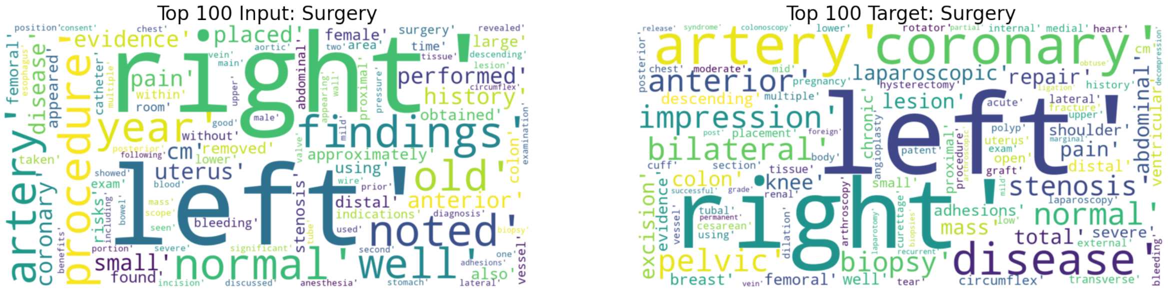
IMPRESSION: Internal hemorrhoids External hemorrhoids Unable to pass scope beyond 25 cm due either to stricture or very sharp bend secondary to multiple surgeries. Unsuccessful Sigmoidoscopy. Otherwise Normal Sigmoidoscopy to 25 cm. External hemorrhoids were found.

EDA

For this project, the EDA focused on exploring unique word frequencies between the input and target sequences by medical specialty and the input and target sequence lengths.

Medical Specialties

To identify any possible patterns or relationships in the top words and to ensure that the words fit within the specialty stopwords were removed.



Top 100 Input: Surgery

Top 100 Target: Surgery

Figure 1: Top 100 unique words for the input and target sequences for the Surgery specialty.

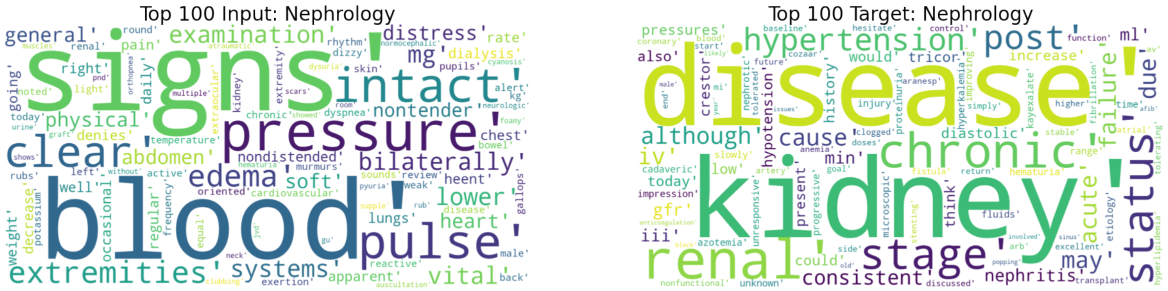


Figure 2: Top 100 unique words for the input and target sequences for the Nephrology specialty.

From the above samples we can see clear patterns in both specialties. Surgery shows signs of delineating between left and right sides of patients and their organs as well as vernacular relating to heart surgery, while Nephrology talks about Kidneys and Kidney disease as expected.

Sequence Lengths

Having previously identified using a T5 model as an excellent candidate, I knew that the maximum length of sequences the model could summarize was 512 tokens. Therefore I decided to investigate the number of input sequences that had more than 512 words. Roughly 2-percent of the input sequences contain more than 512 words. Similarly, I investigated how many of the target sequences were longer than 128 characters, as this is the upper limit on summary length. What I ended up finding was that roughly 4-percent of all target sequences contain more than 128 words.

Both distrubtions are right skewed with the majority of sequence lengths being well below the token limits and having a tail of outlying lengths. These were still used as part of the dataset knowing that any sequence above the maximum length would be truncated.

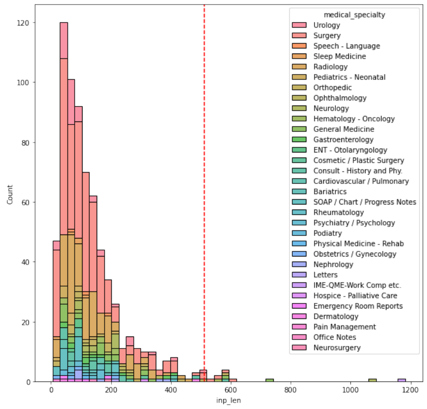


Figure 3: Distribution of input sequence lengths.

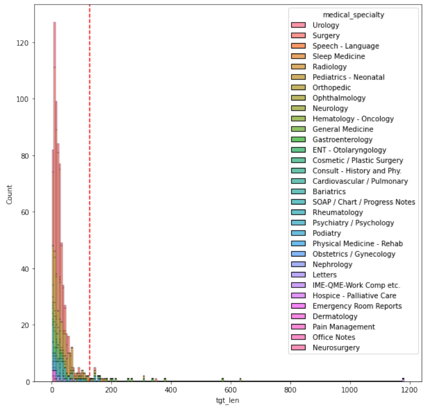


Figure 4: Distribution of target sequence lengths.

It is also worth mentioning here that the model also tokenizes punctuation, and the sequence lengths I measured only included words.

Model Selection

For this project, I explored the impact retraining a T5-small model had on the quality of summaries generated. Using PyTorch and PyTorchLightning modules, I retrained a HuggingFace model for conditional generation using the T5-small checkpoint as the model architecture for 50 epochs. I chose this as a cut-off point because although validation loss continued to decrease, the loss was converging asymptotically.

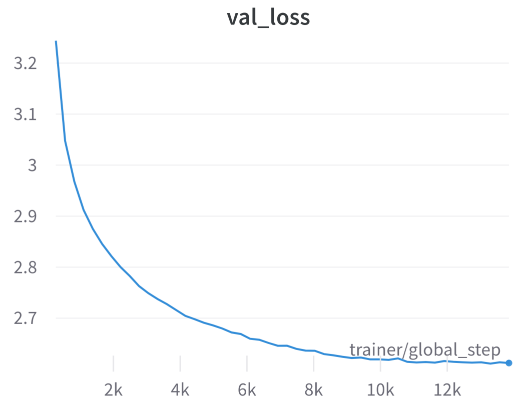


Figure 5: Validation loss quickly declines before reaching a point of asymptotic decay.

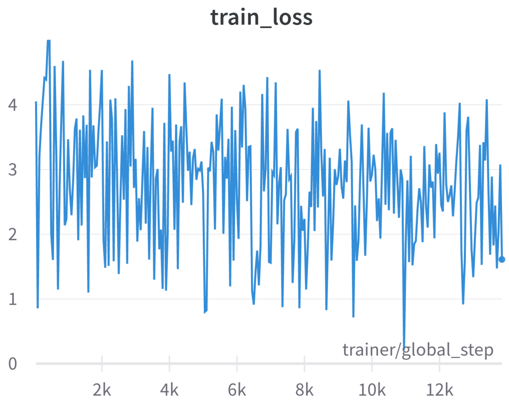


Figure 6: The training loss doesn't seems to fluctuate around a consistent value.

Comparing the performance of the two models was done by bootstrapping the average of the calculated Rouge-L f-scores for each model. It is worth mentioning that it was not always the case that the target summaries extracted from the documents shared a similar structure to the generated summaries. This is an important fact as the Rouge-L metric measures similarity between a generated summary and a provided summary by looking for the longest common subsequences of words, if there are zero words in common then the score will be zero. The 95% CI of mean Rouge-L for the base model was: [0.1305, 0.2006], while the 95% CI of mean Rouge-L for fine-tuned mode was: [0.1247, 0.1907]. A t-test confirmed that based on the substantial overlap of the CIs that there was no significant difference between the results.

Visually comparing a random selection of results showed that both models seemed to struggle on the same types of documents. In some instances the fine-tuned model would receive a lower score but capture more details relevant to the context of the document. In these situations the model generally received a lower score due to the summary having more words. Other instances showed that the fine-tuned model performed better at abstracting context from different sentences and finding situations to shorten the summary as compared to the baseline model. See example below.

Input:
There is diffuse subcutis space edema extending along the posteromedial aspect of the elbow adjacent to the medial epicondyle, extending to the olecranon process and along the superficial aspect of the epicondylo-olecranon ligament. There is no demonstrated solid, cystic or lipomatous mass lesion. There is enlargement with hyperintense signal of the ulnar nerve within the cubital tunnel. There is inflammation with mild laxity of the epicondylo-olecranon ligament. The combined findings are most consistent with a ulnar nerve neuritis possibly secondary to a subluxing ulnar nerve however the ulnar nerve at this time is within the cubital tunnel. There is no accessory muscle within the cubital tunnel. The common flexor tendon origin is normal. Normal ulnar collateral ligamentous complex. There is mild epimysial sheath edema of the pronator teres muscle consistent with a mild epimysial sheath sprain but no muscular tear. There is minimal intratendinous inflammation of the common extensor tendon origin consistent with a mild tendinitis. There is no demonstrated common extensor tendon tear. Normal radial collateral ligamentous complex. Normal radiocapitellum and ulnotrochlear articulations. Normal triceps and biceps tendon insertions. There is peritendinous inflammation of the brachialis tendon insertion but an intrinsically normal tendon

Base model summary:
there is enlargement with hyperintense signal of the ulnar nerve within the cubital tunnel. there is mild epimysial sheath edema of the pronator teres muscle consistent with a mild epimysial sheath sprain but no muscular tear.

Fine-tuned model:
Subcutis space edema along the posteromedial aspect of the elbow adjacent to the medial epicondyle, extending to the olecranon process. There is enlargement with hyperintense signal of the ulnar nerve within the cubital tunnel. The combined findings are consistent with a ulnar nerve neuritis possibly secondary to a subluxing ulnar nerve however the ulnar nerve at this time is within the cubital tunnel.

Recommendations

Experimenting with transfer learning on domain specific NLP problems like summarizing medical documents suggests that fine-tuning improves performance over a baseline model. That being said there is still considerable room for improvement in this case as it was fine-tuned on a relatively small data set. I would recommend finding more data to increase the variety of examples a model is trained on, as well as the model's ability to pick out important pieces of information.

Further Considerations

My initial idea for the project was to use the MIMIC-III data set to produce the training and testing sets. However, due to the need to complete training and apply for credentialing through Physionet I only used the MTSamples Data. In the future I would train a model using that data and compare the results to the model currently in use.

Furthermore, I would like to try and address the issue of document length by exploring the idea of generating summaries for different sections of text, combing the summaries and then resummarzing that to see how much more information and how relevant all of the information is.

A final thought would be to extend the number of tokens available and fine-tune/retrain embedding layers to see how this would affect the results.

Final Product

As a final component the fine tuned model was deployed as a backend inference model using Deployments on PaperSpace Gradient. A front-end client was developed and hosted using Streamlit.