Summarize without Direct Supervision: Extractive Summarization of Medical Notes using Weakly Supervised Learning

Stanford CS224N Custom Project

Hsu-Hang (Eric) Yeh
Department of Biomedical Data Science
Stanford University
ericyeh@stanford.edu

Abstract

Medical notes summarization is challenging because of the scarcity of labelled data. The goal of the project is to develop a method that scores each sentence based on its importance without human annotation data and select the top 10 percents of sentences to produce a succinct summary.

We proposed a weakly supervised scheme, where the model learns to predict a proxy target: near future procedure, and during this process also learns the importance of each sentence. We build an attention-based head model on top of ClinicalBert[1] that takes diagnoses as accessory inputs and predicts near future procedures, both of which can be automatically extracted from a medical database. Using the intermediate attention weights as importance scores, we demonstrated that this approach can boost performance of extractive summarization compared with baseline unsupervised learning, achieving an F1 score of 0.353 and an area under receiver-operator curve of 0.708 on the test set. This study opens new possibilities to other potential heuristics that may help circumventing the need of manually labelled data.

1 Key Information to include

- Mentor: Manan Rai
- External Collaborators: Sophia Wang

2 Introduction

Medical professionals need to read and process huge amounts of medical notes every day. Among various types of notes, progress notes are particularly common ones that are written by doctors to record the evaluation, discussion, and treatment decisions at each patient's visit. Given the time constraints of medical consultation, reviewing previous progress notes is a heavy workload for doctors, especially if the patient has many years of notes. Therefore, automatic summarization of notes that condense multiple documents into a single succinct summary would bring huge improvement in health care workflow.

In general, there can be two different approaches for text summarization: extractive and abstractive. Extractive summarization scores sentences based on their importance and directly select the high-score sentences for summary, whereas abstractive summarization generates new sentences from scratch after digesting the input documents. While abstractive summarization might produce more coherent paragraph, we focused on extractive summarization in this study because medical documents require high accuracy and correctness.

Despite the astonishing improvement in other medical natural language processing (NLP) tasks using deep-learning models[2], medical text summarization remains a challenge due to the lack of labelled training data. Reviewing medical notes and selecting reference summaries require tremendous time investment and domain expertise, and is unlikely to obtained by crowdsourcing for protection of patient privacy.

To circumvent this data scarcity problem, we aim to develop a summarization model without direct supervision by making it learn a different task, in which the labels are readily available without human annotation. This task is designed in a way that the model would need to rely on "valuable" sentences to perform well, and we can approximate the importance of a sentence by how much the model uses it for prediction. This can be viewed as weakly supervised learning where the label in the new task are a "proxy" of the true label. Similar approaches have been demonstrated successfully to generate summaries relevant to a specific query[3], and we aimed to extend this idea to produce general summaries that do not depend on any query word.

To achieve a general summarization task, we devised a different heuristic: predicting near future medical procedures, to infer the importance of a sentence. The model takes two inputs: the diagnoses and the notes, and predicts if any procedure will occur in the near future. This approaches has two advantages: (1) both diagnoses and procedures are well recorded, including the date and the type of them, and easily accessible in any electronic medical health records; (2) physician often use sentences that signal changes in a summary, and a need for procedure usually indicates the patient's condition worsens.

With near future procedures as proxy labels, we experimented with 5 different models with different attention and prediction heads and achieved 12% higher F1 score with our best model than the unsupervised term frequency-inverse document frequency (tf-idf) baseline. We also analyzed the advantages and drawbacks of this approach using generated samples. We hope this proof-of-concept study would inspire the future work on general summarization task of medical notes by designing more comprehensive proxy labels.

3 Related Work

Deep-learning based models have demonstrated good results with extractive summariazation. A model called SummaRuNNer[4], which is composed of a hierarchical recurrent neural network (RNN) with gated recurrent units (GRU), progressively encodes word-level embedding and sentence-level embeddings, and use the sentence embeddings to predicts if it belongs to summaries. NeuSum[5] uses similar GRU-based RNN to encode sentences and add additional attention mechanism at the sentence selection step which achieved better results. Recently, with more powerful transformer-based encoder, [6] proposed similar approach of sentence classifiers with BERT-encoded sentence vectors, achieved better results on DailyMail dataset. While building more powerful encoder seems to be promising, a key issue that prohibits the use of such models on medical texts is the need for large labelled data set.

To avoid the need of manually labelled data, Liu et al. [7] used the intrinsic correlation between medical notes to generate pseudo-labels for training. The key idea was that physicians tend to repeatedly record important terms, such as the diagnoses, in each note and these terms would have higher correlation with future notes. However, this approach has limited value in the summary of progress notes, in which doctors often copy and paste non-important information from last progress notes as well.

Another strategy is to use a query sentence or word to score the importance of each sentence, which is also called query-based summarization. In a recent study[3], McInerney et al. used query-based approach to extract sentences relevant for making diagnoses in radiology reports. They trained the model on a separate task of predicting future diagnosis and used the intermediate results to score the importance of sentences. This circumvents the need of manually labelled reference summary for training due to easily available diagnoses in any medical database. However, the generated summary is query-specific, which means using a different query word would produce

different summaries. We're inspired by this and decide to extend the idea to build a model that generates patient-specific summaries without relying on specific query words.

4 Approach

At a high level, the model is a scoring function that gives each sentence an importance score, given the diagnoses on the same day the sentence is written. After obtaining the scores for all sentences for a given patient, we select the top 10 percent highest scoring sentences as our summary. We used the occurrence of medical procedures in near future to weakly supervised the model to learn the importance score.

Let y_j be the label of the j-th note, which equals 1 if any medical procedure happens within 30 days of the note's date or 0 otherwise. Let n_j be the total number of sentences in j-th note, m_j be the number of diagnoses on the same date of the note, $S_j = \{s_{1,j}, s_{2,j}, ..., s_{n_j,j}\}$ be the set of sentences in j-th note, and $D_j = \{d_{1,j}, d_{2,j}, ..., d_{m_j,j}\}$ be the set of diagnoses of j-th note. Let $g(S_j, D_j)$ be the function that estimates the probability p_j of near future procedure of the j-th note, the objective of training becomes minimizing $\ell(\theta) = -\sum_j y_j \log(g(S_j, D_j))$. From the intermediate calculation of g, we can derive another scoring function f such that the importance score of $s_{i,j}$ is $f_{D_j}(s_{i,j})$.

Let $A = \bigcup_j S_j$ be the set of all sentences of a given patient. The ultimate goal is to find a subset A' such that

$$\{s_{i,j} : s_{i,j} \subset A'\} = argmax_{A' \subset A, |A'| = \lfloor \frac{|A|}{10} \rfloor} f_{D_j}(s_{i,j}) \tag{1}$$

4.1 Baselines

Four different baselines were explored. Since we fix the ratio of selecting sentences at 10 percents, the very basic baseline is to randomly sample 10 percents of sentences with equal probabilities.

The second baseline is to simply perform K-means with euclidean distance as metric on all sentence embeddings of ClinicalBERT[1] using $\lfloor \frac{|A|}{10} \rfloor$ as the number of centroids and extracting the sentences closest to the centroids. The [CLS] token embedding from ClinicalBERT was used as the representative embedding for the whole sentence. We used scikit-learn package of version 1.0.2 to perform K-means algorithm.

The third baseline is to use tf-idf encoding of sentences to compute the importance score. Each term T has a tf-idf score $tf(T, s_{i,j}) \cdot idf(T)$, where $tf(T, s_{i,j})$ is the frequency of T in $s_{i,j}$ and $idf(T) = \log \frac{(1+n)}{(1+df(T))} + 1$, where n is the total number of sentences and $df(T)$ is the document frequency of T. The larger the tf-idf score, the more important the term is. The score of a sentence is simply the summation of all tf-idf scores of each unique term in it: $\sum_{T \in s_{i,j}} tf(T, s_{i,j}) \cdot idf(T)$. We trained the TfidfVectorizer in scikit-learn library on the training data and encoded each sentence in the test set to get their importance scores.

The fourth baseline uses cosine similarity as scoring function, that is: $f_{D_j}(s_{i,j}) = \hat{s}_{i,j}^T \hat{D}_j$, where $\hat{s}_{i,j}$ is the [CLS] token embedding of $s_{i,j}$ from ClinicalBert and $\hat{D_j}$ is $\frac{1}{m_j} \sum_{k=1}^{m_j} \hat{d}_{k,j}$, the average of [CLS] embeddings of D_j.

4.2 Models

We explored 5 different models, which can be grouped into 2 classes: single-direction attention and bi-direction attention between diagnoses and sentences. The 2 classes of model are illustrated in Figure 1.

In **ClinicalBERT-Naive**, **ClinicalBERT-PL**, and **ClinicalBERT-DL** models, we used ClinicalBert as our encoder. The above models utilizes single-direction attention from direction to sentences. Using the same notation, we denote $\hat{s}_{i,j}$ as the [CLS] token embedding of $s_{i,j}$ from ClinicalBert and $\hat{d}_{k,j}$ as the [CLS] token embedding of $d_{k,j}$ from ClinicalBert. Let $\tilde{D}_j$ be the overall

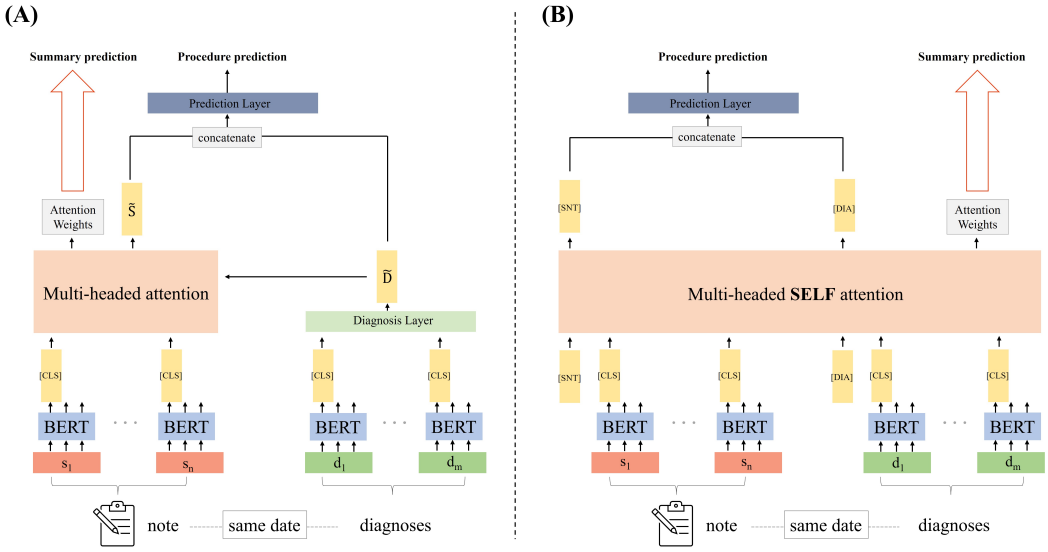
embedding of D_j after passing all $\hat{d}_{k,j}$ through a diagnosis layer. A multi-headed dot product attention of $\tilde{D}_j$ with respect to all $\hat{s}_{i,j}$ were performed to compute a weighted sum of values as output $\tilde{S}_j$. We then concatenate $\tilde{S}_j$ and $\tilde{D}_j$ and pass it through a prediction layer to compute $\hat{y}_j$. In **ClinicalBERT-Naive**, the diagnosis layer is simple averaging and the prediction layer is a single layer neural net.

The naive design apparently is too simple for the model the learn the complex relationship between the embeddings and the label, so in **ClinicalBERT-PL** we modifies the prediction layer into 2-layer neural net with residual connection and applies layer normalization on the input $\hat{s}_{i,j}$ and $\hat{d}_{k,j}$. In **ClinicalBERT-DL**, we tried to improve the expressive power of the diagnosis embedding, using a 2-layer neural net as diagnosis layer to processes each $\hat{d}_{k,j}$ and taking averages of the outputs to form $\tilde{D}_j$.

The next attempt is to improve the BERT encodings themselves. Because ClinicalBert is trained on notes in intensive care units, a place where ophthalmology patients rarely visit, we believed a domain-specific BERT model would have higher power in understand the texts. So in **OphBERT-PL** we replaced the encoder with a BERT model finetuned on ophthalmology notes as encoders, which was generously provided by our external collaborator [Tao, 2022, work in progress], and kept the head model the same as **ClinicalBERT-PL**. In the above models the scoring function $f_{D_j}(s_{i,j})$ is the average across different heads of the attention weights of $\tilde{D}_j$ with respect to $s_{i,j}$.

The single-directional attention did not leverage the full power of attention mechanism to understand the relationship between sentences and the diagnoses. In our last model **Transformer-PL**, we treated all $s_{i,j}$ and $d_{k,j}$ as continuous sequence separated by a [DIA] token and capped by a [SNT] token. The embeddings of the two introduced tokens are randomly initiated and are learnable parameters during training. The entire sequence is fed into a transformer-like model with 4 blocks of multi-headed dot product self-attention. The last hidden states of [SNT] and [DIA] are concatenated and passed through a prediction layer same as in **ClinicalBERT-PL** to make predictions. The scoring function $f_{D_j}(s_{i,j})$ is $w_{[SNT]} + w_{[DIA]}$, where w_X is the average across different heads of the attention weights of X with respect to $s_{i,j}$.

Figure 1: Model architecture. (A) single-direction attention model (B) bi-direction self attention model



5 Experiments

5.1 Data

Ophthalmology progress notes were extracted from Stanford Research Repository (STARR) database. A thousand patients who had underwent any ophthalmic procedures in their history were randomly sampled from the above data. Their de-identified IDs were randomly split into training, validation, and test sets of size 950, 50, 50 patients, respectively. This amounts to 13974, 974, 724 notes in each group, respectively. The diagnoses for each visit were also extracted from the database and matched with the notes on the unique ID for each encounter.

We wrote codes to clean each note by using regex patterns to search and remove de-identified brackets, numbers in the subheadings, the replacement character in unicode, and long spaces. Subsection titles, such as "Family History", were joined using a slash (e.g. Family/History) for easier sentence segmentation. We used python package scispaCy [8] 0.4.0 with model "en_core_sci_md" to split each note into sentences, resulting in 462130, 29855, 22604 sentences in train, validation, and test set, respectively.

5.2 Evaluation method

For each model, a 3-fold evaluation was done to help us analyze the performance of the model. We first checked if it performed well on the proxy task: (1) predicting near future procedures. This is a note-level binary classification problem where each note is assigned a probability. The area under receiver-operator curve (AUROC) was calculated from the predicted probabilities and the true labels in the test set. F1 score using a fixed threshold of 0.1 was also reported on the test set. Secondly, we randomly selected 5 patients from the test set and manually annotated 2 versions of summaries using their corresponding 870 sentences and tested the model prediction performance on (2) only procedure-related sentences and (3) general summary sentences. In (2), only sentences with clear indication of future procedures were labeled positive, whereas in (3) all sentences considered important for a summary were labeled positive. The two versions of summaries do not necessarily overlap. Notice that only (3) is what we're interested in ultimately, but (1) and (2) can help us understand why the model succeeds or fails. Both (2) (3) can be viewed as sentence-level classification. F1-score with threshold of 0.1 and AUROC were reported on the manually labelled sentences. We also included ROUGE score[9], a common metric to evaluate extractive summaries. ROUGE-n refers to the overlap of n-grams between predicted summaries and true summaries, and ROUGE-L refers to overlap of the longest common subsequence. The number of overlaps can be used to derive the precision, recall, and the corresponding F1 scores. The ROUGE scores were computed by python rouge-score package of version 0.0.4.

The baselines were only tested on our true target (3). Notice that AUROC for random selection and K-means were not reported because they don't output probabilities.

5.3 Experimental details

Each input sentence is tokenized by the Hugging Face Tokenizer of ClinicalBert with a maximum length of 128. Longer sentences were truncated and shorter sentences were padded to the maximum length. Due to the limitation of computation resources, we freezed all parameters in pre-trained BERT models and only trained the head model. All models were trained with an AdamW optimizer with beta1 of 0.9, beta2 of 0.95. Each batch of data contains all sentences and diagnoses of the j-th note.

Except the **ClinicalBERT-Naive**, all models were trained for 30 epoches using an exponential decay of learning rates from 0.001 with a decaying rate of 0.9 after each epoch, and the model with the lowest validation loss was saved. **ClinicalBERT-Naive** was only trained for 8 epoches with a fixed learning rate 3e-5 due to limitation of computation resources, but it reached a similar loss on validation set as other models. The number of heads in attention for all models was 8.

5.4 Results

The quantative results were summarized in Table 1. Naive K-means clustering only achieved F1 score of 0.105, which showed no much improvement from random selection. Using cosine similarities, the area under ROC curve (AUROC) was only 0.508 but the F1 score improved to 0.153, reflecting that sentences containing similar meaning as the diagnoses were more likely chosen for the summary. Tf-idf performed reasonably good, even outperforming two of our models, OphBERT-PL and Transformer-PL.

All single-direction models had moderate performance on both the near future procedure prediction and procedure-related summary prediction. Despite having no outstanding performance on the proxy task, the model did learn better to select important sentences. Both **ClinicalBERT-PL** and **ClinicalBERT-DL** outperformed the tf-idf baseline, with the latter achieving 0.353 in F1 score and 0.708 in AUROC. **ClinicalBERT-DL** also had highest ROUGE-1, ROUGE-2, and ROUGE-L F scores. This indicated that while the proxy task is hard to learn, the model was able to incorporate additional knowledge on which sentences to rely upon to make good predictions. Notice that the relatively low performance of procedure-related summary compared with general summary might seem counterintuitive, but this could be due to procedure-related summary contained much fewer positive sentences, only 54% of that in general summary and roughly 5% of all sentences.

Using a ophthalmology-specific BERT model surprisingly decreased the performance in summary extraction. Bi-direction self attention model **Transformer-PL** performed poorly on all 3 evaluation tasks, with F1 score on general summary prediction close to that of random selection.

Table 1. Quantative results of baselines and 5 models on 3 tasks: procedure prediction, procedure-related summary prediction, general summary prediction

	Procedure		**Summary - Procedure**		**Summary - General**				
	F1	AUROC	F1	AUROC	F1	AUROC	Rouge-1 F1	Rouge-2 F1	Rouge-L F1
Baseline									
Random	-	-	-	-	0.094	-	0.241	0.111	0.167
K-means	-	-	-	-	0.105	-	0.391	0.238	0.281
Tf-idf	-	-	-	-	0.235	0.567	0.398	0.293	0.334
Cos-similarity	-	-	-	-	0.153	0.508	0.312	0.224	0.263
Models									
ClinicalBERT-Naïve	0.491	0.719	**0.198**	0.632	0.216	0.596	0.366	0.236	0.278
ClinicalBERT-PL	0.474	0.776	**0.198**	**0.764**	0.294	0.690	0.431	0.326	0.362
ClinicalBERT-DL	0.487	0.737	0.147	0.676	**0.353**	**0.708**	**0.494**	**0.414**	**0.451**
OphBERT-PL	**0.515**	**0.787**	0.143	0.688	0.200	0.702	0.356	0.214	0.262
Transformer-PL	0.319	0.556	0.121	0.590	0.118	0.490	0.337	0.203	0.243

6 Analysis

We performed qualitative analysis by visual inspection of the selected summary sentences. Unfortunately, we can not present the examples here because they may contain sensitive patient health information. One interesting observation is that our models have tendency to select sentences from shorter notes. To visualize this behavior, in Appendix Figure A1 we showed that more sentences were selected from shorter notes in all our models. This phenomenon can be explained by our use of softmax scores as selection criteria. Because the softmax scores were calculated across sentences within a single note, a sentence in a shorter note (i.e. less sentences) was more likely to receive a high softmax score regardless of its actual information content. To solve this problem, we can devise a more complex scoring function that adds a penalty to a shorter note.

Adding more layers to both prediction and diagnosis layer has the best performance in our test set. We believed the diagnosis layer is especially important because the model learns better how to represent the ophthalmology diagnoses, which were rarely seen in original MIMIC data. Comparing the outputs between **ClinicalBERT-PL** and **ClinicalBERT-DL**, we noticed that **ClinicalBERT-PL** somtimes made mistakes by attending to some unrelated sentences, such "Family", whereas **ClinicalBERT-DL** more consistently selected diagnoses-related sentences.

Surprisingly, **Transformer-PL** model, which takes sentences and diagnoses embeddings as inputs, did not perform well, even on procedure prediction. From the outputs we observed that the model erroneously attend to personal information of patients, such as the number of children and the birth place. This can partially explain why the model also performed poorly on the precedure prediction task. One possible cause is that the model is more complicated and contains more weights than the others, and our training data was not big enough, resulting in ineffective learning.

Providing domain-specific BERT model did not seem to improve the performance. The OphBERT model reportedly did not perform better than ClinicalBert on text classification task [Tao, 2022, work in progress]. A possible explanation is that the size of ophthalmology notes corpus might not be large enough for the BERT model to learn effective weights. Turning to Bert models trained on scientific literatures, instead of only clinical notes, might be a reasonable choice. While literatures may lack special abbreviations commonly used for clinical notes, they presumably contain more ophthalmology-related words than MIMIC data.

A limitation of this approach is that the model attends to some sentences related to procedure but not relevant to general summary, such as saying that the patient has signed the informed consent. This sets the upper limit of how this approach can perform on extractive summarization. To avoid this problem, we plan to devise other heuristics in the future that teaches the model to put more attention on patient conditions instead of focusing only on procedure itself.

7 Conclusion

In this study, we devised a weakly supervised learning strategy that uses near future procedures as proxy labels to learn the importance of sentences in medical notes. We demonstrated that the model was able to learn importance scores of sentences using this approach, which circumvents the need of manually labelled data. This could bring inspirations on how to approach this task with other heuristics. We plan to add more heuristic that help the model learn more precise scoring functions and also expand the reference summaries to evaluate the robustness of the model.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[2] Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, et al. Neural natural language processing for unstructured data in electronic health records: a review. *arXiv preprint arXiv:2107.02975*, 2021.

[3] Denis Jered McInerney, Borna Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem Meent, and Byron C Wallace. Query-focused ehr summarization to aid imaging diagnosis. In *Machine Learning for Healthcare Conference*, pages 632–659. PMLR, 2020.

[4] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[5] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. Neural document summarization by jointly learning to score and select sentences. *arXiv preprint arXiv:1807.02305*, 2018.

[6] Yang Liu. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019.

[7] Xiangan Liu, Keyang Xu, Pengtao Xie, and Eric Xing. Unsupervised pseudo-labeling for extractive summarization on electronic health records. *arXiv preprint arXiv:1811.08040*, 2018.

[8] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

A Appendix (optional)

Figure A1. Distribution of note sizes from which sentences are selected