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

Hsu-Hang (Eric) Yeh, MD[1];

[1]Department of Biomedical Data Science, Stanford University, Palo Alto, CA

Problem

- Medical professionals need to read and process huge amounts of medical notes every day. Automatic summarization of notes that condense multiple documents into a single succinct summary brings huge benefits.
- Transformer-based models achieves good performance on text summarization, but requires human annotated data, which is rare in medical notes.

Backgrounds

- Weakly supervised learning can be used to solve the data scarcity problem.
- In a recent study[1], McInerney et al. trained the model on a separate task of predicting future diagnosis and used the intermediate results to score the importance of sentences.
- However, the summary is query-specific. Different queries produce different summaries.

Methods

- We devised a different heuristic: predicting near future procedures, to make the model learn to score the **general** importance of sentences.
- The idea is: if the model relies on some sentences to infer the near future procedure, the sentence might contain important information for summary.
- Objective of training: $\ell(\theta) = -\sum y_j \log(g(S_j, D_j))$. where y_j is the near future label of the j-th note, $S_j = \{s_{1,j}, s_{2,j}, ..., s_{nj,j}\}$ is the set of sentences in j-th note, and $D_j = \{d_{1,j}, d_{2,j}, ..., d_{mj,j}\}$ is the set of diagnoses of j-th note, $g(S_j, D_j)$ is the function that estimates the probability p_j of near future procedure of the j-th note
- 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 subset A' such that

$$\{s_{i,j}: s_{i,j} \in A'\} = argmax_{A' \in A, |A'| = \frac{|A|}{10}} f_{D_j}(s_{i,j})$$

- Data:
  - 1000 ophthalmology patients were randomly extracted from Stanford Research Repository (STARR) database.
  - Their de-identifi ed 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. Notes were cleaned and segmented into sentences.
  - The diagnoses for each visit were also extracted.

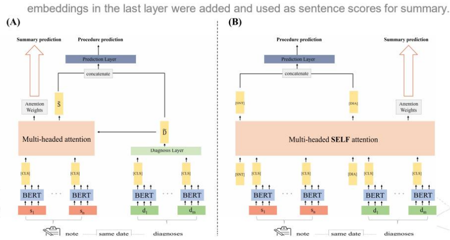
Experiments

Baselines

- Random selection of 10 percents of sentences
- K-means on ClinicalBert[2] sentence embeddings and choose the sentence closest to the center
- Term frequency-inverse document frequency (tf-idf) embedding of a sentence. Tf-idf is a score for each term that represents its importance. The sum of tf-idf score in a sentence is used as sentence scores.
- Cosine similarity between ClinicalBert diagnoses embeddings and each sentence embedding as sentence scores

Models

- **Inputs: Bert embedding of sentences of a note and diagnoses on the same date, Outputs: procedure logit and attention weights to all sentences**
- **(A) single-direction attention**
  - ClinicalBERT-Naïve: diagnosis ClinicalBert embeddings are averaged and prediction layer is a single-layer neural net
  - ClinicalBERT-PL: prediction layer has two layers, residual connection, and layer norm on nputs
  - ClinicalBERT-DL: in addition to PL, diagnosis layer is a two-layer neural net
  - OphBERT-PL: change ClinicalBert to OphBert [Tao, 2022, work in progress] that is trained on opthalmology notes
- In (A) the attention weights of $\bar{D}$ with respect to all sentence embeddings are used as sentence scores for summary.
- **(B) bi-direction attention**
  - Transformer-PL: concatente all sentences and diagnoses in single sequence and introduce [SNT] and [DIA] token, whose last hidden states are used for prediction.
- In (B) the attention weights of [SNT] and [DIA] with respect to all sentence embeddings in the last layer were added and used as sentence scores for summary.



Results

- 3 different evaluation: (1) procedure prediction, (2) pure-related summary (only sentences related to procedure is selected), and (3) general summary (our primary goal) – (1)(2) are only for analysis purpose and (3) is of primary interest
- Both ClinicalBERT-PL and ClinicalBERT-DL outperformed the tf-idf baseline. ClinicalBERT-DL also had highest ROUGE-1, ROUGE-2, and ROUGE-L F1 scores.
- Despite having no outstanding performance on the proxy task, the model did learn better to select important sentences.

	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

Analysis

- The models have tendency to select sentences from shorter notes.
  - This phenomenon can be explained by our use of softmax scores because a sentence in a shorter note (i.e. less sentences) was more likely to receive a high softmax score.
- 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].
  - Can be due to the small size of ophthalmoogy notes corpus.

Conclusions and Discussion

- Weakly supervised learning strategy that uses near future procedures as proxy labels can help the model learn the importance of sentences in medical notes.
- 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

References

1. 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.
2. 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.