An Automated Approach to Identify Key Phrases in Patient Notes

By: Jinyoung Hur, Taewan Kim, Qiwei Lin
University of Chicago

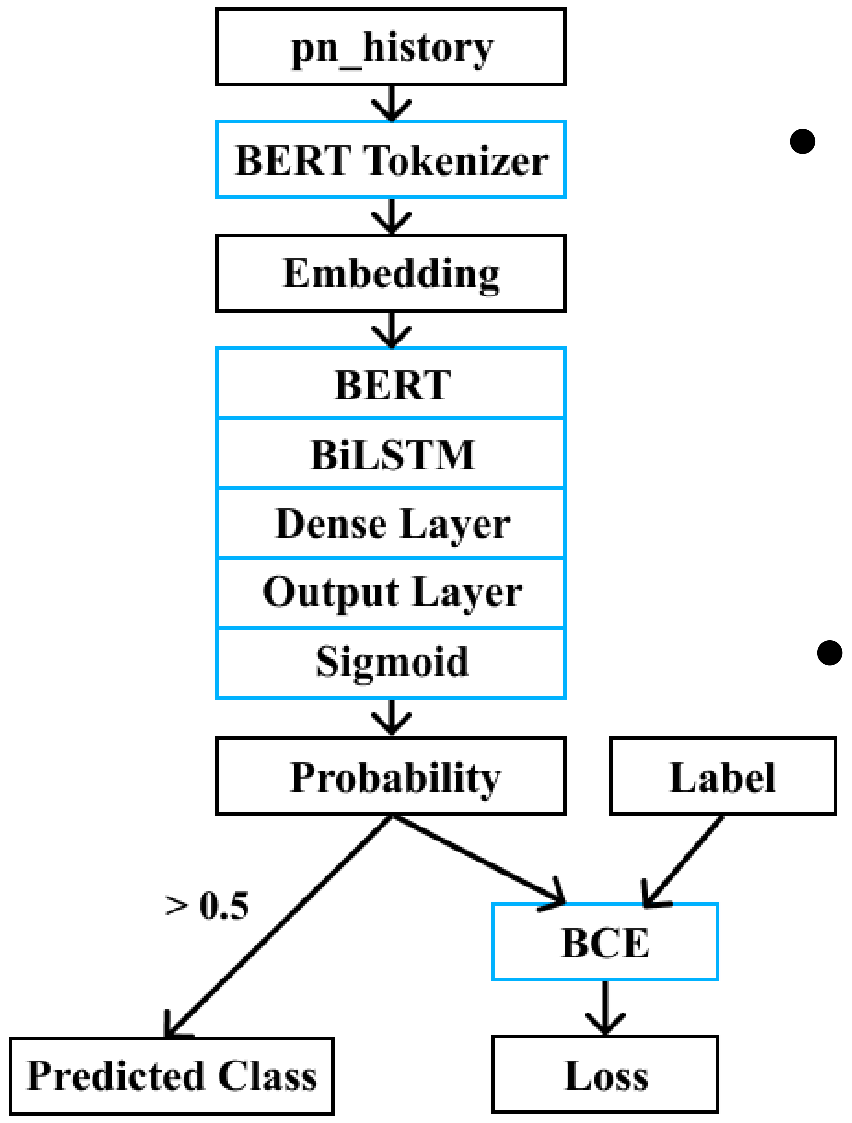
Overview

Natural Language Processing (NLP) can be applied to identify important medical concepts in patient notes, reducing a significant amount of time spent by physicians and serving as a more transparent and interpretable evaluation method for medical trainees. We develop an automated approach based on domain-specific BERT and Bi-LSTM that achieves satisfactory performance on mapping clinical concept features to words expressed in clinical patient notes.

Objective

- Preprocess a collection of patient notes from Step 2 Clinical Skills examination data from National Board of Medical Examiners (NBME)
- Train token-level classifiers using patient note and clinical case feature data to extract locations of medical phrases within a patient note.
- Compare various models to select the one with the highest micro-averaged F1 score

Methods

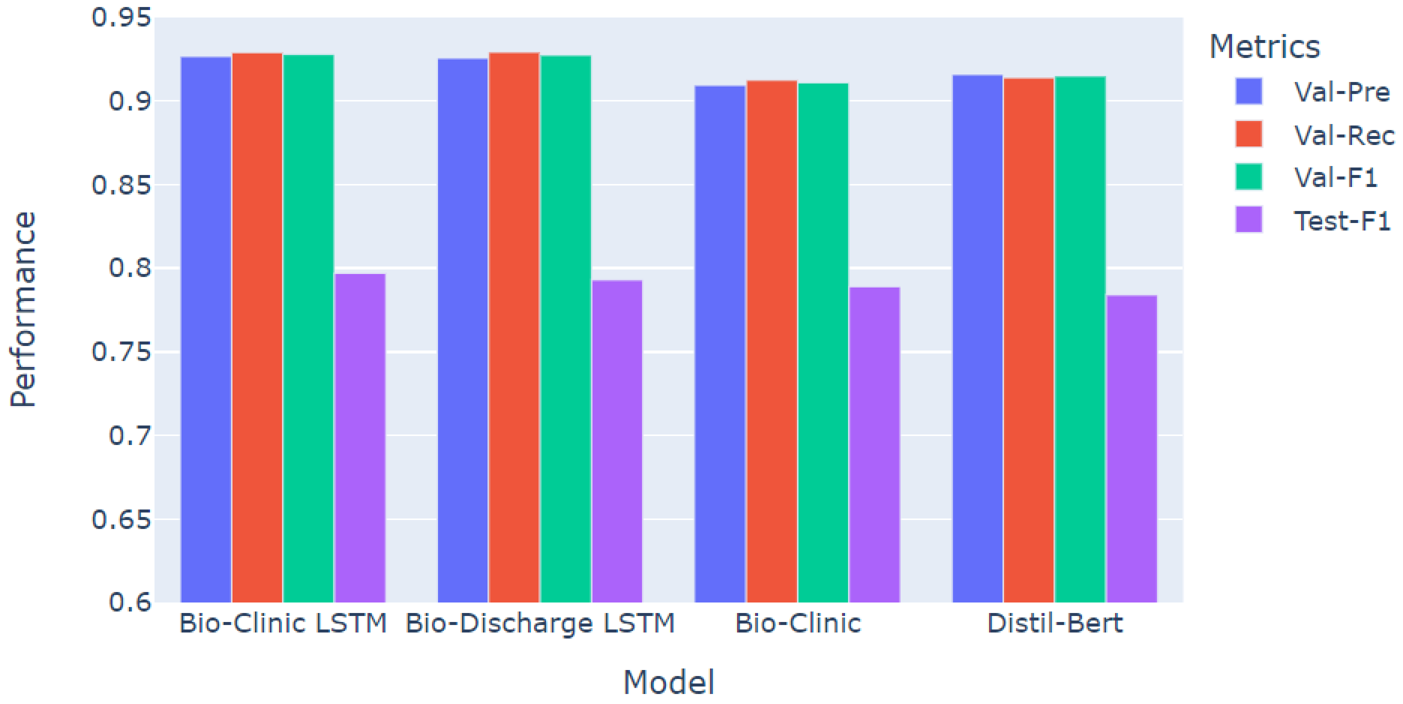


- For simpler models, Bi-LSTM, dense layers and Relu layers are removed, dropout is increased to 0.5.

- Hyperparameters:
  - LSTM-layers: 1
  - Hidden Size: 256
  - Dropout: 0.1
  - Learning rate: 1e-4

Results

Model Performance



- Bio-Clinical BERT + Bi-LSTM model has the highest F-1 score on the validation **(0.928)** and test set **(0.797)**. But it is lower than the state-of-the-art result **0.89**.
- Large performance gaps indicate over-fitting
- Simpler Models without Bi-LSTM achieves comparable performance within fewer epochs
- Using domain-specific BERT only boosts the performance by a small margin, compared to BERT pre-trained on general corpora

Table: Bio-Clinical BERT by Case on Validation Set

Case #	Precision	Recall	Micro-F1
0	0.851	0.919	0.884
1	0.916	0.914	0.915
2	0.937	0.935	0.936
3	0.937	0.945	0.941
4	0.919	0.948	0.933
5	0.922	0.834	0.876
6	0.932	0.974	0.952
7	0.926	0.969	0.947
8	0.963	0.941	0.952
9	0.930	0.954	0.942

- The model has a relatively lower performance on case 0 and 5
- For features in case 5, the performance is low on Associated-nausea (Recall = 0.38, F1 = 0.53) and Associated-throat-tightness (Recall = 0.61, F1 = 0.75)

Error Analysis of the Best Model

Ground Truth Annotation

2 week follow up **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** following ER visit **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** for palpitations. Labs **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** drawn at ER visit were **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** shown to be within normal limits **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** . Her first occurrence of palpitations **EPISODES OF HEART RACING** was 3 weeks ago and seemed to come on suddenly with no known trigger.

Prediction Annotation

2 **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** week follow up **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** following ER visit for palpitations. Labs **EPISODES OF HEART RACING** drawn at ER visit were shown to be within normal limits. **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** Her first occurrence of palpitations **EPISODES OF HEART RACING** was 3 weeks ago and seemed to come on suddenly with no known trigger. Palpitations **EPISODES OF HEART RACING** were first occurring once every few days but now occur more often

Models are likely to make mistakes on:

- Annotations with gaps in between
- Annotations with highly correlated trigger words

Conclusion

- Using BERT pre-trained on medical corpora and Bi-LSTM achieves around **0.8** F-1 score on the test data. Future work can focus on making the performance more generalizable.
- Medical phrases with negation and large heterogeneity in expression remain hard to predict
- Future work could incorporate unlabeled data selected by active learning to increase the number of training instances (currently 1000 patient notes).

References

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

2. Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings.

3.Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining.

4. Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2021. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction.