Combining Transfer Learning and Structured Medical Domain Knowledge for Clinical Semantic Textual Similarity

David Chang[1], Eric Lin, MD[2], Cynthia Brandt, MD, MPH[3], Andrew Taylor, MD[4]

[1]Computational Biology and Bioinformatics Program, Yale University, New Haven, CT; [2]Department of Psychiatry, Yale University; [3]Yale Center for Medical Informatics, Yale University; [4]Department of Emergency Medicine, Yale University

Introduction

Semantic textual similarity (STS) is a natural language understanding task that assesses the ability of natural language processing (NLP) systems to estimate the semantic equivalence of two snippets of text on a predefined scoring scale. While recent advances in NLP and deep learning have led to significant performance improvements in STS and related tasks in the general English domain, such progress has only begun to transfer over to the clinical domain. Drawing from these recent advances, we developed a system for the task of clinical STS for the 2019 n2c2/OHNLP Shared-Task challenges.

Methodology

Our model architecture combines a text encoder based on BERT and a GCN-based graph encoder, with a final linear layer to generate a score between 0 and 5. Figure 1 shows a diagram of the system.

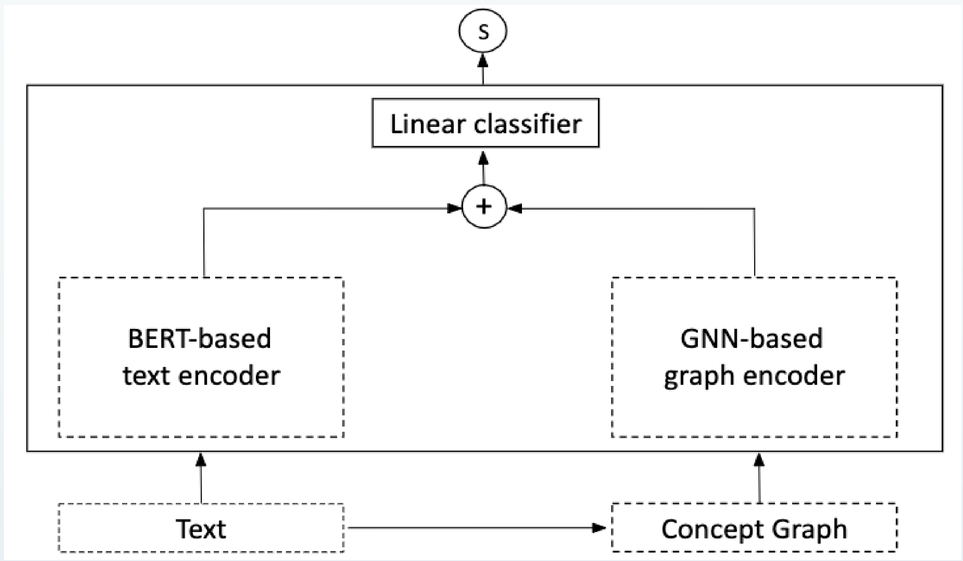


Figure 1. Graphical summary of overall system. Text is used to generate corresponding concept graphs. The text encoder and the graph encoder process the inputs separately, and their outputs are concatenated for the final scoring layer.

We use MetaMap to extract clinical concepts from the text and use their relations from the UMLS Metathesaurus to construct a concept graph for each sentence. Given the low precision of MetaMap, manual review was conducted to exclude candidate concepts that were spurious. The text is passed to the text encoder, and the corresponding graphs are passed to the graph encoder. The outputs of the encoders are concatenated and passed to a linear layer to produce the final score for clinical semantic similarity. We also devised a simple data augmentation strategy to append strings from the MetaMap output to the input text as a way of adding clinically relevant information. We used knowledge distillation and an ensemble based on both clinical and general variations of BERT (MT-DNN, NCBI BERT, Clinical BERT) for the final submissions.

Results

An ablation study for our system (Table 1) showed that all of those components contributed positively to the final performance to varying degrees. We produced our final three submissions in the following way: an ensemble of models trained on the 5-fold split datasets; an ensemble trained on all available training data; and an average of the two. The best performance of 0.8784 Pearson correlation came from the third submission (Table 2). It is worth noting that most of the performance improvement relative to the base BERT model came from ensembling. Considering the limited size and quality of the data, it is difficult to make claims about the general effectiveness of any of the components used. Following attempts by a clinical expert and a non-clinical expert to manually score the test set, we came to appreciate the model's ability to perform as well as it did.

Model Configuration	Pearson Corr.
TE	84.35
TE + GE	84.87
TE + GE + DF	85.15
TE + GE + DA	85.05
TE + GE + DF + DA	85.25
TE + GE + DF + DA + EN	87.51
TE + GE + DF + DA + EN + KD	87.86

Table 1. Ablation study showing contributions of text encoder (TE), graph encoder (GE), discriminative fine-tuning (DF), data augmentation (DA), ensemble (EN), knowledge distillation (KD).

Submission	Pearson Corr.
Ensemble with 5-fold splits	87.76
Ensemble with all data	87.78
Average of both	**87.84**

Table 2. Results for the final three submissions.

Conclusion

We present a system for clinical STS based on transfer learning and the incorporation of structured medical domain knowledge using GNNs. We developed the system with minimal hyperparameter tuning and relied more on ensembling to mitigate the small and noisy nature of the dataset. Overall, the clinical STS track of the 2019 n2c2/OHNLP Shared Tasks was an opportunity to narrow the gap between general domain NLP and clinical NLP.

Acknowledgements

This project was made possible by NIH Training Grant 5T15LM007056-33 and NIMH Training Grant 2R25MH071584-12.

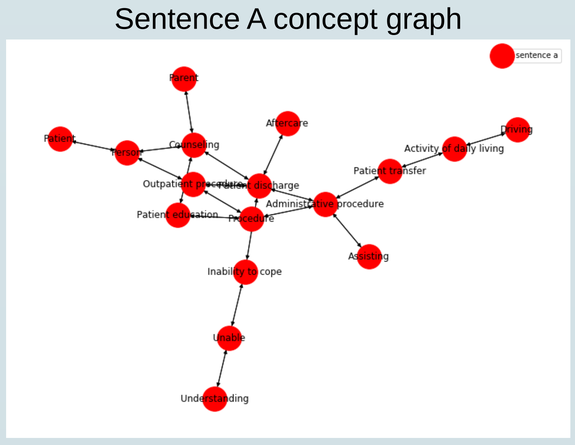
Data Visualization

Sentence A: patient discharged to home, ambulating without assistance, family driving, accompanied by parent, above person's verbalized understanding of discharge instructions and follow-up care

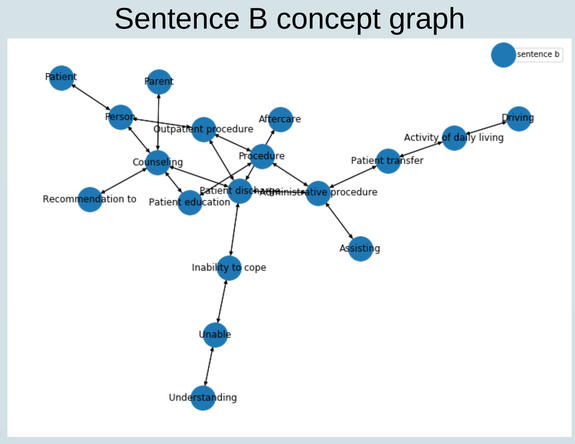
Sentence B: patient discharged to home ambulating without assistance, family driving, accompanied by parent, discharge instructions given to patient, above person's verbalized understanding of discharge instructions and follow-up care

Similarity score: 4.75

Sentence A concept graph



Sentence B concept graph

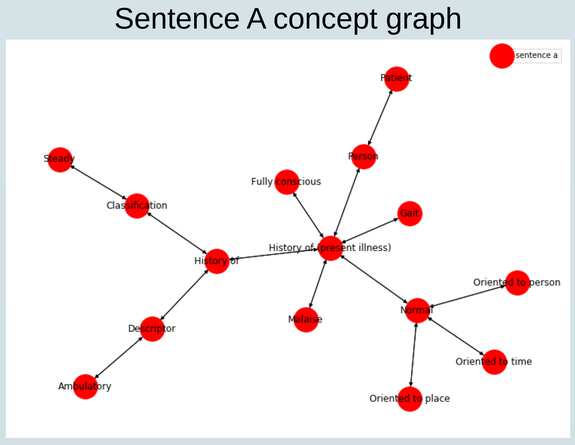


Sentence A: patient arrives ambulatory, gait steady, history obtained from patient, patient appears generally ill, patient cooperative, alert, oriented to person, place, and time

Sentence B: patient arrives via hospital wheelchair, unsteady gait, inability to ambulate, history obtained from patient, patient appears comfortable, patient cooperative,, alert, oriented to person, place, and time

Similarity score: 3.15

Sentence A concept graph



Sentence B concept graph