Efficiency of Various Embedding Techniques in Clinical Notes

Problem Statement

Expand on different embedding techniques like GloVe, ELMo, ClinicalBERT etc. by implementing these on our dataset, and evaluating their performance by using these embeddings in a predictive model.

Background

Most word embeddings are represented in the Euclidean space, which sometimes makes them unable to capture hierarchical structure observed in certain corpora (or they may require high dimensional embedding dimensions in order to capture this complexity). Clinical notes are temporal in nature, which makes Word2vec, GloVe, and FastText, that learn language through windows of context hard to apply to data that exhibits long-term dependencies. A good example of this if a treatment plan at the end of a long paragraph is related to symptoms mentioned at the beginning, those methods may not be able to capture it. Also, if a patient has multiple clinical notes that depict progression of a disease, due to the time-series nature of the notes the sequence might also not be captured.

Clinical Notes present a specific challenge in terms of developing a model including very high dimensionality, sparsity, and complex linguistic and temporal structure. To mitigate this, we need to develop efficient representations on these clinical notes at the patient-level. This could possibly also reduce the time one spends in feature engineering tasks related to complex sequential unstructured data.

Dubois et. al. preformed 3-level evaluation on clinical notes: word-level, note-level, patient-level. Table below shows word-level comparison of various embedding methods.

Embedding method	May-Treat (%)	May-Prevent (%)
GloVe300-W10-R1	7.83	8.51
GloVe-100-W7-R2	6.01	08.09
GloVe-300-W4-R2	8.81	**10.64**
GloVe-300-W7-R2	8.25	10.21
GloVe-500-W7-R2	9.23	10.21
GloVe-300-W10-R2	**10.49**	9.79
GloVe-300-W4-R3	6.57	7.23
MCEMJ	8.25	6.81
Cross-channel	5.45	2.55
MaxGRU200-MCEMJ	7.27	5.53
MaxGRU300	1.26	0.43

In their paper they note that it would be interesting to develop a new expression of the GloVe or word2vec objective function that takes into account the specific structure of these notes (sample negated words during negative sampling, sample windows from the set of words at each iteration, etc.). They also suggest exploring other options for the RNN supervision.

In an ideal situation the labels would be less sparse and would only use the note's content; so perfect labels would be some very high-level aggregation of the concepts present in the note. In our project we would like to build upon some of the findings presented in the papers referenced below.

References

1. https://www.sciencedirect.com/science/article/pii/S2590177X19300563
2. Efficient Representations from Clinical Text- https://arxiv.org/pdf/1705.07025.pdf
3. Learning Effective Embeddings from Medical Note
   https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2744372.pdf
4. Contextual Embeddings from Clinical Notes Improves Prediction of Sepsis
   https://www.medrxiv.org/content/10.1101/2021.03.02.21252779v1.full

Extra

Modeling high-cost tasks like ER, ICU stay durations using in-hospital mortality and stay duration using clinical notes. We can also frame this problem as co-morbidity problem using MTL
https://www.nature.com/articles/s41467-021-20910-4

http://tjn.mit.edu/pdf/whats-in-a-note.pdf
MTL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568068/
https://psb.stanford.edu/psb-online/proceedings/psb19/ding.pdf