A Transformer-based Machine Learning Framework using Conditional Random Fields as Decoder for Clinical Text Mining

Lifeng Han[1], Valerio Antonini[2], Ghada Alfattni[1], Alfredo Madrid[1], Warren Del-Pinto[1], Judith Andrew[1], William Dixon[1], Meghna Jani[1], Ana María Aldana Gelves[1], Robyn Hamilton[1], Karim Webb[1], and Goran Nenadic[1]

[1]The University of Manchester, UK & [2] Dublin City University, Ireland

Abstract

Clinical Natural Language Processing (NLP) methods are increasingly used in different healthcare applications, including identification of drug exposure, disease severity progression, etc. While the majority of published models use statistical modelling or LSTMs, here we explore a new clinical NLP framework that uses state of the art Transformer models in combination with Conditional Random Fields as decoder. We also use graph-based label propagation and test the framework for the drug information extraction task.

Project Overview

Interest in NLP methods is increasing in many healthcare applications, partially due to encouraging results from various shared tasks, such as those organised by i2b2 and n2c2. The extracted information includes diseases, symptoms, drugs, effects, and relationships between them. Proposed methods most often include conventional high-performing statistical models such as conditional random fields (CRFs), and neural models as Recurrent Neural Networks (RNNs) and Bi-directional Long-Short Term Memories (Bi-LSTMs) [1]. While the accuracy of such machine learning models have reached a level notably higher than previous methods, there is still a gap compared to human performance. For instance, a recent evaluation on drug relation extraction using Bi-LSTM+CRF shows that the *micro* and *macro* accuracy suffers from lower recall (0.84+) in comparison to higher precision scores (0.93+) [1]. Transformer-based neural models have been proved to achieve better learning capacity and have shown better results than LSTMs in other NLP applications such as machine translation, but they have not been widely used in processing electrical health records [2, 3]. We explored using the state-of-the-art neural NLP Transformer models to replace the Bi-LSTM as the model encoder (see Figure 1). To further address the data needs for tasks where the available training data with manual labels is scarce, we explored using a graph-based label propagation method to extend the labelled data for model training using lexical similarity based graph construction [4, 5]. The results will be demonstrated using the standard shared task data (n2c2) on both drug entity recognition and relation extraction.

Accepted to HealTAC2022 as Poster Presentation
https://healtac2022.github.io/programmes/

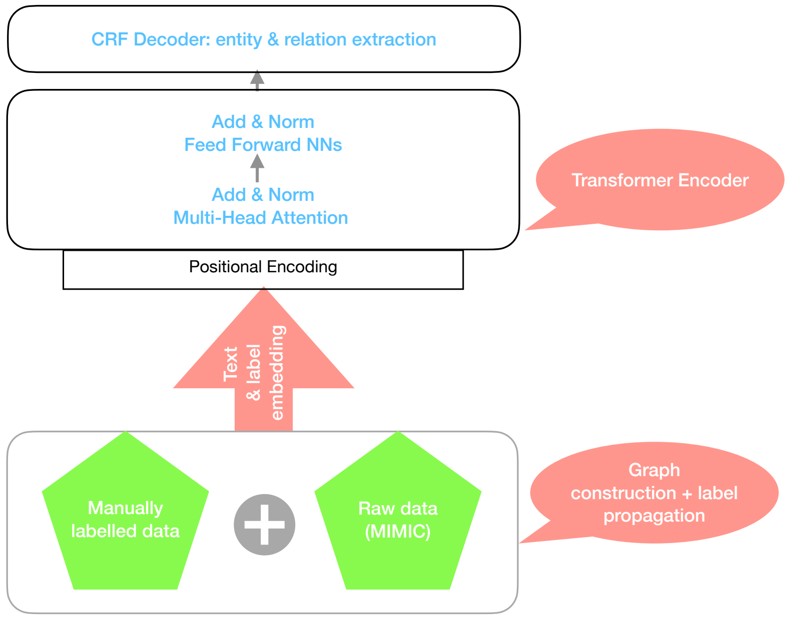

Figure 1: Transformer Encoder + CRF Decoder with Graph Construction and Label Augmentation for Training Data Augmentation (*Appendix*).

References

[1] Alfattni G, Belousov M, Peek N, Nenadic G, et al. Extracting drug names and associated attributes from discharge summaries: Text mining study. JMIR medical informatics. 2021;9(5):e24678.

[2] Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. JMIR Med Inform. 2019 Sep;7(3):e14830.

[3] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. .

[4] Han ALF, Zeng X, Wong DF, Chao LS. Chinese named entity recognition with graph-based semi-supervised learning model. In: Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing; 2015. p. 15-20.

[5] Zeng X, Wong DF, Chao LS, Trancoso I. Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2013. p. 770-9.