Poster number: 19
e: lifeng.han@manchester.ac.uk

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

Lifeng Han[3], Valerio Antonini[4], Ghada Alfattni[1,3], Alfredo Madrid[1,3,5], Warren Del-Pinto[1,3], Judith Andrew[1,3], William G. Dixon[1,2], Meghna Jani[2], Ana María Aldana[2], Robyn Hamilton[2], Karim Webb[2], Goran Nenadic[3]
[1]Centre for Epidemiology Versus Arthritis, University of Manchester, UK; [2]Department of Rheumatology, Salford Royal Hospital, Salford, UK; [3]School of Computer Science, The University of Manchester, UK; [4]Dublin City University, Ireland; [5]Rheumatology Department, and Health Research Institute (IdISSC), Hospital Clínico San Carlos, Madrid, Spain

HealTAC 2022: the 5th Healthcare Text Analytics Conference

Motivations

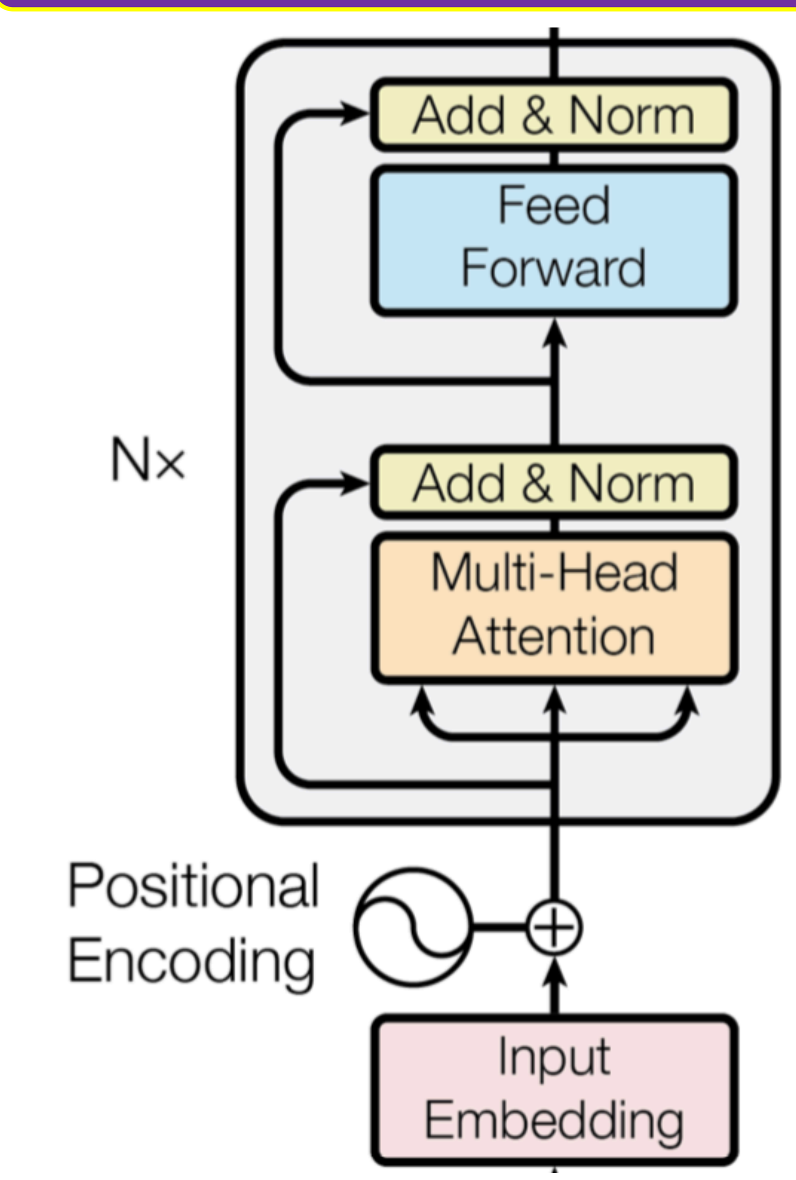
Clinical Natural Language Processing (NLP) methods are increasingly used in different healthcare applications, including identification of drug exposure, disease severity progression, relation extraction, etc.

However, the majority of published models use either statistical modelling or neural network based models such as LSTMs. To take advantage of the strength from both paradigms and further improve the model performances in this domain, we explore a new clinical NLP framework that uses state of the art Transformer neural models as encoder in combination with Conditional Random Fields (CRFs) as decoder.

To overcome the data scarce issue where the manually annotated clinical data is hard to acuqire, we propose to use graph-based label propagation method to extend labelled dataset for model learning. We will test the framework for the drug information extraction task from n2c2 challenges.

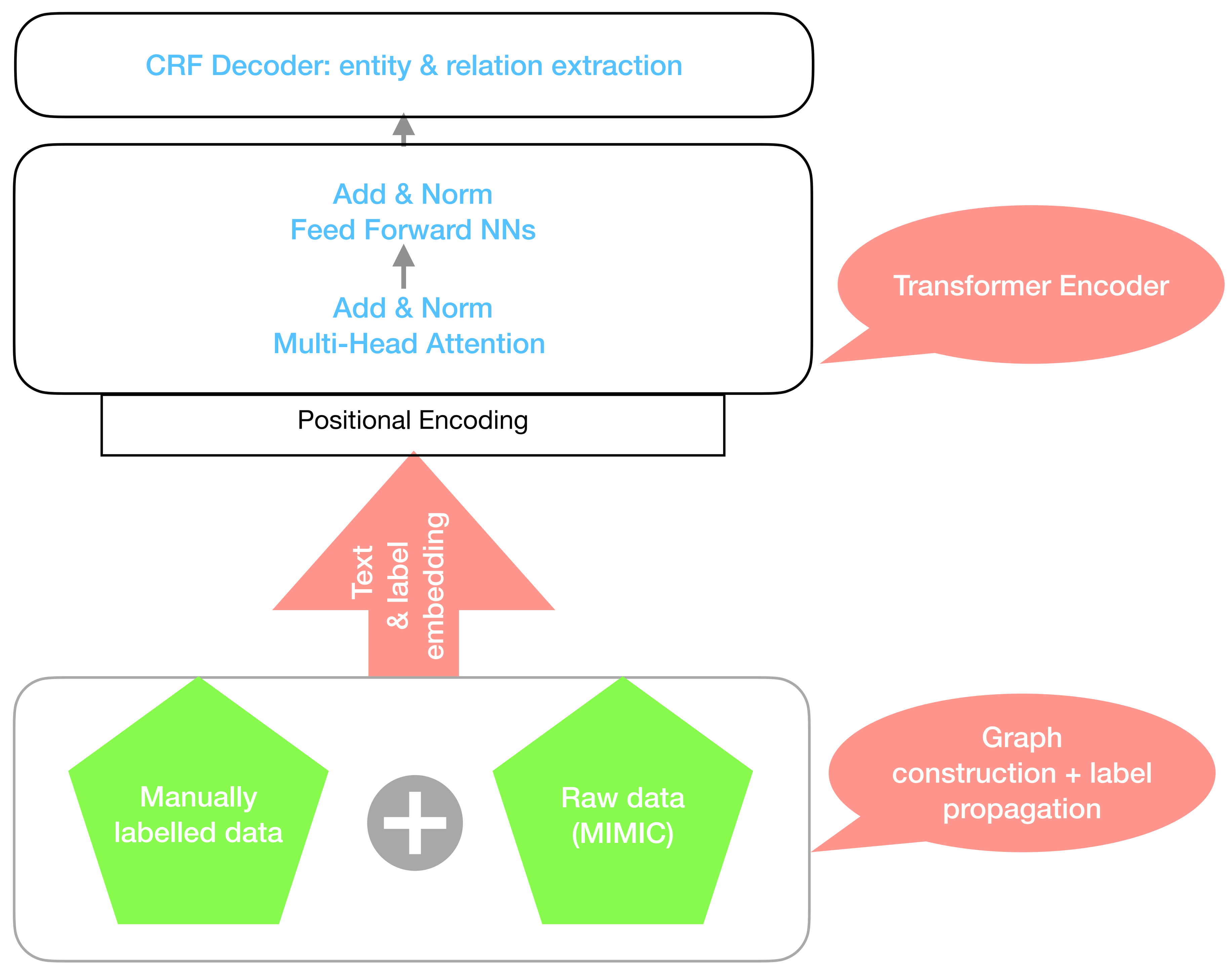
Git: https://github.com/poethan/TransformerCRF

Transformer Components



- Input embedding includes both text and label embedding.
- Positional encoding is a replacement of positional ruction included in other neural networks such as RNN and CNN.
- Each encoder layer of Transformer (Vaswani et al. 2017) includes a multi-head attention and fee-forward network, both having a linear normalisation afterwards.
- The output of Transformer encoder will feed directly into CRF input

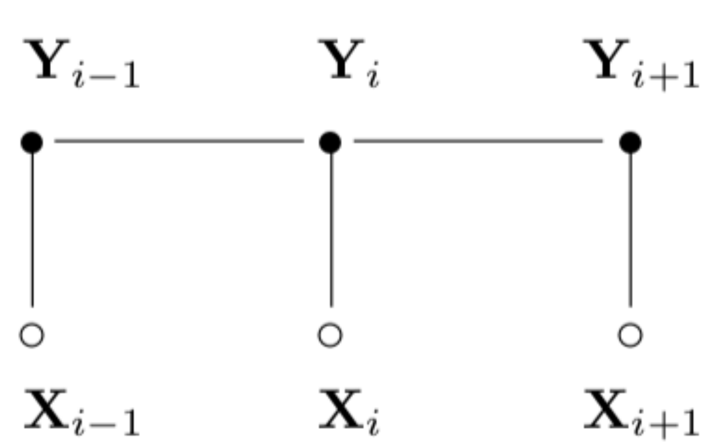
Methodological Design



CRF Components

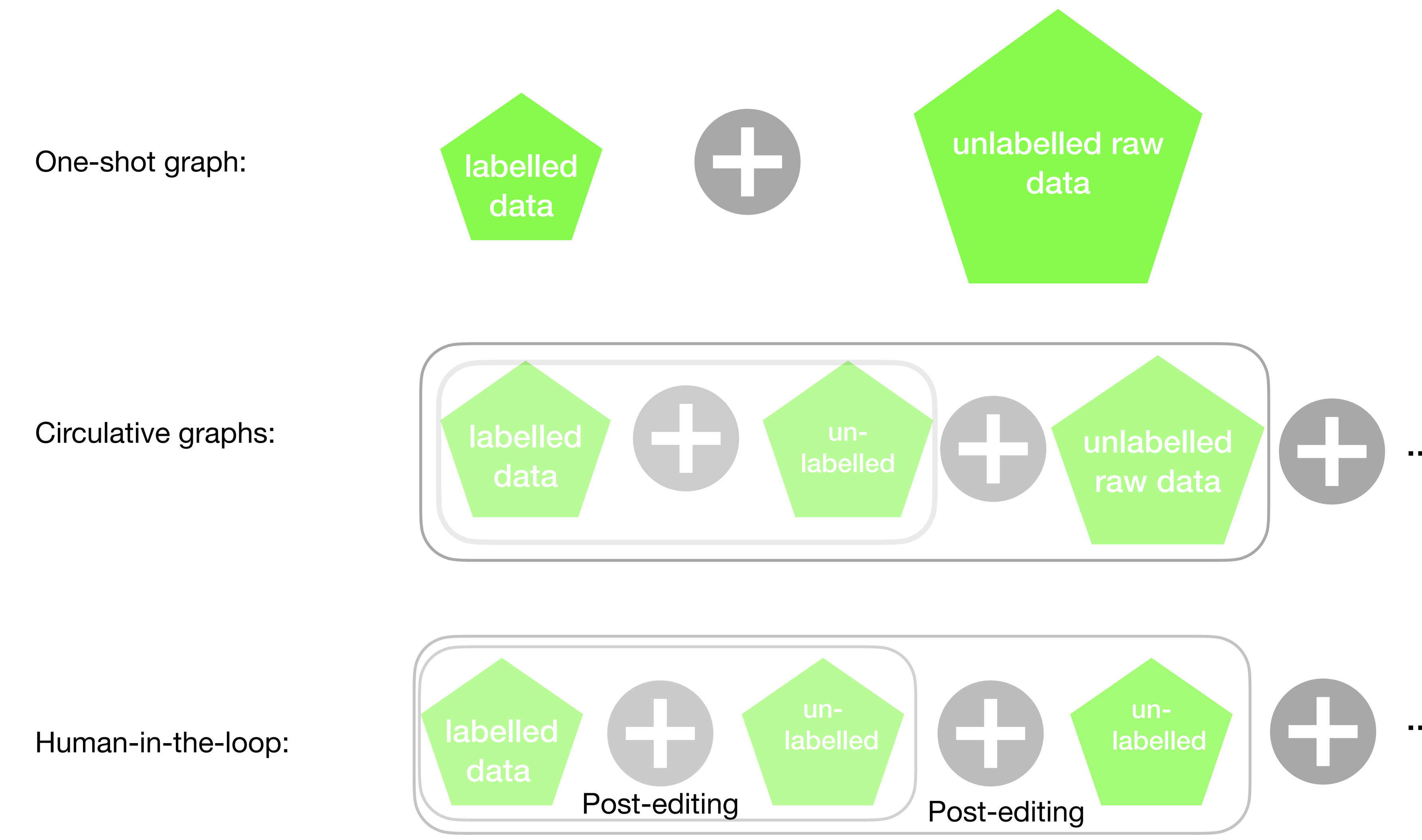
An application from fundamental theorem of random fields, CRFs are un-directional graph models firstly designed for sequence segmentation and labelling by Lafferty et al. (2001)

Assuming X is the current lexicon and Y is the label of X to be predicted, the value of Y can be conditioned on both lexicon X and neighbouring labels e.g. using *'i'* as a variable representing current position:



Different from statistical NLP, in this neural model application, the input of X will be embedded word vectors instead of symbolic lexicons.

GBSSL Components: Graph-construction + Lable Propgation



We design different strategies for graph-based semi-supervised learning (GBSSL) to take advantages of both high-quality annually labelled data and large amount of raw data.

- One-shot graph: this is a straight-forward graph construction connecting all labelled and unlabelled lexicons using their lexical similarities. Then using graph based label propagation to project the labelling tags from labelled lexicons to the raw data.
- Circulative graphs: each time step we build a graph with the same size of existing labelled data (original and newly acquired) and raw data.
- Human-in-the-loop post-editing: we integrate the fixed number of raw data at each time step, and carry out human post-editing to correct the auto-labelling errors. We time-step moves on, the errors are expected to reduce to a tolerable level for the model.

References

Lafferty et al. (2021) CRFs: "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" ICML2001 https://dl.acm.org/doi/10.5555/645530.655813

Vaswani et al. (2017) Transformer: "Attention is All you Need" NIPS2017 https://papers.nips.cc/paper/2017

Han et al. (2015) GBSSL: "Chinese Named Entity Recognition with Graph-based Semi-supervised Learning Model" in SIGHAN2015 https://aclanthology.org/W15-3103.pdf (our paper)