BioNLP 2019 Project Proposal Deep C. Patel, Akshay Shah

Our project is to tackle the i2b2 2010 challenge. The challenge is basically to extract medical entities from unannotated clinical texts such as medical problems, treatments, and tests. The task can be categorized under the umbrella of classic NLP task known as Named Entity Recognition and comes with a fair share of challenges such as difficulty to collect expertise based rich annotated medical texts required for state of the art deep learning techniques to give productive results.
The following text briefs us about the current methods to perform entity extraction in medical text. In the traditional machine learning/deep learning approach, Bi-directional LSTM combined with Conditional Random Fields (CRF) works well on the medical text as they employ both word and character level bidirectional LSTMs thus giving out more in-depth contextual representation. However, BiLSTM has been seen to struggle because of polysemous words [1], resulting in crucial false positives, for instance, BiLSTM CRF mistakenly labels the gene name “BRCAI” as a disease entity, because there are disease entities such as “BRCAI-abnormalities” or “brcai-deficient” and so on [1]. In order to tackle false positives, [1] CollaboNet uses the architecture of multiple BiLSTM-CRF models where each model is regarded as a single-task model (STM) and each STM is trained on the particular dataset and considered as an expert on particular entity type. These experts then transfer knowledge to other experts. With a purview to address the lack of training data in biomedical text, the concept of transfer learning is used in the form of contextualized language representations as discussed in the ELMo [2] and BERT [3]. BERT is based on the concept of multi-layer bi-directional transformer which is very significant when trained on two tasks, namely, pre-trained randomly masked tokens and predicting if two sentences follow each other. BioBERT [4] is built on top of BERT to build a model using annotated medical corpora like PubMed, PMC, or combination of both and can take the shape through various approaches using fine-tuning embeddings by leveraging transfer learning. Currently, BioBERT claims to be SOTA for the 2010 i2b2/va challenge entity extraction task with an 86.84 F score.

Augmented memory architectures such as NTM [5] and DNC [6] have been developed by researchers. They were shown improving the learning performance on several tasks that were previously learned using Recurrent Neural Networks. The unique thing about these augmented memory architectures is that, unlike RNNs, they read and write in external memories. It gives them the advantage of working with larger memories which can be manipulated selectively. In contrast, RNNs have temporary memory and do not provide selective read and write operations on it. Going one step further, the DNC also records the temporal linkages between each entity written into the memory. During reading from the memory, this temporal relation can be exploited to establish the sequential relation between entities in both forward and backward direction, thus working analogously to BiLSTMs.

As mentioned previously, the BiLSTMs have struggled with polysemous words because one word only has one representation in the static embeddings [7]. However, when the word is compared against the context during the runtime, it can reveal its true sense and the improvements can be made. Considering the temporal linkages being recorded by the DNC during runtime, we hypothesize that the problem of polysemous words can be tackled by it. Looking at the above features of the neural architectures with augmented memories, we think that they could improve the results of medical NER task when compared with solving it using BiLSTMs + CRF. Therefore, we propose to solve the i2b2 2010 medical entity recognition challenge by using the DNC.

References

1. Yoon, W., So, C. H., Lee, J., & Kang, J. (2019). CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. BMC bioinformatics, 20(10), 249.
2. Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).
3. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
4. Lee, Jinhyuk, et al. "Biobert: pre-trained biomedical language representation model for biomedical text mining." arXiv preprint arXiv:1901.08746 (2019).
5. Graves, Alex, et. al. “Neural Turing Machines.” arXiv preprint arXiv:1410.5401v2 [cs.NE] (2014).
6. Graves, A, et. al. “Hybrid Computing Using a Neural Network with Dynamic External Memory.” Nature 538, 471-476 (2016).
7. Xiaohai, https://talkai.blog/2019/03/06/from-bow-to-bert/ (2019).