Medical Name Entity Recognition using Hybrid LSTM and Rule Based Technique

Minhah Saleem, Amal Muhammad Saleem, Azka Rehman, Muhammad Usman
5F, Wooil Bldg., 623, Gangnam-daero,Gangnam-gu, Seoul, Korea 06524

Abstract

We purpose a hybrid rule base LSTM model for medical name entity recognition. The architecture consists of a CRF based BiLSTM neural network with rule based False positive reduction. To incorporate unknown word during inference we used FastText model [1]. The complexity of network is carefully selected so that model learns a generalised pattern from training set and provide reliable and consistent results on test set

Introduction

Recognizing Named Entities in medical records is not same as other Name Entities recognition tasks because medical records have their own format of storing information about patient's disease, and previous diagnosis. Wording used in medical records is a lot different from other text documents and many abbreviations such as CXR, PA, mg and alot of chemical notations are used which are difficult to understand even for non professionals. The publicly available data for medical NER is quite low as compared to other NER datasets. Since data is low we cannot use any complex network because it would overfit and we will not be able to make relaibile inference. We cannot use transfer learning because there is a lot of difference between medical records and other text documents. So we have to design a network by carefully selecting it's parameter so that network learns a generalized pattern from training data.

Archtecture

Pipeline consists of a FastText model [1] with the embedding vector of size 300 and moving window of length 15. It is used to incorporate unknown words during inference. Main deep learning network includes a tokenization layer, an embedding layer with embedding size of 450, a Bi directional LSTM with 150 units and a time distributed Dense layer of 100 units. Finally a CRF layer with 3 unit to predict medicine, non medicine, and medicine related word. Network is optimised with CRF loss. After this we performed False positive reduction using a list of common non medical English words and compared predicted annotation with this list.

Experimentation

We have experimented with different architectures including BERT [2], BioBERT [3], attention LSTM [4], convolutional LSTM [5], and Bi Directional LSTM [6]. These networks are trained and tested on dataset provided by competition organizers. BiLSTM give the highest F1 score among these networks. So we modify BiLSTM architecture to further increase it's performance. We have tested different loss function for optimising the network and CRF loss was giving best performance. The number of parameters for networks are optimized so that network does not overfit or underfit. Complexity of network is kept minimum without underfitting so that network learns the general distribution of the training dataset. We train our network for 200 epochs because around 150 epoch model learning starts to converge. We have used word level embedding in our network because output labels are for words and make it easier for the network to learn word to word relation. Word level embedding has an issue that if an unknown word appears during inference tokenizer will not have any token to represent that word and it pass and empty string to represent that word. It affects the performance of the network. We have used the FastText model[1] to replace unknown words with most similar meaning word which we can tokenize. After analysing results, we found that deep learning model was generating some false positives which were very easy to detect like bat, teeth. So a list of common non medical English words was generated manually. We used a rule based technique in which we compared each medicine entity predicted by model with the list of common English words. If the predicted word is in that list we change its annotation to non medicine.

Conclusion

The performance criteria for this competition was to F1 score. Our hybrid model achieved f1 score of 92.499. The main advantage that our network have is consistency in results because we have not used any complicated network which tends to overfit or have partially learned parameters due to early stopping.

References

1. Athiwaratkun B, Wilson AG, Anandkumar A. Probabilistic fasttext for multi-sense word embeddings. arXiv preprint arXiv:180602901. 2018.
2. Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nature cell biology. 2008;10(5):593-601.
3. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.
4. Jin Y, Xie J, Guo W, Luo C, Wu D, Wang R. LSTM-CRF neural network with gated self attention for Chinese NER. IEEE Access. 2019;7:136694-703.
5. Cho M, Ha J, Park C, Park S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. Journal of biomedical informatics. 2020;103:103381.
6. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:150801991. 2015.