Biomedical Named Entity Recognition with Attention

Hao Wu
University of Illinois at Urbana-Champaign
Urbana, IL
haow11@illinois.edu

Peilun Zhang
University of Illinois at Urbana-Champaign
Urbana, IL
peilunz2@illinois.edu

Abstract

Biomedical named entity recognition (BioNER) is a specific sequence to sequence (seq2seq) task. Given a sequence of words, the goal is to obtain the best or most possible sequential labels, such as genes, chemicals and diseases. Currently, most sequence labeling methods heavily rely on Bi-directional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF) models. Transformer that utilizes attention mechanism shows promising results in handling seq2seq problems, which provides us with insights to modify the BiLSTM-CRF models to utilize attention mechanism to conduct BioNER. We propose an Attention-CRF model to selectively encode the representation of sequential words, and recognize the sequence of labels. We evaluate the performance of our learning framework on 15 BioNER datasets compared with state-of-the-art BioNER frameworks and baselines. We further compare and analyze the time complexity of BiLSTM-CRF and Attention-CRF as seq2seq models.

1 Introduction

With more and more literature produced in the biomedical area, building an efficient toolkit to automatically extract knowledge from documents becomes increasingly necessary. To handle this requirement, BioNER plays an important role that aims to recognize biomedical entities, such as genes, chemicals, and diseases without too many human efforts. Besides, it is also beneficial for many downstream applications, like relation extraction, event detection and document summarization.

Generally, BioNER is regarded as a sequence labeling task, which means we should assign the best label sequence to the input word sequence. Traditional methods often require handcrafted features to locate specific entity types. For instance, we can frequently see the suffix '-ase' more in proteins than diseases [Habibi et al., 2017]. However, generating such amount of domain-specific features for BioNER system requires domain knowledge and cannot be directly adapted to recognize new entity types. Rather than manually designing entity-specific features, recent works mainly focus on developing an automatic learning algorithm to extract latent features of sequential words to feed a conditional random field (CRF) [Lafferty et al., 2001] layer to predict the output label sequence. Habibi et al. [2017] followed the suggestion from Lample et al. [2016] and proposed BiLSTM-CRF model to predict the label sequence with completely agnostic knowledge to every type of the entity. This neural network model only requires golden dataset with entity labels and pre-trained word embedding from large, domain-specific corpus (e.g. all PubMed abstracts). Wang et al. [2018] improves the ability of BiLSTM-CRF network with multi-task learning by sharing BiLSTM layers .

Currently, with the fast development of attention mechanism (Vaswani et al. [2017], Bahdanau et al. [2014], Luong et al. [2015]), researchers work on building new model architecture for feature encoding that eschews recurrence but instead relies on only the attention mechanism to encode the sequential information about input corpus. Vaswani et al. [2017] proposed one novel sequence transduction model, called Transformer, based solely on attention mechanism. Lee et al. [2019] applied

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

a pre-trained biomedical language representation model based on BERT architecture [Devlin et al., 2018] which performs well in three representative biomedical text mining tasks, including BioNER.

In this paper, we combine the idea of BiLSTM-CRF [Habibi et al., 2017] model and self-attention mechanism [Vaswani et al., 2017], by replacing feature encoding layer of BiLSTM with a self attention layer [Vaswani et al., 2017] to represent character level information and word level information. Outputs of those layers are then still fed to a CRF layer to do the final label sequence predictions. We call it Attention-CRF model. This method completely eschews the sequential computation, and only needs to parameterize the attention assigned on each unit (word or character) to encode the unit information. We evaluate and compare the performance of Attention-CRF with other 4 neural network architecture in 15 benchmark biomedical dataset used by [Crichton et al., 2017].

2 Background

2.1 Problem definition of Name Entity Recognition (NER)

Given a word sequence $\mathbf{w} = \{w_1, w_2, ..., w_n\}$, we should assign the best label sequence $\mathbf{l} = \{l_1, l_2, ..., l_n\}, l_i \in L$ to each word separately, where L includes every possible labels according to IOBES schemes (see 4.1). For example, if we are given a sentence "Selegiline - induced postural hypotension...", the best label sequence should be like "S-Chemical O O B-Disease E-Disease...". Because in this sentence, "selegiline" is a kind of chemical and "postural hypotension" is a disease, whereas other words are out of chunks.

2.2 Long Short-Term Memory (LSTM)

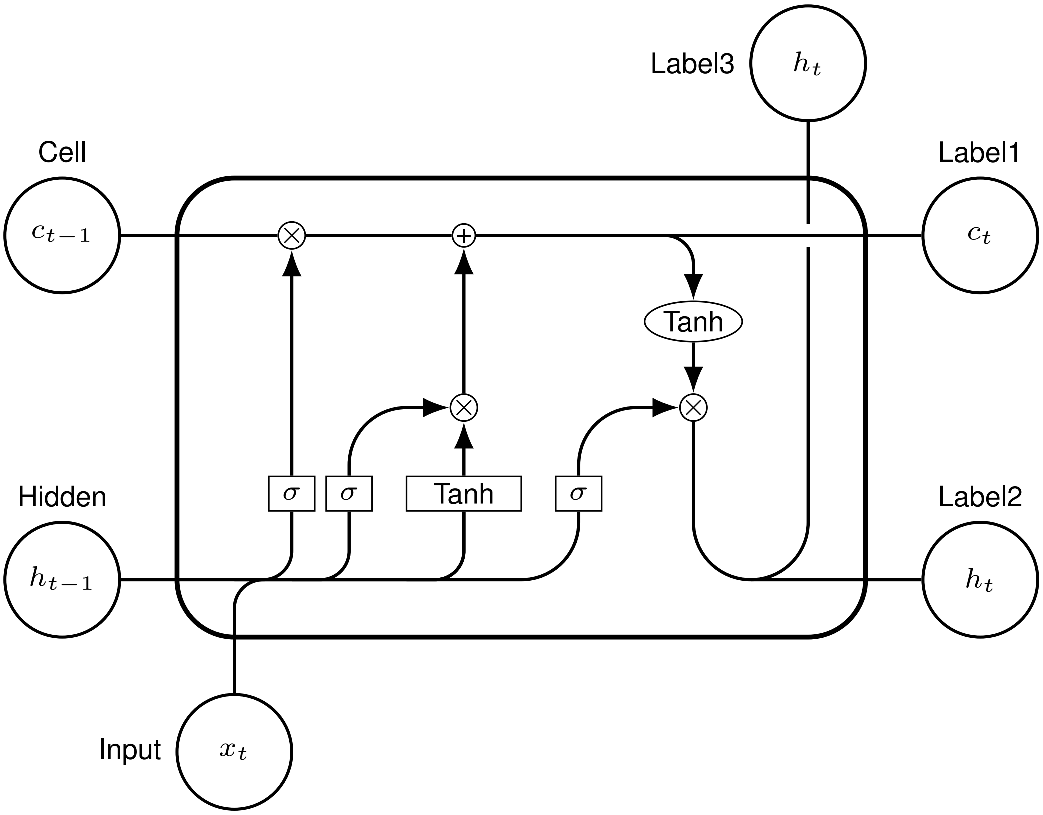

Figure 1: Architecture of Long Short-Term Memory Neural Network.

Recurrent neural networks (RNNs) are a kind of neural networks that conduct operations upon sequential data. The input to a RNN is a sequence of vectors $(x_1, x_2, ..., x_t)$, where each vector x_i is a representation of an element in the input sequence. The output generated by the network is another sequence $(h_1, h_2, ..., h_t)$ that represents some information about sequence at each step, where each vector h_i is a hidden state vector. Although RNNs can learn long dependencies, previous work has also found out that RNNs, in practice, appears to be biased towards the most recent inputs [Bengio et al., 1994] because when training a RNN using back-propagation, the gradients can tend to zero or infinity due to the usages of numbers with finite precision.

Long short-term memory (LSTM) neural network is a designated type of RNN that aims to model dependencies between sequence elements [Hochreiter and Schmidhuber, 1997]. Figure 1 illustrates an architecture of LSTM neural network. At step t of recurrent calculation, c_t represents memory stored in cell, h_t represents the output hidden state, x_t is the input vector, σ denotes element-wise sigmoid function, $tanh$ denotes element-wise hyperbolic tangent function, $+$ and $\times$ denotes element-wise summation and multiplication. At step t of the recurrent calculation, the inputs of the network are x_t, $c_{\text{t-1}}$, $h_{\text{t-1}}$ and the outputs of the network are c_t, h_t. Previous work [Wang et al., 2018] uses the following implementation:

$$\begin{aligned}
i_t &= \sigma(W^i x_t + U^i h_{t-1} + b^i) \\
f_t &= \sigma(W^f x_t + U^f h_{t-1} + b^f) \\
o_t &= \sigma(W^o x_t + U^o h_{t-1} + b^o) \\
g_t &= tanh(W^g x_t + U^o h_{g-1} + b^g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot tanh(c_t)'
\end{aligned}$$

where i_t, f_t and o_t are referred as input, forget, and output gates. Initially, h_0 and c_0 are zero vectors and the parameters to train are W^j, U^j and b^j for $j \in \{i, f, o, g\}$.

2.3 Bi-directional Long Short-Term Memory (BiLSTM)

LSTM model computes a sequence of hidden state vectors that represents the previous context of the sentence at each word. It is intutive that the subsequent context can also provide useful information.

Inspired by this idea, Bi-directional Long Short-Term Memory (BiLSTM) improves the LSTM model by feeding inputs twice with different directions, one in the original direction, the other in the reversed direction [Graves and Schmidhuber, 2005]. Then, the outputs from each directions are aggregated together as the final output, which captures the dependencies from not only the previous elements but also the subsequent elements of the sequence.

2.4 Bi-directional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF)

A very simple tagging model can use the output hidden state vectors as features to make independent tagging decisions. However, a limitation of this strategy appears when the dependencies across the labels are very strong. NER is a typical example of such task because the grammatical structure has significant affects on the labels but the structure will not be modeled with independent assumptions. It is useful and important to model the dependencies across output labels. Previous work utilizes a conditional random field to conduct this task and has shown a promising result [Lafferty et al., 2001]. For an input sequence, $X = (x_1, x_2, ..., x_n)$, and output sequence $y = (y_1, y_2, ..., y_n)$, a score is defined as:

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}$$

where n is the length of the sequence, k is the number of distinct labels, P is an $n \times k$ matrix of the output from the BiLSTM layer, $A_{i,j}$ is the transition probability from label i to label j. There are two additional labels that represents *start* and *end* of the sequences and therefore A is a matrix with dimension $(k + 2) \times (k + 2)$.

The training process maximizes the log-probability of the correct label sequence:

$$log(p(y|X)) = s(X, y) - log(\sum_{\widetilde{y} \in Y_X} e^{(s(X, \widetilde{y}))})$$

where Y_X are defined as all possible label sequences for a input sequence X. During the prediction stage, the model predicts the output sequence by:

$$y^* = \underset{\widetilde{y} \in Y_X}{\operatorname{argmax}}\, s(X, \widetilde{y})$$

2.5 Attention

Inspired by the observations that certain alignment exists between the input sequence and output sequence, which means that each step of generating a token is greatly related to a certain part of the input sequence, attention mechanism aims to allow the model to refer back to the input sequence [Young et al., 2018]. An attention network keeps a set of hidden state representations that scale with the size of the input sequence. The model performs selection over the representations to allow the model to maintain a variable-length memory [Kim et al., 2017].

For input sequence $(x_1, x_2, ..., x_n)$, let z be a categorical latent variable with sample space {1, 2, ..., n} that represents the selection among input, and let q be a query. The mechanism aims to generate a context c with the input sequence and a query, by accessing to a distribution $z{\sim}p(z|x,q)$ that can be derived from applying softmax function to vectors of alignment scores, where it conditions p on the input x and a query q. Then, the context over a sequence is defined as:

$$c = \mathbb{E}_{z \sim p(z|x,q)}[f(x, z)]$$

where $f(x, z)$ is a compatibility function. Two types of compatibility functions are majorly used, dot-product compatibility function [Luong et al., 2015] [Vaswani et al., 2017] and multi-layer perceptron compatibility function [Bahdanau et al., 2014].

3 BioNER Model with Attention

4 Experiment

4.1 Dataset

The datasets we use in this paper to evaluate the performance of BioNER models were introduced by Crichton et al. [2017]. They provided 15 datasets focusing on biomedical name entity types. These datasets contain several entity types about biomedical and they are all publicly accessible. The details of datasets above are listed in Table 1, including the name of datasets, name entity types, and entity counts.

Following the data splitting metrics of Lample et al. [2016], we separate the datasets into three sections: training section, development section , and test section. We use training section to optimize our neural networks, whose hyperparamters are tuned on development section. And the performance of the model will be evaluated in the test section. The training set will include most data items, around 70% of all data items, and the development and test set will contain around 10% and 20% datasets respectively. The size of development section should not be too large, because it is used to tune the hyperparameters by many times of repeated experiments.

The sequence of labels is encoded by IOBES schemes. According this schemes, each word can be annotated with 5 kinds of label. The IOBES format is following IOB format (inside-outside-beginning). It is a tagging format for tagging tokens from sequences.

- I: I- prefix indicates the tag is in the chunk
- O: O- prefix indicates the tag is outside a chunk
- B: B- prefix means the tag is the beginning the chunk and is followed by tags with same entity types
- E: E- prefix means the tag is the end of the chunk
- S: S- prefix indicates single tag of the chunk

Table 1: The datasets and details of their annotations

Dataset	Entity Types	Entity Counts
AnatEM	Anatomy NE	13,701
BC2GM	Gene/Protein NE	24,583
BC4CHEMD	Chemical NE	84,310
BC5CDR	Chemical, Disease NEs	Chemical: 15,935; Disease: 12,852
BioNLP09	Gene/Protein NE	14963
BioNLP11EPI	Gene/Protein NE	15,811
BioNLP11ID	4 NEs	Gene/Protein: 6551; Organism: 3471; Chemical: 973; Regulon-operon: 87

BioNLP13CG	16NEs	Gene/Protein: 7908; Cell: 3492;
		Cancer: 2582; Chemical: 2270;
		Multi-tissue structure: 857; Tissue: 587;
		Cellular component: 569; Organ: 421;
		Organism substance: 283;
		Pathological formation: 228;
		Amino acid: 135; Organism: 1715;
		Immaterial anatomical entity: 102;
		Organism subdivision: 98;
		Anatomical system: 41;
		Developing anatomical structure: 35

BioNLP13GE	Gene/Protein NE	12,057

BioNLP13PC	4 NEs	Gene/Protein: 10,891;
		Chemical: 2487; Complex: 1502;
		Cellular component: 1013

CRAFT	6 NEs	SO: 18,974; Gene/Protein: 16,064;
		Taxonomy: 6868; Chemical: 6053;
		CL: 5495; GO-CC: 4180

Ex-PTM	Gene/Protein NE	4698

JNLPBA	5 NEs	Gene/Protein: 35,336;
		DNA: 10,589; Cell Type: 8639;
		Cell Line: 4330; RNA: 1069

Linnaeus	Species NE	4263
NCBI-Disease	Disease NE	6881

4.1.1 Dataset Benchmarks

4.2 Pre-trained Word Embedding

Moen and Ananiadou [2013] created a bunch of word vectors trained from the entire available biomedical scientific literature, a text corpus of over five billion words. It includes three separate data sources: (1) Abstracts form PubMed dataset (PubMed); (2) Full-text documents from PubMed Central (PMC); (3) English Wikipedia dump (Wiki).

Habibi et al. [2017] studied the impact of different word embeddings trained from different combination of data sources listed above. They reported that the Wiki-PubMed-PMC embeddings achieve the best performance in all their evaluations. This dataset provides a set of word vectors with 200 dimensions induced on a combination of PubMed and PMC texts with texts extracted from a recent English Wikipedia dump [Moen and Ananiadou, 2013].

4.3 Evaluation Metrics

All evaluation results are conducted on each test dataset. Wang et al. [2018] mentioned *exact match* is one solution to evaluate the prediction, which means only that the predicted entity type and entity boundary are the same as the ground truth will be considered as "matched". After that, we calculate the precision, recall, micro-F1 and macro-F1 scores for each dataset independently to check the performance of each model on different dataset. Moreover, we also provide these metrics for each name entity to show the performance of each model on handling different entity types. We want to figure out whether each model is sensitive to the specific entity types. For example, if the entity count is small, the model will fail to capture the information about that entity type and give a bad result.

We compare the set of false positives (FPs) and false negatives (FNs) of the different BioNER methods (Habibi et al. [2017], Wang et al. [2018]) for error analysis. Thus, in the result section, we will show the error samples of each model and the sequence of labels together.

4.4 Baseline Methods

In the experiment, we will compare our Attention-CRF model with 4 previous neural network models:

1. Vanilla BiLSTM-CRF (VBC): A three layer BiLSTM-CRF architecture was employed by Lample et al. [2016] and Habibi et al. [2017]. They took the characters of each word in to the first BiLSTM layer to produce a character-level representation for this word. Then the character-level vector was concatenated with word embedding and fed into the second BiLSTM layer to produce label distribution. Finally, a CRF layer was added to maximize the log probability of label sequence.
2. Single Task Model (STM): Compared with vanilla BiLSTM-CRF, Wang et al. [2018] proposed STM to handle out-of-vocabulary (OOV) words. They still adopted three layer BiLSTM-CRF architecture but considered the character sequence of the input sentence rather than word as the input of the first BiLSTM layer.
3. LM-LSTM-CRF (LLC):
4. Tie or Break (ToB): Shang et al. [2018] regarded the NER task as entity span detection and entity type prediction problem. It had no CRF layer and was reported being more efficient than BiLSTM-CRF models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, March 1994. doi: 10.1109/72.279181.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4, July 2005. doi: 10.1109/IJCNN.2005.1556215.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14): i37–i48, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. *CoRR*, abs/1702.00887, 2017. URL http://arxiv.org/abs/1702.00887.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015. URL http://arxiv.org/abs/1508.04025.

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44, 2013.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. *arXiv preprint arXiv:1809.03599*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752, 2018.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.