CreateMoMo

2017-09-12

CRF Layer on the Top of BiLSTM - 1

Outline

The article series will include:

- **Introduction** - the general idea of the CRF layer on the top of BiLSTM for named entity recognition tasks
- **A Detailed Example** - a toy example to explain how CRF layer works step-by-step
- **Chainer Implementation** - a chainer implementation of the CRF Layer

Who could be the readers of this article series?

This article series is for students or someone else who is the beginner of natural language processing or any other AI related areas, I hope you can find what you do want to know from my articles. Moreover, please be free to provide any comments or suggestions to improve the series.

Prior Knowledge

The **only thing** you need to know is what is Named Entity Recognition. If you do not know neural networks, CRF or any other related knowledge, please **DO NOT** worry about that. I will explain everything as intuitive as possible.

1. Introduction

For a named entity recognition task, neural network based methods are very popular and common. For example, this paper[1] proposed a BiLSTM-CRF named entity recognition model which used word and character embeddings. I will take the model in this paper for an example to explain how CRF Layer works.

If you do not know the details of BiLSTM and CRF, just remember they are two different layers in a named entity recognition model.

1.1 Before we start

we assume that, we have a dataset in which we have two entity types, **Person** and **Organization**. Therefore, in fact, in our dataset, we have 5 entity labels:

- B-Person
- I- Person
- B-Organization
- I-Organization
- O

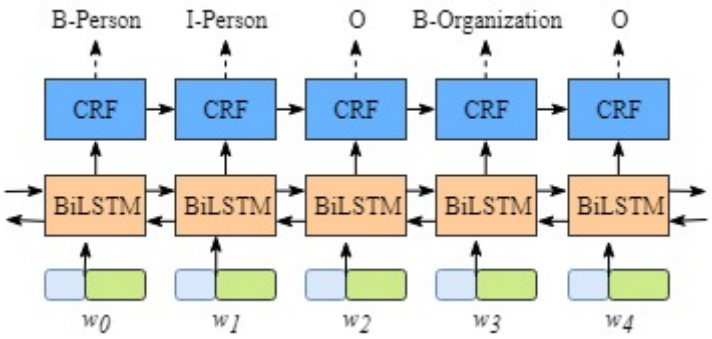
Furthermore, $ x $ is a sentence which includes 5 words, w_0,w_1,w_2,w_3,w_4. What is more, in sentence x, [w_0,w_1] is a Person entity, [w_3] is an Organization entity and others are “O”.

1.2 BiLSTM-CRF model

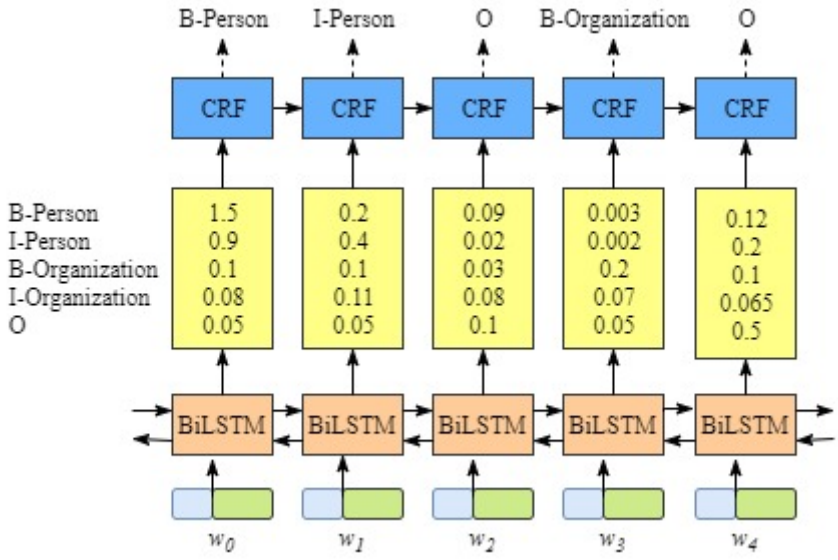
I will give a brief introduction of this model.

As shown in the picture below:

- **Firstly**, every word in sentence x is represented as a vector which includes the word’s character embedding and word embedding. The character embedding is initialized randomly. The word embedding usually is from a pre-trained word embedding file. All the embeddings will be fine-tuned during the training process.
- **Second**, the inputs of BiLSTM-CRF model are those embeddings and the outputs are predicted labels for words in sentence x.



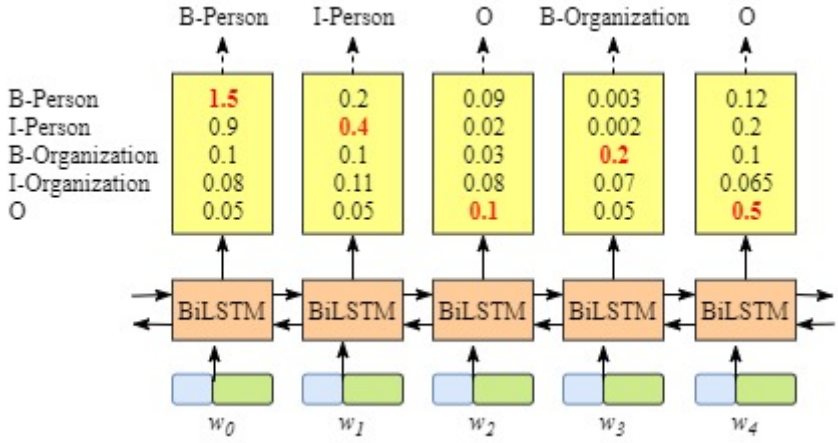
Although, it is not necessary to know the details of BiLSTM layer, in order to understand the CRF layer more easily, we have to know what is the meaning of the output of BiLSTM Layer.



The picture above illustrates that the outputs of BiLSTM layer are the scores of each label. For example, for w_0，the outputs of BiLSTM node are 1.5 (B-Person), 0.9 (I-Person), 0.1 (B-Organization), 0.08 (I-Organization) and 0.05 (O). These scores will be the inputs of the CRF layer.

1.3 What if we DO NOT have the CRF layer

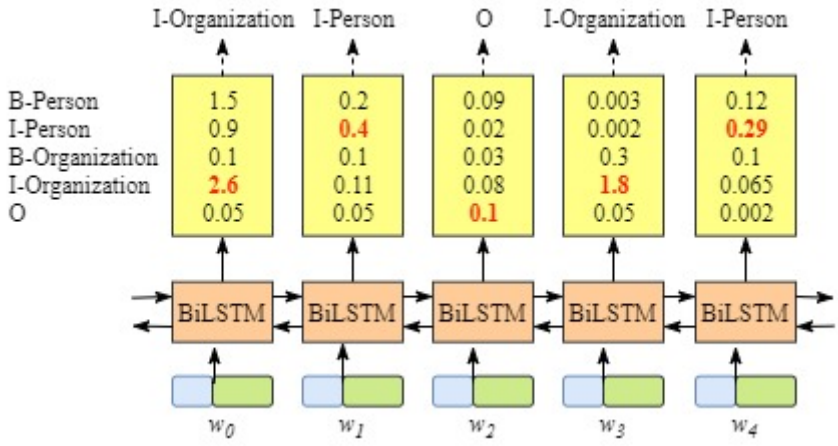
You may have found that, even without the CRF Layer, in other words, we can train a BiLSTM named entity recognition model as shown in the following picture.



Because the outputs of BiLSTM of each word are the label scores. We can select the label which has the highest score for each word.

For instance, for w_0, “B-Person”, has the highest score (1.5), therefore we can select “B-Person” as its best predicted label. In the same way, we can choose “I-Person” for w_1, “O” for w_2, “B-Organization” for w_3 and “O” for w_4.

Although we can get correct labels for sentence x in this example, but it is not always like that. Try again for the following example in the picture below.



Obviously, the outputs are invalid this time, “I-Organization I-Person” and “B-Organization I-Person”.

1.4 CRF layer can learn constrains from training data

The CRF layer could add some constrains to the final predicted labels to ensure they are valid. These constrains can be learned by the CRF layer automatically from the training dataset during the training process.

The constrains could be:

- The label of the first word in a sentence should start with “B-“ or “O”, not “I-“
- “B-label1 I-label2 I-label3 I-…”, in this pattern, label1, label2, label3 … should be the same named entity label. For example, “B-Person I-Person” is valid, but “B-Person I-Organization” is invalid.
- “O I-label” is invalid. The first label of one named entity should start with “B-“ not “I-“, in other words, the valid pattern should be “O B-label”
- …

With these useful constrains, the number of invalid predicted label sequences will decrease dramatically.

Next

In the next section, I will analyze the CRF loss function to explain how or why the CRF layer can learn those constrains mentioned above from training dataset.

Looking forward to seeing you soon!

References

[1] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C., 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360. https://arxiv.org/abs/1603.01360

When you reprint or distribute this article, please include the original link address.

Comments Share

NEWER
CRF Layer on the Top of BiLSTM - 2

ARCHIVES

RECENT POSTS