

MODEL:CRF+LSTM

CONTENTS

1. Introduction	2
2. Model Building	2
3. Embedding layer	2
4. Dropout layer	2
5. Bi LSTM layer	2
6. CRF layer	2
7. CRF	2
References	4
List of Todos	4

Date: 2021-01-16.

2020 Mathematics Subject Classification. NLP.

Key words and phrases. Machine Learning, Data Mining, ...

1. INTRODUCTION

2. MODEL BUILDING

Main ideas:

- 1.The token is mapped from a discrete one hot representation to a low dimensional space to become a dense embedding.
- 2.The embedding sequence of sentences is input into LSTM, and the feature is extracted automatically by neural network.
- 3.Using CRF to predict tag at sentence level.

3. EMBEDDING LAYER

Input: the ID list corresponding to the word and the characteristics of word segmentation information

Objective: to transform words into low dimensional dense vectors

Operation: train the n-dimensional word vector model in advance, and get the n-dimensional vector of each word through query. The n-dimensional vector is output to dropout layer together with the feature vector of segmentation information. The pre trained embedding can capture the global word similarity.

4. DROPOUT LAYER

Input: the n-dimensional vector corresponding to the word and the feature vector of segmentation information

Objective: to alleviate over fitting

5. BI LSTM LAYER

Input: the n-dimensional vector corresponding to the word

Objective: to extract sentence features automatically

Operation: the char embedding sequence of each word is taken as the input of each time step of Bi LSTM, and then the hidden state sequence of forward LSTM output and the hidden state sequence of reverse LSTM output in each position are spliced according to the position to get the complete hidden state sequence. Here you can capture some rules inside the word. Combining with pre training embedding, we can get better word embedding.

6. CRF LAYER

Input: word for the tag sequence rating

Objective: sequence annotation at sentence level

Operation: score again; the score is divided into two parts, LSTM output score plus CRF transfer matrix (i.e. transfer score from one label to another label). After scoring, softmax is used to get the normalized probability.

7. CRF

For example, You have many photos of Xiao Ming at different times of the day, from Xiao Ming getting up with his pants to taking off his pants and going to bed. Now the task is to sort these photos. For example, if some photos are taken during a meal, they should be labeled as eating; if some photos are taken during a run, they should be labeled as running; if some photos are taken during a meeting, they

should be labeled as meeting.

Simple method: The question is, what are you going to do? A simple and intuitive way is to train a multi classifier regardless of the time sequence of these photos. It is to use some labeled photos as training data to train a model and classify them directly according to the characteristics of the photos. For example, if the photo was taken at 6:00 a.m. and the picture is dark, label it sleeping; if there is a car in the photo, label it driving.

Is this feasible? Because we ignore the important information of the time sequence between these photos, our classifier will be defective. For example, if there is a picture of Xiao Ming with his mouth closed, how can it be classified? Obviously, it's difficult to judge directly. You need to refer to the picture before you shut up. If the previous picture shows Xiao Ming is eating, then the picture of shutting up is probably Xiao Ming chewing food to swallow, so you can label it as eating. If the previous picture shows Xiao Ming singing, then the picture of shutting up is probably a snapshot of Xiao Ming singing, so you can label it as singing. The label of the song.

CRF: Therefore, in order for our classifier to have better performance, we must take into account the label information of the adjacent photos when classifying a photo. This is where conditional random field (CRF) can play its role!

Very simple, is to give each word in a sentence marked part of speech. For example, the sentence "Bob drank coffee at Starbucks" after indicating the part of speech of each word is like this: "Bob drank (verb) coffee (noun) at (preposition) stars".

Next, we use conditional random fields to solve this problem.

Take the above sentence as an example, there are five words, we will: (noun, verb, noun, preposition, noun) as a tagging sequence, called L. there are many optional tagging sequences, for example, l can also be like this: (noun, verb, verb, preposition, noun). We want to select the most reliable one from so many optional tagging sequences as our interpretation of this sentence. The label of. How to judge whether a labeled sequence is reliable or not? If we give each marking sequence a score, the higher the score, the more reliable the marking sequence is. We can at least say that if there is a marking sequence after the verb in the marking, we should give it a negative score!! We can define a set of characteristic functions, use this set of characteristic functions to score a annotation sequence, and then select the most reliable annotation sequence. That is to say, each feature function can be used to score a annotation sequence. The final score of the annotation sequence is the combination of the scores of all feature functions in the set on the same annotation sequence.

REFERENCES

LIST OF TODOS