**CRF-based Clinical Entity Recognition with Medical Features**

Zelin Liu

11703910026

# **Abstract**

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories. NER is the basic of many NLP researches. A clinical document contains vital information about patient’s healthcare in unstructured free text format. Here we will introduce a CRF-based model to recognize clinical entity. Our model will provide

# 1. Introduction

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

There are many approaches to this problem. HMM(hidden markov model) and CRF(conditional random field) are two ways which are used most. And CRF combined with LSTM is becoming more and more popular since the revolution of the deep learning. However, in our model, we want to go deeper into the clinical futures instead of only using a deep learning framework and keep on tuning the parameters. And in fact, we did find some useful medical features and a better performance than using CRF.

Part 2 will make a brief introduction on CRF model. Part 3 will discuss the medical features we put in the CRF model. Part 4 will show the results and Part 5 will summarize this paper.

# 2. CRF model

Conditional Random Fields (CRF) are unidirectional graphical models, used to calculate the conditional probability of values on designated output nodes, given already assigned values to the input nodes.

The common used CRF is called as linear chain CRF.

The modelling equation is



The detailed part of CRF model can be found in many papers and also in our slices. And the training process of the CRF model could be finished using the python libray, sklearn-crfsuite. So we don’t have to spend too much efforts on this.

# 3. Clinical Features

We can add many features into CRF models as the templates. The templates can be used in unigram or bigram and so on. Here we mainly use unigram and bigram features.

# Medical Dictionary

This feature is more likely to the rule-based feature. We have create a medical dictionary which contains many medical words. Since clinical entity recognition would be more professional than the normal one, it makes sense to use medical dictionary. We take it as a binary unigram feature.(True or False)

# Prefix and Suffix

This feature might be used in many kinds of entity recognition. And in our clinical field, many diseases or treatments do share same prefix and suffix. So we use suffix as unigram feature.

# Stemming

It is a little bit same as the prefix feature we mentioned above. However, they are not the same. For example, hypertension and hypertensive, tachycardia and tachycardic. We take it as a unigram feature.

# Head of a noun phrase (Deprecated)

Let us give an example. For the same word ‘fever’, it falls into different categories of entity in the two sentences: ‘I got fever’ and ‘They are having a fever debate’. So will take advantage of nltk’s pos\_tag function and find whether a word is at the head of a noun phrase.

However, the experiment results show that this feature doesn’t work well and we finally eliminate this feature.

# PoS tags

This feature plays an important role in almost all kinds of entity recognition. We can also use nltk’s function to realize this. We use this as unigram feature and also left and right bigram feature.

# Others

There are some other features we can take advantage of.

If a word is all uppercased or is first-char-uppercased, it will more likely to be a clinical entity.

Also we will put emphasis on the word contains ‘-’ or ‘/’ or is numeric or date.

We can take these as many features.

# 4. Results

# Data Set

|  |  |  |
| --- | --- | --- |
|  | Corpus(7：3) | Test set |
| #sentences | 16315 | 27626 |
| #problem | 7073 | 12592 |
| #test | 4608 | 9225 |
| #treatment | 4844 | 9344 |

And actually each word has been put on a separate line and there is an empty line after each sentence. The first item on each line is a word, the second NER tag.

The NER tags have the format B-TYPE which means that the word is the first word of a phrase of type TYPE and I-TYPE which means the word is inside a phrase of type TYPE. A word with tag O is not part of a NER phrase.

# Training

We use the sklearn-crfsuite library to train the CRF models, we also use the features what we have mentioned above. We use the gridsearch and cross-validation to find a better parameters c1,c2.

# Results

Finally, we get c1 = 0.08 and c2 = 0.05 will be the best. And the result on dev.eval is

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | F1 |
| B-problem | 0.847 | 0.793 | 0.819 |
| I-problem | 0.791 | 0.775 | 0.783 |
| B-test | 0.870 | 0.786 | 0.826 |
| I-test | 0.819 | 0.754 | 0.785 |
| B-treatment | 0.857 | 0.771 | 0.812 |
| I-treatment | 0.743 | 0.703 | 0.722 |
| average | 0.820 | 0.770 | 0.794 |

And the result for test.eval is

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 |
| 91.68% | 71.98% | 80.22 | 75.88 |

# 5.Summary

To sum up , in this paper we have introduced a CRF-based clinical entity recognition algorithm. We get rid of deep learning or neural networks.

We have found many features from the medical or clinical aspects. Which will be more meaningful than only tuning parameters. And we do improve F1 score for about 4% on the F1-score.

However, the result is not as good as the one using neural networks. So we have to admitted that LSTM did perform well in NLP and what we will do next is to combine CRF+LSTM with these clinical features.

# 6.Reference

<http://f.dataguru.cn/thread-737880-1-1.html>

<https://www.jianshu.com/p/c36974c8aa7d>

<https://blog.csdn.net/qq_35014150/article/details/73826930>

<https://www.zhihu.com/question/35866596>

<https://www.ezdi.com/wp-content/uploads/2014/04/ClinicalNamedEntityRecognition.pdf>

<http://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html#features>

<https://link.springer.com/chapter/10.1007/978-3-319-73706-5_12>

https://en.wikipedia.org/wiki/Named-entity\_recognition