CRF-based Clinical Named Entity Recognition using clinical NLP Features

Parth Pathak
Easy Data Intelligence(ezDI)
parth.p@mediscribes.com

Raxit Goswami
Easy Data Intelligence(ezDI)
raxit.g@mediscribes.com

Gautam Joshi
Easy Data Intelligence(ezDI)
gautam.j@mediscribes.com

Pinal Patel
Easy Data Intelligence(ezDI)
pinal.patel@mediscribes.com

Amrish Patel
Easy Data Intelligence(ezDI)
amrish.p@mediscribes.com

Abstract

A clinical document contains vital information about patient’s healthcare in unstructured free text format, so Information Extraction and Named Entity Recognition are essential to extract meaningful information from this free clinical text. Here we propose a CRF-based supervised learning approach using customized clinical features set to recognize named Entity. The experiment was carried out on i2b2 shared task 2010 data, to recognize three types of named entity (Problem, Treatment and Test). For inexact match, we achieved 0.966 precision, 0.883 recall and **0.923 F-Score**, while for exact match, 0.889 precision, 0.813 recall and **0.849 F-score**. Our approach worked better than all the supervised and hybrid models, while it gave almost similar result to the semi-supervised models used in the shared task. This showed that supervised learning with better feature selection can give as accurate result as semi-supervised learning.

1 Introduction

Electronic Medical Records (EMR) contains only 20 to 25% of patient information in the structure format, while the rest of the patient information resides as a free text inside a clinical document. This clinical free text has both formal and informal linguistic style, so a state of the art Natural Language Processing Engine is required to extract this information. Our aim is to create a linguistically reach NLP engine, which can handle the peculiarity of clinical text. In this, paper we will try to explain our approach towards named entity recognition task and also evaluate NER module our NLP engine.

Unified Medical Natural Language System (UMLS) is the largest medical knowledge resource available. In the past, many traditional NLP engines have used rule based dictionary lookup methods, with UMLS as a base dictionary, to detect NERs. However these approaches fetch very low recall, due to the fact that dictionary lookup can never capture all the lexical and linguistic variants of a medical term, and also due to the fact that clinical documents contains a lot of abbreviation which may vary depending upon a physician’s writing style. So approaches involving machine learning algorithms like Conditional Random Fields (CRF), Support Vector Machine (SVM), and Maximum Entropy Markov Model (MEMM) have been used which can not only utilize textual as well as contextual information to detect NER but also lower the dependency on dictionary lookup. But these approaches also failed to improve accuracy after a certain point, because most of them used traditional NER feature, and refrained from taking the advantage of peculiarity of clinical Data. So here the approach presented by us, uses the unique feature set specifically customized for clinical NLP.

As Conditional Random Fields (CRF) are widely used across multiple domains to detect Named Entities, it was best available choice for our experiment. CRFs are undirected discriminative probabilistic graph model, which have efficient procedure for complete, non-greedy finite state inferences and training. We have used i2b2 shared task data to find out 3 different type of Named Entity (see table 1).

Named Entity	Example
Problem	hypertension, cancer
Treatment	CABG,Endoscopy,Aspirin
Test	Echocardiogram, Blood pressure

Table-1: Named Entity Type and its example

2 Related Work

Over the past many years several NLP tools like cTAKES, MedLEE, and metaMap have used a rule based dictionary lookup method using UMLS meta-thesaurus as a base dictionary. But result by Karin Schuler (2008) showed that, these methods can only fetch a very low F-score of 0.56 for exact matches. Several other approaches based on machine learning algorithms have been tried out. Yefang Wang deployed a voting strategy on the top of three cascading classifier (SVM, CRF and MEMM) and got F-Score of 0.832 for exact matches. But it is very difficult to improve results of cascading classifiers. Xu Y got F-score of 0.848 by combining the rule based method with machine learning. Roberts et al broke NER task into two parts, in the first part they trained SVM to detect NER boundary and in the second part they trained CRF to identify concept and got F-score of 0.796. deBruijin B et al used a semi-supervised approach to detect Named Entity and got F-score of 0.852. But in semi-supervised methods it is very difficult to predict the number of clusters required.

3 Conditional Random Fields

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

BIO (begin-in-out) annotation method was used to annotate different categories, where B_Category_Type represents starting of Entity and I_category_Type represents continuity of an Entity and O is used for all other words. CRF++ a simple and customizable implementation of CRF for segmenting and sequencing the data, was used to train as well as tag the data. In the next portion we will try to summarize the theory behind CRF.

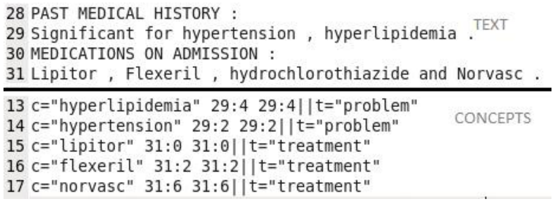
Let O= {o_1, o_2, …, o_T} be a observed input sequence, i.e. sequence of words of a sentence in clinical document. Let S be a set of FSM states each associated with some label *l*, where *l ε* {classification categories like problem, procedure, Medicine}. Let s= {s_1, s_2, …, s_T} sequence of state for given sentence. By Hammersley-Clifford theorem, the conditional probability of a state sequence given an input sequence will be:

$$P_\Lambda(\mathbf{s}|\mathbf{o}) = \frac{1}{Z_\mathbf{o}} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right)$$

where Z_0 is a normalization factor over the all state sequence, which ensures that all the probability distribution sums up to 1.Generally computing Z is intractable but there are a few methods available which shows how to approximate it. $f_k(S_{t-1},S_t,O,t)$ is a feature function over its argument. A feature function can be explained by following example in clinical context: suppose binary feature *stop words* always has value 0, but it changes to 1 if only if S_{t-1} has any one of the six NE categories and S_t has the category “Other” and observation O at position *t*, has a word, which appeared in stop word dictionary. Higher the value of λ makes their corresponding more likely, so in the above example weight of the λ_k should be positive. In general view, feature function f_x can ask powerful arbitrary questions about previous or next sequence of input words and value of k can range from -∞ to +∞.

4 Data

For i2b2 shared task, Partner’s Healthcare, Beth Israel Deaconess Medical Center and University of Pittsburg Medical Center contributed the data. There were 426 manually annotated files, out of which 170 files were used for training and 256 files were used for the testing. Annotation was done for three basic categories Problem, Treatment and Test. Breakdown for different concept is shown in Table 2. For each clinical text file, its respective annotation was done in a concept file as shown figure 1.1, where each line in concept file represents a single concept, which can be traced back to text file using begin-end token number.



```
28 PAST MEDICAL HISTORY :
29 Significant for hypertension , hyperlipidemia .
30 MEDICATIONS ON ADMISSION :
31 Lipitor , Flexeril , hydrochlorothiazide and Norvasc .
```

```
13 c="hyperlipidemia" 29:4 29:4||t="problem"
14 c="hypertension" 29:2 29:2||t="problem"
15 c="lipitor" 31:0 31:0||t="treatment"
16 c="flexeril" 31:2 31:2||t="treatment"
17 c="norvasc" 31:6 31:6||t="treatment"
```

Figure 1.1 : Annotation technique for different categories

	Problem	Treatment	Test	Total
Training	7073	4844	4608	16525
Testing	12592	9344	9225	31161

Table 2: Training and Testing data breakdown

5 Feature sets

Stemming: There can be many variant of the same medical entity in the clinical text, like hypertension and hypertensive, tachycardia and tachycardic, so a basic stemmer, as a unigram feature, was used to generalize different variants.

Part of Speech Tags & Chunks (PoS tags): PoS tags with chunks play an important role in deciding the boundary of a named Entity. Unigram as well as left and right bigrams were used as features.

Head of the Noun phrase: Consider following examples: i) the patient has *diabetes*.
ii) The patient was given *diabetes education*.

In the first example *diabetes* should be annotated as a disease, while in the second example the whole phrase *diabetes education* should be annotated as a Finding. In many examples the head of the noun phrase becomes a deciding factor for classifying a named Entity. A binary (true/false) unigram was used as a feature.

Prefix and Suffix: Many diseases and treatments share same prefix or suffix, like Adrenalectomy, Sclerotomy, and Osteotomy all shares a common suffix "-tomy". Unigram suffix and prefix were used as features.

Section Headers: A clinical note is often divided into relevant segments called Section Headers, like History of Present Illness, Current Medicines, and Lab Data. These section headers provide very useful information at the discourse level. After analyzing more than 10,000 clinical documents, we have classified section headers into more than 40 hierarchical categories. In the clinical NER task there are quite a few named entities like vitamin B_{12}, glucose, insulin which can fall under multiple categories depending upon context where knowledge about section header can be very helpful. Unigram section header id was used as a feature for all the tokens.

Orthographic Features: General orthographical binary (true/false) unigram features like whole word capital, First char capital, Numeric values, Dates, words contains hyphen or slash, medical units (mg/gram/ltr etc) were used as features.

Stop words: From the initial result we found that sometimes Part of Speech tags or Chunks are not always enough for detecting Entity Boundaries, so some prepositions and conjunctions were added in the stop word list. A binary (true/false) unigram was used as a feature.

Dictionary Search: A binary (true/false) unigram feature was used to check whether the word is present in the medical dictionary or not.

Abbreviation and Acronym: Abbreviations in clinical text varies from domain to domain, from clinic to clinic and from physician to physician. It is very difficult to find list of all the valid abbreviation from a medical dictionary, so a binary classifier was trained on SVM to detect whether given entity is abbreviation or not and was used as a unigram feature in this task.

6 Results

The evaluation task was done using two different measures:

Exact micro-averaged precision, recall, and F-Measure: where phrase boundaries and concept type matches exactly and i) correct boundary with incorrect type get no credit. ii) Incorrect boundary with correct type gets no credit iii) incorrect boundary with incorrect type gets no credit. For exact matches we got 0.889 precision, 0.813 recall and 0.849 F-score. Contribution of feature by adding different feature progressively is as shown in Table 3.

Inexact micro-averaged precision, recall and F-score: Concept tagged overlaps with the ground truth concepts at at-least one part. For inexact match, we achieved 0.966 precision, 0.883 Recall and 0.923 F-Score. Table 4 shows comparison of our output with rest of the participants of i2b2 shared task.

System By	Method	Exact F-Meas	In-exact F-Meas
deBrujin et al	Semi-supervised	0.852	0.924
Parth et al	**Supervised**	**0.849**	**0.923**
Jinag et al	Hybrid	0.839	0.913
Kang et al	Hybrid	0.821	0.904
Gurulingappa et al	Supervised	0.818	0.905
Patrick et al	Supervised	0.818	0.898
Tori& Lue	Supervised	0.813	0.898
Jonnalagadda &Gonzalez	Semi-supervised	0.809	0.901
Sassaki et al	Supervised	0.802	0.887
Roberts et al	Supervised	0.788	0.884

7 Conclusion and Future Work

Results of our experiments showed that, Feature selection is very important in improving the accuracy clinical NERs.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J.Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114-133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK

http://jamia.bmj.com/content/19/5/824 Xu Y

http://www.cs.brandeis.edu/~marc/misc/proceedings/lrec-2008/pdf/764_paper.pdf Karin Schuler

http://www.aclweb.org/anthology-new/W/W09/W09-4507.pdf Yefeng Wang