CRF-based Clinical Named Entity Recognition using clinical NLP Features

Parth Pathak
Easy Data Intelligence
parth.p@ezdi.us

Raxit Goswami
Easy Data Intelligence
raxit.goswami@ezdi.us

Gautam Joshi
Easy Data Intelligence
gautam.joshi@ezdi.us

Pinal Patel
Easy Data Intelligence
pinal.patel@ezdi.us

Amrish Patel
Easy Data Intelligence
amrish.patel@ezdi.us

Abstract

A clinical document contains vital information about patient's health such as diseases, symptoms, drugs and treatments in unstructured free text format. Hence, the Named Entity Recognition (NER) is essential to extract meaningful information from this free clinical text. Traditional NLP methods for NER use either a dictionary lookup approach or a Machine learning (ML) approach with textual features. However, they fail to get a decent accuracy due to lack of use of domain specific features. Here we propose a CRF-based supervised learning approach which uses the domain knowledge in form of clinical features. The experiment was carried out on i2b2 shared task 2010 data, to recognize three types of named entity. For inexact match, we achieved F-Score of 0.923, and for exact match F-score of 0.849. Our approach gave better accuracy than all the state-of-the-art approaches which use supervised and hybrid models. 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. A clinical document is divided into numbers segments called section header that contains information about patient's chief complaints, past history, lab data, medicines, and current diagnosis. These section headers have both, very short sentence fragments as well as very long descriptive sentences. These sentences can have both formal and informal linguistic style.

Unified Medical Natural Language System (UMLS) is the largest medical knowledge resource available. For NER, many traditional NLP engines have used rule based dictionary lookup methods, with UMLS as a base dictionary. However these approaches have shown very low recall, mainly because dictionary lookup fails to capture all the lexical and linguistic variants of a medical term e.g. hypertension can also be written as hypertensive disease or high blood pressure. Also one abbreviation can expand to multiple meanings e.g. MR can be expanded as Mister or Mental Retardation or Mitral Valve Insufficiency. To overcome this limitations, various ML approaches have been proposed such as Conditional Random Fields - CRFs (J. Lafferty et al, 2001), Support Vector Machine (T. Joachims et al, 1998), and Maximum Entropy Model (A. Berger et al, 1996), which utilize textual and contextual information, while lowering the dependency on dictionary lookup. But these approaches failed to improve accuracy after a certain point, because most of them did not utilize the peculiarity of clinical text. Previous approaches have used linguistic features like Stemming, Prefix, Suffix, PoS tags and Chunks. However, they fail to take

advantage of the domain specific features like Section Headers, customized stop words, dictionary search, Abbreviation and acronym.

In this paper, we have proposed a novel approach, which uses domain specific knowledge in the form of clinical features along with linguistic features. We have used CRF as supervised learning models which are widely used across multiple domains to detect Named Entities. i2b2 shared task data was used to find out the different types of Named Entities (see table 1).

Named Entity Type	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(Guergana et al, 2010), MedLEE(C. Friedman et al, 2004), and metaMap(A. Aronson et al, 2001) have used a rule based dictionary lookup method using UMLS meta-thesaurus as a base dictionary. But result by Karin et al (2008) showed that, these methods can only fetch a very low F-score of 0.56 for exact matches. Several other ML approaches have also been tried out. Yefang Wang et al (2009) deployed a voting strategy on the top of three cascading classifier (SVM, CRF and MEMM) and obtained F-Score of 0.832 for exact matches. But it is very difficult to improve results of cascading classifiers. Xu Y et al(2012) got the F-score of 0.848 by combining the rule based method with machine learning. Roberts et al(2010) 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(2010) used a semi-supervised approach to detect Named Entity and got the F-score of 0.852. However in semi-supervised methods it is very difficult to predict the number of clusters required.

3 Using CRF for Clinical Text

CRFs are unidirectional graphical models, used to calculate the conditional probability of values on designated output nodes, using already assigned values to the input nodes.

BIO (begin-in-out) annotation method was used to annotate different categories, where B_Category_Type represents beginning of an Entity and I_category_Type represents continuity of an Entity and O is used for all other words. In the next portion we will try to summarize the theory behind CRF.

Let $O = \{o_1, o_2, \ldots, o_T\}$ 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 belongs to* {classification categories like problem, test, treatment}. Let $s = \{s_1, s_2, \ldots, s_T\}$ sequence of state for given sentence. By Hammersley-Clifford theorem, the conditional probability of a state sequence given an input sequence will be:

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

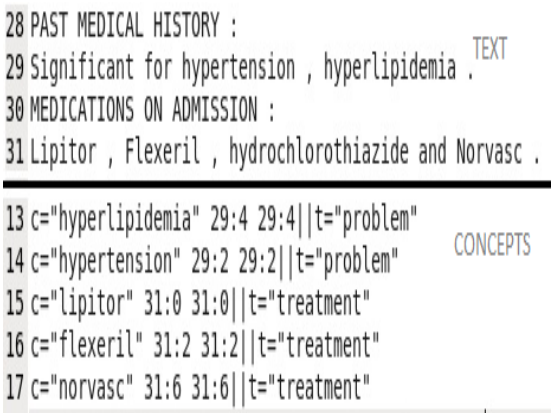
where Z_0 is a normalization factor over the all state sequence, which ensures that all the probability distribution sums up to 1. Generally computing Z_0 is intractable, but there are several methods available to approximate it. $f_k(S_{t-1},S_t,O,t)$ is a feature function over its argument. A feature 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 and only if S_{t-1} has any one of the three 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 value of λ makes their corresponding f_k more likely, so in the above example weight of the λ_k should be positive. In general view, feature function f_k can ask powerful arbitrary questions about previous or next sequence of input words and value of k can range from $-\infty$ to $+\infty$.

CRF++, a simple and customizable implementation of CRF for segmenting and sequencing the data, was used to train as well as tag the data.

4 Experimental Setup

For i2b2 shared task (Özlem Uzuner et al[3]), Partner's Healthcare, and Beth Israel Deaconess 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. Also there were 267 unannotated files. Annotation was done for three basic categories: Problem, Treatment and Test. Breakdown of different concepts property, is shown in Table 2. For each clinical text file, its respective annotation is given in a concept file as shown figure 1.1. Each line in concept file represents a single concept, which has four attributes: 1) concept name (i.e. hypertension, hyperlipidemia) 2) concept type (i.e. problem, treatment, test) 3) Begin line number and token number (29:4 means 29th line 4th token) 4) End line number and token number. Each line in text file represents a single sentence. All the section headers are written in capital letters having a colon at the end. (i.e. PAST MEDICAL HISTORY, MEDICATION ON ADMISSION)



"Figure1.1. Annotation technique"

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

"Table2. training and testing data breakdown"

5 Feature sets

The feature sets were divided into three basic categories. 1) Textual Feature (Stemming, Prefix, Suffix, Orthographical Features) 2) Linguistics Features (PoS & Chunks, NP Head) 3) Clinical Features (Section Headers, customized stop words, dictionary search, Abbreviation and acronym). The next part summarizes all the 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. But section header classification in itself is a big task. Same section header can have multiple variant as shown in Table3. Sections like Review of System or Past Medical History can have many sub section. After analyzing more than 10,000 clinical documents, we have developed a database of 5500 different section, which are classified into more than 40 hierarchical categories. But using only section header dictionary for classification, fetch many false positives. So a simple Hidden Markov Model was used to take advantage of section header sequences and remove false positives. 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.

HISTORY OF PRESENT PROBLEM
HPI
HOPI
Brief HPI
HISTORY OF PRESENT ILLNESS
CURRENT HISTORY OF PRESENT ILLNESS
UPDATED HISTORY OF PRESENT ILLNESS
PRESENTING CONCERNS AND HISTORY OF PRESENT ILLNESS
UPDATED HOPI

"Table3. Variants of HPI section"

Dictionary Search: Unified Medical Natural Language System (UMLS) along with Lexicon Variant Generator (LVG) was used as dictionary to detect different concepts. However traditional UMLS dictionary search do provide many false positive, so to limit number of false positives we used rule based filters.

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. Abbreviation Disambiguation is a very big challenge in clinical NLP. We used LRABR as our base dictionary to detect abbreviation. On top

of it, a simple binary classifier trained on SVM was used to detect whether given entity is valid abbreviation or not. This is very helpful in removing false positives.

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.

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

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

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

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 correct type gets one credit ii) Correct boundary with incorrect type gets no credit. iii) Incorrect boundary with correct type gets no credit. iv) Incorrect boundary with incorrect type gets no credit. For exact matches we got 0.889 precision, 0.813 recall and 0.849 F-score. As shown in Table 3, our result is better than all the other supervised and hybrid approaches used in the shared task.

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, which is also similar to the best result (deBrujin et al[10]) of i2b2 shared task.

System By	Method	Exact F-Meas	In-exact F-Meas
deBrujin et al	Semi-supervised	0.852	0.924
Our approach	**Supervised**	**0.849**	**0.923**
Jinag et al	Hybrid	0.839	0.913
Kang et al	Hybrid	0.821	0.904
Gurulingappa et al[4]	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

" Table 4: shared task results"

7 Conclusion and Future Work

We proposed a novel approach for Clinical named entity recognition, which uses domain knowledge in form of clinical features along with existing linguistic features. Our approach outperformed all the other approaches, which were using only linguistic features. Hence, it can be concluded that, effective use of domain knowledge in the form of feature creation is very important for solving any problems with machine learning algorithm. In future, we plan to extend this approach for extracting the drug and its attributes along with their relationship with diseases.

References

Karin Kipper-Schuler, Vinod Kaggal, James Masanz, Philip Ogren and Guergana Savova. "System evaluation on a named entity corpus from clinical

notes", *Language Resources and Evaluation Conference, LREC. 2008.APA.*

Wang, Yefeng, and Jon Patrick. "Cascading classifiers for named entity recognition in clinical notes." *Proceedings of the workshop on biomedical information extraction. Association for Computational Linguistics, 2009.*

Özlem Uzuner, Brett R South, Shuying Shen, Scott L DuVall. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text 2011".*Journal of the American Medical Informatics Association 18.5 (2011), 18(5), 552-556.*

Gurulingappa H, Hofmann-Apitius M, Fluck J. "Concept identification and assertion classification in patient health records". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2,2010.*

Patrick JD, Nguyen DHM, Wang Y, et al. "I2b2 challenges in clinical natural language processing 2010". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.*

Torii M, Liu H. "BioTagger-GM for detecting clinical concepts in electronic medical reports". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.*

Jonnalagadda S, Gonzalez G. "Can distributional statistics aid clinical concept extraction". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.*

Sasaki Y, Ishihara K, Yamamoto Y, et al. "TTI's systems for 2010 i2b2/VA challenge". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data.* Boston, MA, USA: i2b2, 2010.

Roberts K, Rink B, Harabagiu S. "Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data.* Boston, MA, USA: i2b2, 2010.

deBruijn B, Cherry C, Kiritchenko S, et al. NRC at i2b2: "one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.

T. Joachims, C. Nedellec, and C. Rouveirol. "Text categorization with support vector machines: learning with many relevant". *In Machine Learning: ECML-98 10th European Conference on Machine Learning, Chemnitz, Germany. Springer,1998.*

A. Berger, V. Della Pietra, and S. Della Pietra. "A maximum entropy approach to natural language processing". *Computational linguistics, 22(1):39–71, 1996.*

J. Lafferty, A. McCallum, and F. Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In *machine learning-international workshop then conference*, *pages 282–289, 2001.*

A. Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program". *In Proceedings of the AMIA Symposium, page 17. American Medical Informatics Association,* 2001

C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. "Automated encoding of clinical documents based on natural language processing". *Journal of the American Medical Informatics Association, 11(5):392–402, 2004.*

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications". *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

Jiang M, Chen Y, Liu M, et al. "Hybrid approaches to concept extraction and assertion classification - vanderbilt's systems for 2010 I2B2 NLP Challenge". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.*

Kang N, Barendse RJ, Afzal Z, et al. "Erasmus MC approaches to the i2b2 Challenge". *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.*

Yan Xu,Kai Hong,Junichi Tsujii, Eric I-Chao Chang2 "Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative" *clinical discharge summaries ,2012.*