Understanding Medical Named Entity Extraction in Clinical Notes

Aman Kumar[1], Hassan Alam[1], Rahul Kumar[1], Shweta Sheel[1]
[1]BCL Technologies, San Jose, CA

Abstract - *Clinical notes contain extensive knowledge about patient medical procedures, medications, symptoms etc. In this paper we present an integrated approach to processing textual information contained in the clinical notes. We extract three major medical entities namely symptoms, medication and generic medical entities from patient discharge summaries and doctors notes from the I2B2 dataset. Quick access to structured information of these entities may help medical professionals in providing better and cost-effective care.*

Keywords: *Medical Named Entity Extraction, Machine Learning, Natural Language Processing*

1 Introduction

In recent years, various medical facilities and individual care providers have embraced electronic medical record (EMR) or some other version of electronic data management system. Understanding various facets of patient history is critical to speedy and economical treatment.

Clinical notes have been analyzed in greater detail to harness important information for clinical research and other healthcare operations, as they depict rich, detailed medical information. In this paper we describe a machine learning system which extracts medical named entities of three categories namely symptom, medication and generic medical condition. In this study we do not analyze text in the bio-medical journals or research papers. We analyze clinical text from doctor's notes and records that are generated during patient interviews. Clinical notes pose challenge for natural language processing in that they contain short phrases, abbreviations, acronyms etc. Sometimes there are instances of ungrammatical constructions in these notes.

This paper is organized as follows. First we provide a brief overview of various medical named entity extraction techniques and tools. We then describe the machine learning method that we use to build the medical named entity extraction system. Then we go over the experiment design followed by a section on results and discussion.

1.1 Related Studies

Named entity extraction is a type of information retrieval which focuses on identifying instances i.e., names of various types of entities. For example, cancer would be an instance of disease; swelling would be an instance of symptoms and so on. One of the earliest NER models was based on decision tree [1]. In this paper [1], Sekine used features such as part-of-speech tags extracted by a morphological analyzer, character based information and specialized dictionary. This system was developed for Japanese. Another early work was done by Bikel, Schwartz and Weischedel [2]. Authors used Hidden Markov Model (HMM) to identify named entity. Primary features like bi-gram and orthographic features like word case, word shape etc. were used. Borthwick [3] in his PhD thesis used maximum entropy (MaxEnt) algorithm.

McCallum and Li [4] developed Conditional Random Fields based algorithm to extract NER in coNLL-2003 shared task competition. Sarawagi and Cohen [5] propose a semi Markov CRF (Conditional Random Field) algorithm for named entity extraction. Cohen and Sarawagi [5] further extended the semi Markov model with use of dictionary and notion of similarity function. Naidu and Sekine [6] provide wide overall survey of NER research.

Aranson [7] developed MetaMap to map bio-medical concepts from Unified Medical Language System (UMLS). FriedMan et al. [8] developed a NLP system based on MetaMap which extracts various entities from clinical notes such as temporal information, corresponding codes by matching with UMLS etc. They extract concepts in semantic form based on various predefined frames. Minard et al. [9] presented and compared multiple approaches based on domain-knowledge and machine-learning techniques to Medical Entity Recognition. They show that the hybrid approach based on both machine learning and domain knowledge obtains the best performance. Li, Schuler and Savova [10] have used both CRF and SVM based for model extraction of *disorder* in clinical text. In this paper Authors have extracted a dictionary from SNOMED-CT [11]. SNOMED-CT is a map of concepts and relationships. In total it contains almost 360,000 and 1 million relationships. It is classified into categories such as procedure, body part etc. Several such models has been developed which uses variation of statistical models aided with dictionary based system. Wang & Patrick [12] present a cascaded system to do NER on clinical notes. Their system consists of two of CRF and SVM respectively. Patrick & Li [13] developed a CRF based methods to extract medication from clinical text. Meystree et al. [14] provide an overview of recent developments in clinical information retrieval field.

2 Methodology

Our medical named entity extraction system is modeled with a CRF model. In the literature it has been consistently demonstrated that CRF models are best performing models for sequential labeling. For the present study we extended and modified the Stanford NER package [15]. We prefer Stanford NER package because it can be seamlessly integrated with existing NLP tool suite such as parser and morphological analyzer.

CRFs are undirected graphical models which can be interpreted as conditionally trained finite state machines. While based on the same exponential form as maximum entropy models, they have efficient procedures for complete, non-greedy finite-state inference and training.

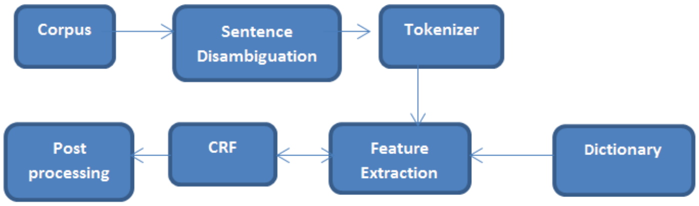
Here are the definitions:

Let *G=(V,E)* be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then **(X,Y)** is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that *w* and *v* are neighbors in *G*.
If the graph *G* = (*V*, *E*) of Y is a tree, the conditional distribution over the label sequence Y = y, given X = x, by fundamental theorem of random fields is:

$$p_\theta(y|x) \propto \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x)\right)$$

Where
$\theta = \lambda_1, \lambda_2, \ldots, \lambda_n, \mu_1, \mu_2, \ldots, \mu_n$ are weights to be estimated by the model.

The architecture of our system is given below.



The I2B2 corpus was annotated with the following tags.

1. MNE-S: refers to medical named entity symptom.
2. MNE-M: refers to medical named entity medication.
3. MNE: This tag refers to generic medical named entity which may not qualify under previously defined categories but are useful to physicians in understanding patient clinical record.

The I2B2 corpus was annotated in Begin-Inside-Outside (BIO) format. We noticed that several clinical notes had a lot of discrepancies with respect to sentence boundary. So we used Stanford CRF model to first break the corpus into sentences and then these sentence were broken into tokens. We curated an extensive dictionary of medical conditions, symptoms and medications using SNOMED-CT.

2.1 Feature Extraction

A brief overview of features extracted is described below.

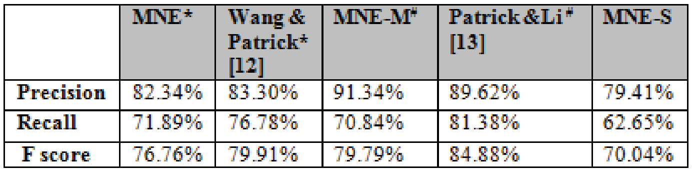
1. Word based features: We extracted N-gram features to understand word context. We used a window of N=1, 2, 3, 4. We also used word shape as a feature. This was appropriate to understand some tokens like temperature and concepts like dosage, frequency etc. We also used morphological features such as prefix and suffix of words.
2. Semantic Knowledge: Known acronyms and synonyms were extracted.
3. Orthographic features: Spelling variations and spelling corrections were extracted.
4. Parse tree features: We used Stanford parser to extract semantic features. We tagged tokens with part-of-speech tags. We extended these features in N-gram fashion (N=1,2..). We also introduced distance of tokens from numeric quantifiers, if present in a sentence.
5. Dictionary based features: We used Boolean features - if present token is a medication or a symptom or a specific medical condition etc.
6. To further reduce noise related to abbreviations such as q.i.d, Dr. etc., we developed a lexicon. Such tokens were represented with specific features during model development.
7. Extensive regex based features to identify time entities were used. Presence of such entities indicates context for specific medical events (such as pain in the night etc.) in the clinical notes.
8. Character level features: Presence of special characters like /, @, -, :, mixed cases, alphanumeric characters in a word etc were harnessed. Each such occurrence was represented with a unique Boolean flag.
9. Aspell dictionary check: We introduced a special Boolean flag if the word was present in Aspell English dictionary.

3 Experiment and Evaluation

We extracted corpus from I2B2 dataset (https://www.i2b2.org/NLP/DataSets/). In total we sampled 2100 sentences from the corpus. Our motive for sampling was that we wanted a system that was not overly representative of a specific entity such as symptom. We wanted a balanced system that had equal representation of all entities under investigation. These sentences were then annotated by a physician in BIO format for each entity. We split the corpus in 70/30 ratio for training and testing respectively. Training of the model was done with 10-fold cross validation.

Table 1 shows the evaluation scores of our system. We use standard Recall, Precision, and F-score to measure the performance of the system.

Table 1. Experiment Results



	MNE*	Wang & Patrick* [12]	MNE-M#	Patrick &Li# [13]	MNE-S
Precision	82.34%	83.30%	91.34%	89.62%	79.41%
Recall	71.89%	76.78%	70.84%	81.38%	62.65%
F score	76.76%	79.91%	79.79%	84.88%	70.04%

*, $^{\#}$ represent same category.

In the results presented above, we show that our system has done extremely well in simultaneous extraction of multiple entities from clinical notes. The comparable systems although perform marginally better, they only extract a single type of entity unlike ours where we extract three entities – MNE-M, MNE-S, and base MNE. Also, existing systems do not extract MNE-S (Medical Named Entity – Symptom).

4 Conclusions

In this paper we presented an integrated machine learning system to extract three major types of medical named entities. We believe such an integrated system is more practical than systems that extract entities in isolation.

Our future efforts will be toward building a larger granular set of medical entity extraction, which will further improve multi-class medical entity extraction system. Such a system can be used for preparing medical reports based on patient's records such as patient discharge summary and doctor's notes.

5 References

[1] Sekine, S. 1998. Nyu: Description of the Japanese NE System Used For Met-2. In Proc. Message Understanding Conference

[2] Bikel, D. M., Schwartz, R., & Weischedel, R. M. An algorithm that learns what's in a name. Machine learning 34, no. 1-3 (1999): 211-231.

[3] Borthwick, A. A maximum entropy approach to named entity recognition. PhD diss., New York University, 1999.

[4] McCallum, A & Wei L. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pp. 188-191.

[5] Sarawagi, S. & Cohen, W. W. Semi-markov conditional random fields for information extraction. In Advances in Neural Information Processing Systems, pp. 1185-1192. 2004.

[6] Cohen, W. W., & Sarawagi, S. Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 89-98. ACM, 2004.

[6] Nadeau, D. & Sekine, S. A survey of named entity recognition and classification. Lingvisticae Investigationes 30, no. 1 (2007): 3-26.

[7] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program

[8] Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated Encoding of Clinical Documents Based on Natural Language Processing. Journal of the American Medical Informatics Association : JAMIA, 11(5), 392–402.

[9] Minard AL, Ligozat AL, Ben Abacha A, et al. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. J Am Med Inform Assoc. 2011; 18(5):588–93

[10] Li, D., Kipper-Schuler, K., & Savova, G. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In Proceedings of the workshop on current trends in biomedical natural language processing, pp. 94-95. Association for Computational Linguistics, 2008.

[11] Bos, L., and K. Donnelly. "SNOMED-CT: The advanced terminology and coding system for eHealth." Stud Health Technol Inform 121 (2006): 279-290.

[12] Wang, Y. & Patrick, J. Cascading classifiers for named entity recognition in clinical notes. In Proceedings of the workshop on biomedical information extraction, pp. 42-49. Association for Computational Linguistics, 2009.

[13] Patrick, J., & Li, M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. Journal of the American Medical Informatics Association, 17(5), (2010):524-527.

[14] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. Extracting information from textual documents

in the electronic health record: a review of recent research. Yearb Med Inform 35 (2008): 128-44.

[15] Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363-370. Association for Computational Linguistics, 2005.