Recognising Named Entity of Medical Imaging Procedures in Clinical Notes

Wei Ruan, Won-sook Lee
School of Electrical Engineering and Computer Science (EECS)
University of Ottawa
Ottawa, Canada
{wruan013, wslee}@uottawa. ca

***Abstract*—Information on medical imaging procedures in free-text clinical notes plays a significant role in help diagnosis process. In this paper, we present a named entity of the medical imaging procedure recognition system based on conditional random fields (CRF) model with word-based, part-of-speech, Metamap semantic and et.al features. The system is trained and tested on a manually labelled dataset from I2B2 with the F_1 score up to 0.923 for recognizing medical imaging procedure entities. Our system can be customized by simply defining new medical named entities, which has been proved on medication recognition with the F_1 score from 0.870 to 0.937.**

***Index Terms*—Medical Named Entity Recognition, Information Extraction**

I. Introduction

This paper presents an extending work for improving pictorial visualization of EMR (Electronic Medical Records) summary system developed by Ruan.W and et.al [1]. Since current EMR contain large amounts of free-text contents and various tables to show numerous health data, it limits users (no matter doctors or patients) from promptly determining medical conditions or quickly finding desired information to a certain extent. They aim to tackle this as information visualization and extraction problems by the creation of easy and intuitive user interfaces for visualizing medical information. The pictorial visualization of EMR summary system (PVEMRSS) they developed has an ability to extract medical information, such as medications, physical exams and human body location, from clinical notes and visualize them on an interface spatially and temporally.

In this study, we have expanded the type of medical information extracted for improving PVEMRSS. We use the technique of named entity recognition based on conditional random fields (CRF) model with word-based, Part-of-Speech and Metamap semantic features to extract entities of medical imaging procedures. Medical imaging procedures are one of the non-invasive diagnostic tests (but some of them involve exposure ionizing radiation), which allow doctors to see inside the body in order to determine the best treatment options for patients. According to the study of Fazel.R et.al [2] and Dorfman.A et.al [3], it shows that the use of medical diagnostic imaging procedures has grown rapidly, even has led to concern about low-dose ionizing radiation exposure in general population, which means that medical imaging procedure plays an important role in the process of diagnosis.

Figure 1 is a sample of how a doctor describes a patient's abdominal ultrasound. The sentence is short. However, it contains information about human body parts, medical imaging procedures, performance date and medical imaging results. To a large extent, these information could help future diagnosis. For example:

- A human body part is likely a sick organ;
- From performance date doctors can easily get to know how old the history illness is to see if there is any unusual comparing with new images;
- Medical imaging results could help the doctor determine further treatment for patients obviously;

Hence, it is essential to extract information of medical imaging procedures from text-free clinical notes.

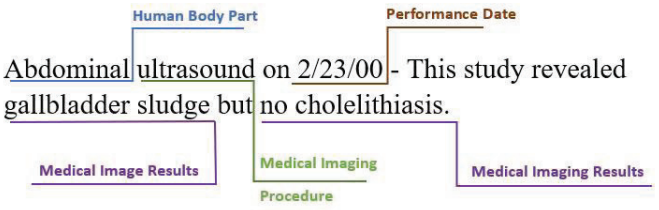

Fig. 1. A sample of how a doctor describes a patient's ultrasound information in clinical notes.

Since Metamap semantic features can classify each token into a specific type and group, our system can be easily customized by simply defining new medical named entities. We train and test our system for medication entity recognition on medication dataset and the results show that the performance is satisfied with F_1 score from 87.0% to 93.7%.

II. Related Works

Named-entity recognition (NER), also known as entity identification, entity chunking and entity extraction, is a process of information retrieval that aims to locate and identify a string of text into pre-defined categories. Sudha Morwal and et.al [4] proposed to achieve the task of NER by using Hidden Markov Model (HMM) which is a popular statistical Markov model. In this model, there are two distributions [5]: transition distribution $P(yt|yt-1)$, which tells how adjacent y value are related, and the observation distribution $P(x|y)$,

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

which tells how observed x valued related to hidden y values. HMM has a strong statistical foundation with efficient learning algorithms. However, it is only dependent on every state and its corresponding observed object.

Hai Leong Chieu and et.al [6] presented a maximum entropy-based named entity recognizer using global information. The features they used for training the model consists of 2 classes: local and global. Local features are based on neighboring tokens including token itself while global features are extracted from other occurrences of the same token in the whole document.

CRF stands for Conditional Random Fields, which is another probabilistic model for segmenting and labelling sequence data [7]. CRF offers several advantages over HMM and also avoid a fundamental limitation of maximum entropy Markov models (MEMM). This is a reason why CRF is widely used in the task of named entity recognition for information extraction. McCallum and Li [8] applied Conditional Random Fields (CRF) algorithm to extract named entities in coNLL-2003 shared task competition. A semi Markov CRF algorithm for named entity recognition is proposed by Sarawagi and Cohen [9]. They (Cohen and Sarawagi) [10] further extended the semi Markov model using dictionary and notion of similarity function.

As mentioned in the introduction section, information extraction of clinical notes is crucial for health giver for the process of diagnosis. Medical named entity recognition is still a popular topic in the field of health and informatics. Aman Kumar and et.al [11] presented an integrated approach to extracting medical entities from patient discharge summaries by modeling the system using Conditional Random Field (CRF). Other than basic word-based features, they added Boolean features to define if the present token is a medication or symptom or a specific medical condition with an extensive dictionary of medical terminologies (medical conditions, symptoms and medications) using SNOMED-CT. Finally the F_1 score they obtained of each medical named entity is around 75% compared with Patrick and Li [12] who proposed medication entity recognizer based on CRF as well with the results of 84.44%. Similarly, Yefeng Wang [13] presented a methodology of recognizing clinical named entities by using CRF with orthographic, lexical and semantic features. The author compared a rule-based and CRF-based system which achieved a F_1 score of 64.12% and 81.48% respectively.

III. Methodology

Our system of recognizing named entity of medical imaging processing is modeled based on CRF. CRF (Conditional Random Fields) is defined by Lafferty, MaCallum and Pereira [7]. It is able to efficiently achieve the task of sequence segmentation and labeling with discriminatively trained models [7]. One of the advantages of CRF is that the model is flexible enough in terms of feature selection and it is not necessary for features to be conditionally independent. We collected and manually tagged dataset from I2B2 and used 70% of them for training and the rest for testing. The features we extracted for trained CRF model are word-based features, POS features, semantic features and et.al (see B *Feature Extraction*).

1. *MRI_MDI is contraindicated_B-ST due to pacemaker.*
2. *Prostatic_B-HMB MRI_MDI showed no_B-RVL evidence_I-RVL of extracapsular extension .*
3. *Abdominal_B-HMB ultrasound_MDI on 2/23/00_B-MID - This study revealed gallbladder_B-RVL sludge_I-RVL but no_B-RVL cholelithiasis_I-RVL.*
4. *Renal_B-HMB ultrasound_MDI was scheduled_B-ST during the Intensive Care Unit stay.*
5. *Discharge medications will include Colace_B-MED 100_I-MED mg_I-MED p.o. b.i.d., Lasix_B-MED 40_I-MED mg_I-MED p.o. b.i.d., NPH_B-MED insulin_I-MED 14_I-MED units_I-MED subcutaneously b.i.d. .*

Fig. 2. Samples of manually labeled sentences. *MRI_MDI* means MRI is labeled as *MDI*. Words without any tag are considered as "No chunk" with tag "O".

A. Dataset

We sampled 1100 sentences related to medical imaging diagnostic procedures from the I2B2 2009 Medication Challenge corpus. Each token of sentences is manually classified and labeled with a tag corresponding its category (see fig 2). Table 1 illustrates all the tags and their descriptions. For comparing the performance of our system with [11] [12], we also extracted 500 sentences associated with medications entities from the I2B2 corpus for training and testing of the medication named entity recognition.

TABLE I
CATEGORIES AND TAGS OF OUR NER SYSTEM USING IOB FORMAT.

Tag Name	Reference
O	No any chunk
MDI	Medical imaging procedures
HMB*	Human body parts
MID*	Date of procedure performed
HMB*	Performed medical imaging results
ST*	Status of the current procedure
MED*	Medications and drugs

PS: * means the tag has "B-" and "I-" prefix. The "B-" prefix before a tag indicates that the tag is the beginning of a chunk, and an "I-" prefix before a tag indicates that the tag is inside a chunk.

B. Feature Extraction

Metamap is a powerful tool to recognize the UMLS (Unified Medical Language System) concepts in text. For feature extraction, we mainly use the semantic types that each token is classified into by Metamap. The following is the details of the features we extracted:

- Word-based features: Lowercase of the present word is considered as one of the features. Other than that, there are some other word-based features, such as word shape, title or not-title (whether all the case-based characters in the string following non-casebased letters are uppercase and all other case-based characters are lowercase.), prefix and suffix of present words.

- POS features: Stanford POS (Part-Of-Speech) tagger [14] [15] is a famous POS tagger for assigning parts of speech to each word.
- Semantic features: Metamap is used to identify the semantic type and semantic group of present word. For example: "MRI" is classified as *Diagnostic procedure* and *Procedures*, and tagged as "diap" and "PROC", in semantic type and semantic group respectively. If there is no any proper Metamap tags for present word, the word is then tagged as "none" and "NONE" for semantic type and group.
- We also consider if the present word is the first string of the sentence or last string.
- In order to understand the present word $x[i]$ and its context, we extracted the $x[i-2]$, $x[i-1]$, $x[i+1]$, $x[i+2]$ and their corresponding word-based features, semantic features and POS features as the features of current word.

IV. Results and Evaluation

Our system is trained on the dataset with 1100 sentences associated with medical imaging procedures, plus 500 medication-related sentences from I2B2. We split the dataset in 70/30 ratio for training and testing respectively. Precious, recall and F_1 score metrics are used to evaluate our model (See equation 1,2,3, TP,FP and FN stands for true positive, false positive and false negative respectively).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{3}$$

The results (Table 2) show that our system has different performance on different entities. For recognising "MDI", medical imaging procedures, the model works very well with the F_1 score of 92.3%, while it just has 35.7% accuracy on extracting "I-MID". The average precious, recall and F_1 score, however, are acceptable with the value of 85.6%, 86.6% and 85.4% respectively.

TABLE II
The values of precious, recall and F_1 score of each named entity.

Entities	Precision	Recall	F_1 Score
O	0.867	0.953	0.913
MDI	0.907	0.940	0.923
B-HMB	0.837	0.837	0.837
I-HMB	0.875	0.636	0.737
B-MID	0.917	0.647	0.759
I-MID	0.950	0.667	0.805
B-MIR	0.754	0.655	0.630
I-MIR	0.786	0.505	0.615
B-ST	0.545	0.316	0.400
I-ST	0.556	0.263	0.357
Avg	0.856	0.866	0.854

TABLE III
Our results of named entity of medication comparing with Kumar.A [11], Patrick.J [12] and Wang.Y [15].

	B-MED(ours)	I-MED(ours)	Kumar[8]	Patrick[9]	Wang.Y[13]
Precious	0.967	0.833	0.913	0.896	0.833
Recall	0.908	0.909	0.708	0.814	0.768
F_1 **Score**	**0.937**	**0.870**	0.798	0.849	0.799

For evaluating our system, we also compare the results with Kumar.A [11], Patrick.J [12] and Wang.Y [15] for medication entity recognition. From the Table 3, it can be easily seen our system has a better performance.

V. Conclusion

In this paper, we present a customizable Metamap-semantic-feature-based named entity recognition system of extracting entities of medical imaging procedures and medications for improving a previous work of Pictorial Visualization of EMR Summary System [1]. We used the most well-known CRF model for our system. For feature extraction, other than the word-based and POS features, we also applied Metamap to give each token extra tags of semantic types and semantic groups instead of giving a Boolean feature to identify the tokens categories [11]. Finally, our model is evaluated with precision, recall and F_1 score metrics and average values 85.6%, 86.6% and 85.4% obtained respectively. In addition, we compared the results of medication entity recognition with previous works and it shows our model has a better performance.

References

[1] Ruan, Wei, et al. "Pictorial Visualization of EMR Summary Interface and Medical Information Extraction of Clinical Notes." 2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). IEEE, 2018.

[2] Fazel, Reza, et al. "Exposure to low-dose ionizing radiation from medical imaging procedures." New England Journal of Medicine 361.9 (2009): 849-857.

[3] Dorfman, Adam L., et al. "Use of medical imaging procedures with ionizing radiation in children: a population-based study." Archives of pediatrics & adolescent medicine 165.5 (2011): 458-464.

[4] Morwal, Sudha, Nusrat Jahan, and Deepti Chopra. "Named entity recognition using hidden Markov model (HMM)." International Journal on Natural Language Computing (IJNLC) 1.4 (2012): 15-23.

[5] Dietterich, Thomas G. "Machine learning for sequential data: A review." Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Berlin, Heidelberg, 2002.

[6] Chieu, Hai Leong, and Hwee Tou Ng. "Named entity recognition: a maximum entropy approach using global information." Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002.

[7] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[8] McCallum, Andrew, and Wei Li. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons." Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003.

[9] Sarawagi, Sunita, and William W. Cohen. "Semi-markov conditional random fields for information extraction." Advances in neural information processing systems. 2005.

[10] Cohen, William W., and Sunita Sarawagi. ”Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods.” Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
[11] Kumar, Aman, et al. ”Understanding Medical Named Entity Extraction in Clinical Notes.” (2014).
[12] Patrick, Jon, and Min Li. ”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.
[13] Wang, Yefeng. ”Annotating and recognising named entities in clinical notes.” Proceedings of the ACL-IJCNLP 2009 Student Research Workshop. Association for Computational Linguistics, 2009.
[14] Toutanova, Kristina, and Christopher D. Manning. ”Enriching the knowledge sources used in a maximum entropy part-of-speech tagger.” Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13. Association for Computational Linguistics, 2000.
[15] Toutanova, Kristina, et al. ”Feature-rich part-of-speech tagging with a cyclic dependency network.” Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.
[16] 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.