Semantic based Clinical Notes Mining for Factual Information Extraction

Musarrat Hussain
Department of Computer Science and Engineering, Kyung Hee University, South Korea
musarrat.hussain@oslab.khu.ac.kr

Dong-Ju Choi
Department of Surgery, Seoul National University Bundang Hospital, South Korea
djchoi@snubh.org

Sungyoung Lee*
Department of Computer Science and Engineering, Kyung Hee University, South Korea
sylee@oslab.khu.ac.kr

***Abstract*—The existence of a substantial amount of clinical notes has raised significant demand for clinical text processing and information extraction. Clinical notes are one of the most common forms of clinical documentation and an abundant source of patient information. A list of diseases mentioned in the patient clinical notes is one of the essential factual information for experts. However, checking this information manually is a cumbersome and time consuming task. Therefore, we propose an automatic mechanism that extracts disease relevant information from clinical notes using natural language processing techniques. The methodology extracts disease information that is diagnosed in the patient and neglects the negated disease names presented in the notes. The initial evaluation results of the methodology on MTSamples provided dataset with 97.81% accuracy show its effectiveness and applicability for the mentioned goal. The methodology benefits and assists human experts by extraction disease relevant information from patient notes in a minimum time frame.**

***Index Terms*—Factual Information Identification, Clinical Notes Mining, Information Extraction**

I. Introduction

Clinical notes are a substantial source of patient information. It describes a complete detail of a patient including its previous and current health status, all the diseases screened and diagnosed, and clinical test results. Clinical notes are a type of clinical documentation which includes discharge summaries, progress notes, admission notes, operative notes, and primary care notes [1]. The purpose of clinical notes is to keep track of patient health status. Whenever, a patient visits to health care providers, the physician thoroughly check the previous clinical notes of the patient before taking any necessary steps to get complete details about the patient. One of the primary information of interest for healthcare providers is the diseases that are diagnosed and mentioned in these notes. Manually checking of these notes are time consuming and cumbersome task. Therefore, an automatic technique is required to extract disease specific information from clinical notes and save human experts time and burden. The main issue in clinical notes mining is the identification of diagnosed diseases mentioned in the patient clinical notes because the disease name can be mentioned in the form of negation. The symptoms present in a patient may indicate multiple diseases, therefore, the physician screens various options for taking a final diagnose decision. All the screened and diagnosed diseases are written in the notes. However, the experts normally only search for the diagnosed disease in the clinical notes.

In this research, our primary focus is to extract disease information from patient clinical notes using word semantics. After necessary preprocessing steps, including sentence extraction, tokenization, case transformation, and bi-gram generation, the system finds the semantic types of each token using medical knowledge source Unified Medical Language System (UMLS) meta thesaurus [2]. The system evaluates the semantic type of each token and identifies a list of initial candidate diseases. A term is considered a possible disease term if its semantic type is the same as the disease semantic type. After we find a list of diseases mentioned in notes, the disease list is then evaluated to check the existence and negation of each disease using NegEx [3] because of some diseases mentioned in the clinical notes are scanned but were not diagnosed in the patient. We filter out negative disease from the disease list and a final list of all the diseases diagnosed in a patient is presented to the human expert.

The proposed methodology is beneficial for healthcare providers to instantly process patient notes and extract a list of previously diagnosed diseases in a patient. The methodology can also be generalized for extracting other information of interest, including medication information, temporal event detection.

II. Related Work

Information extraction from clinical notes is one of the key research areas [1]. A lot of efforts have been made and a number of applications including cTAKES [4], and HITEx [5] have been developed to leverage natural language processing (NLP) techniques for information extraction embedded in clinical notes [6]. Information for integrating biology and the bedside (i2b2) has organized various challenges related to concept extraction, assertion classification, and relation classification [7] to extracted precious information from clinical notes. Some of the applications related to the proposed one are described as follows.

H. Xu et al. developed a tool MedEX [8] intended to extract medications and drug information from patient discharge summaries and visit notes. MedEx accurately extracts drug names and other related information including drug

strength and frequency. The system processes the clinical narratives in three steps; preprocessing, semantic tagging and parsing. Preprocessing split the narrative into sentences, semantic tagging break down each sentence into tokens, and tag each token with its semantic category like drug name, drug strength, etc. While parsing parses the textual sentences into the structured form using chart parser and context-free grammar. The parse structured notes are then used to find out drugs related information.

Y. Xu et al. [9] proposed and designed an end-to-end system for the i2b2 temporal relation extraction challenge. The objective of the system was to extract temporal information that can be utilized in building a timeline for each event. The timeline helps in tracking patient status, diseases, disease causes, operations, and side-effects of drug related information. The system consists of machine learning based event extractor, name entity extractor, and temporal content extractor. The author trained and test a support vector machine and Markov logic networks, machine learning models for achieving the aforementioned goal.

III. Proposed Methodology

In this research, we present clinical notes mining methodology for disease extraction. The methodology identifies and extracts disease information from textual clinical notes in three steps: preprocessing, disease identification, and negated disease filtration as shown in Figure 1. Preprocessing step prepare notes to subsequent steps ready format (tokens in our case). *Disease identification* step identify and extract disease names mentioned in the notes. While *negated disease filtration* step analyze the sentences containing the disease name to check its affirmation or negation. The negated diseases are filtered out and the final list of the disease present in the patient notes is presented. The details of each step are provided in the subsequent subsections.

A. Preprocessing

Preprocessing is one of the essential steps for each text mining applications. Generally, preprocessing transform the provided text to a desired and subsequent process ready format. In the proposed methodology, the primary objective of this component is to split the textual clinical note to sentences and identify the tokens of each sentence. The individual token is not semantically beneficial, therefore, we considered the n-gram technique to process a token in a given context. The $n = 2$ produced efficient result. Therefore, we create tokens with bi-gram. The overall essential preprocessing steps performed include loading a note, split it into sentences, tokenized each sentence, transform the case of each token to lowercase, and identify the bi-gram of the tokens. We did not remove stop words from the sentences as it helps in disease negation identification. The preprocessed sentences and tokens are then processed by other components for disease identification.

B. Disease Identification

The objectives of *disease identification* is to examine the patient preprocessed clinical notes and extract all diseases (scanned and diagnosed) mentioned in the notes. We achieve this goal by utilizing the semantic type of tokens. We identify semantic types of each token from meta thesaurus of UMLS dictionary. We also have identified a list of semantic types that are considered as diseases. The identified semantic type of each token is mapped with the disease semantic type list. A token semantic type matched with disease semantic type is considered as disease term. After evaluation of all tokens semantic types, a list of possible disease terms are extracted from the provided text.

C. Negated Disease Filtration

A disease mentioned in patient clinical notes is not necessarily the one diagnosed in the patient. Clinical service providers scan for various diseases based on signs and symptoms and the initial information provided by the patient in the early stages of disease diagnosis. However, they reach the final decision of a disease diagnosed after various test results and clinical evaluations. All the necessary steps and diseases evaluated are mentioned in patient notes. Therefore, we device *negated disease filtration* to re-evaluated and check the negativity of the early extracted possible disease list. This component checks the sentence of each extracted disease and reviews its negativity by utilizing NegEx [3]. NegEx provides information about the positive and negative aspects of the disease. We filter out all diseases that are used in negative aspects in the clinical notes. The final list contains only the disease list that is diagnosed in the patient in various stages.

IV. Results

We evaluate our proposed methodology on a subset of clinical notes provided by Medical Transcription Sample (MTSamples) [10] to check its effectiveness in terms of accuracy. The dataset utilized consists of 1050 notes in 20 various clinical domains including urology, nephrology, radiology, orthopedic, and cardiovascular etc. In the initial settings of evaluation, we only apply preprocessing steps and disease identification which extract disease information with 86.23% accuracy. The analysis of the result shows that there are diseases which are scanned but were not diagnosed in the patient. However, our methodology has detected those diseases as well, which is not of interest to the healthcare providers. To avoid this, we add *Negated Disease Filter* to the methodology which increases the accuracy to 97.81%. The confusion matrix and detail measure of the proposed methodology evaluation is shown in Table I and Table II respectively.

TABLE I
CONFUSION MATRIX OF PROPOSED METHODOLOGY

	True Positive(TP)	**True Negative (TN)**
Predicted Positive	3289	49
Predicted Negative	337	14023

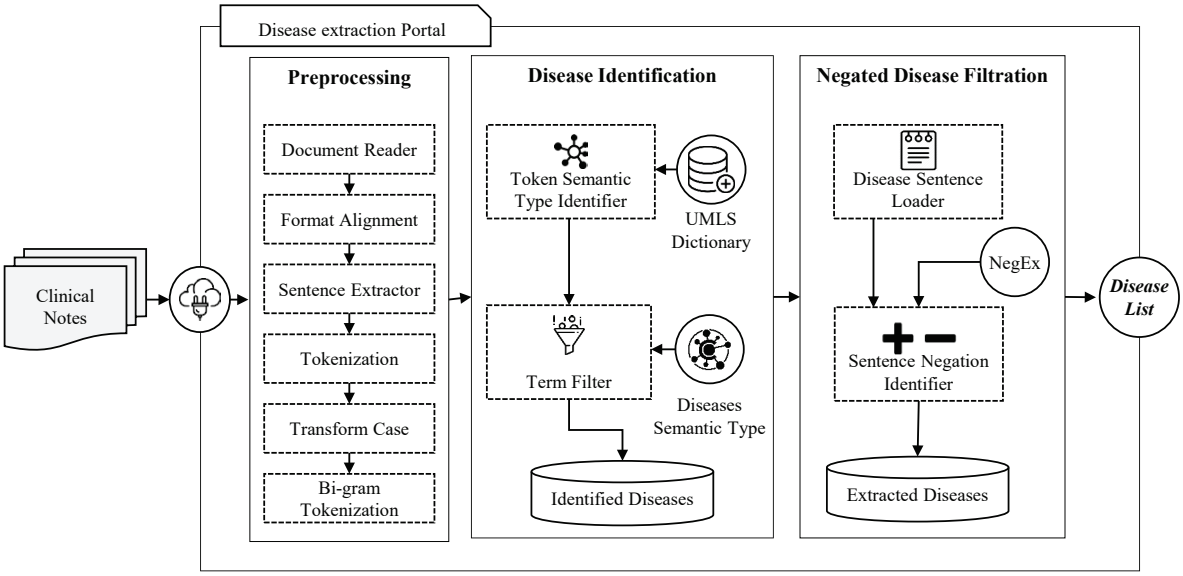

Fig. 1. Proposed methodology for disease extraction.

TABLE II
DETAILS OF RESULTS ACHIEVED

Measure	**Value(%)**	Derivations
Sensitivity	90.71	TPR = TP / (TP + FN)
Specificity	99.64	SPC = TN / (FP + TN)
Precision	98.50	PPV = TP / (TP + FP)
Negative Predictive Value	97.65	NPV = TN / (TN + FN)
False Negative Rate	09.29	FNR = FN / (FN + TP)
Accuracy	97.81	ACC = (TP + TN) / (P + N)

The primary benefit of the methodology is to reduce the burden of healthcare providers, reduce the time required for scanning all the notes, provided accurate patient health status in terms of disease.

V. CONCLUSION AND FUTURE WORK

Clinical notes provide credible, accurate and detailed information about patient health. The information presented in these notes can assist healthcare providers to get complete detail about a patient. However, these notes are written in textual format which is one of the main causes of it under-utilization. To overcome this deficiency, we proposed a semantic analysis based technique for disease extraction from clinical notes. The proposed technique analyze and extract disease diagnosed in a patient which can help physician by providing initial health status of the patient and can guide him for taking further necessary steps. In the future, we are going to extend and test the methodology generalization for extracting any factual information including sing and symptoms, drug information, activities taken, and follow up scheduled.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629) supervised by the IITP(Institute for Information & communications Technology Promotion), IITP-2017-0-00655, NRF-2016K1A3A7A03951968, and NRF-2019R1A2C2090504.

REFERENCES

[1] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn *et al.*, "Clinical information extraction applications: a literature review," *Journal of biomedical informatics*, vol. 77, pp. 34–49, 2018.

[2] "Unified medical language system (umls)," https://uts.nlm.nih.gov/metathesaurus.html, accessed: 2019-07-19.

[3] "Extending negex with kernel methods for negation detection in clinical text."

[4] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[5] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC medical informatics and decision making*, vol. 6, no. 1, p. 30, 2006.

[6] Y. Wang, S. R. Steinhubl, C. Defilippi, K. Ng, S. Ebadollahi, W. F. Stewart, and R. J. Byrd, "Prescription extraction from clinical notes: towards automating emr medication reconciliation," *AMIA Summits on Translational Science Proceedings*, vol. 2015, p. 188, 2015.

[7] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[8] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "Medex: a medication information extraction system for clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, no. 1, pp. 19–24, 2010.

[9] Y. Xu, Y. Wang, T. Liu, J. Tsujii, and E. I.-C. Chang, "An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 849–858, 2013.

[10] "Medical transcription sample (mtsamples)," https://www.mtsamples.com, accessed: 2019-08-28.