SIGNIFICANT INFORMATION EXTRACTION OF MEDICAL DATA THROUGH BOOTSTRAPPED PATTERN LEARNING AND ARTIFICIAL INTELLIGENCE TECHNIQUES

[1]Manisha Mali, [2]Dr. Niraj Sharma
[1]Research Scholar, [2]Associate Professor
[1,2]Dept. of CSE, SSSUTMS, Sehore, MP, India

Abstract— Medical (also known as clinical) data refers to health-related information associated with regular patient care or as part of a clinical trial program. There are many categories of such data; These categories include radiology reports, pathology investigation reports, Doctor's advice or prescriptions, and many more. Among these reports, radiology reports and pathology reports are the most critical medical documents that a diagnostician looks into, especially in case of emergency. They provide the physicians with crucial information regarding the patient's condition and help them take immediate action in critical situations. However, these reports exist in the form of unstructured text, which makes them time-consuming for humans to interpret. So there is a need to extract the medically necessary information and their importance from these reports. We have proposed a system that will extract the medically significant information in structured form and a level of importance and classify the overall report into critical or non-critical categories, which help doctors identify potential high-priority reports. Our proposed system is based on Bootstrapped Pattern Learning for information extraction and Artificial Intelligence Techniques for classifying data into critical and non-critical categories.

Index Terms: Information Extraction, Clinical data, *Bootstrapped Pattern Learning*, Artificial Intelligence

I. INTRODUCTION

With the document sets for each domain at hand, we can start extracting information from them. When performing automatic IE from texts, the first decision is whether the extraction process should be supervised or unsupervised. In unsupervised IE, the algorithm's outcome is defined only by the processed documents themselves, with only a few possibilities influencing the algorithm´s outcome. It is easier to define the desired output format in supervised IE by training the system with a hand-labeled (or differently created) reference standard. The automatic labeling algorithm learns from the provided training data and uses the acquired knowledge to label the remaining documents the way the training data was labeled. For this technique, the documents from the training corpus and the remaining documents must share the same structure. Electronic health records (EHR), Administrative data, Claims data, Patient / Disease registries, Health surveys, and Clinical trials data are the six major types of clinical data. Digitalization of Health Information has brought up the rapid adoption of Electronic Health Record systems at hospitals and clinics. Large amounts of detailed longitudinal patient information, including lab tests, medications, disease status, and treatment outcomes, have been accumulated and are available electronically and become valuable data sources for clinical and translational research. The best-known form of clinical data is the electronic health record, which contains lots of important information. An electronic health record is the digital version of a person's medical information and history maintained over time by a healthcare provider. Medical data found in an electronic health record can include general numerical information, such as vital signs like heart rate, respiratory rate, and temperature, Diagnostic-related information, like laboratory test results from blood tests, genetic tests, culture results, and so on. It can also include imagery like x-rays, Treatment information. For instance, is the person receiving any medication? If so, how much (what dose) and how often? and so on. Other forms of clinical data include 1. Administrative data, i.e., non-clinical research data focused on record-keeping surrounding a service, such as a hospital discharge information. This can be part of an electronic health record as well. 2. Claims data, which is information regarding insurance claims. 3. Patient/disease registries, which help collect and track clinical information of defined patient populations. 4. Health surveys can help evaluate or tally statistics like saying the most common chronic illnesses a nation faces. 5. Clinical trial data, which is clinical information that has been gathered thanks to experiments associated with clinical research. A well-known challenge when using EHR data for research is that large amounts of detailed patient information are embedded in the clinical text (e.g., clinical notes and progress reports). Information extraction, one of the popular Natural language processing (NLP) technologies, can unlock information embedded in clinical text by automatically extracting desired information (e.g., cancer stage information, disease characteristics, and pathological conditions) from the text. Many successful studies applying information extraction techniques have been reported, ranging from phenotyping, detecting adverse events, improving healthcare quality, facilitating genomics research such as gene-disease association analysis, and pharmacogenomic studies. Clinical information extraction applications can be developed using either symbolic techniques or statistical machine learning. Applications built based on symbolic techniques involve handcrafted expert rules, such as regular expressions and logic rules. It has been shown effective in the clinical domain due to the clinical sublanguage characteristics. However, rule-based applications can be expensive and cumbersome to develop, requiring the collaboration between NLP experts and healthcare professionals, and the resultant applications may not be portable. Meanwhile, machine learning approaches are efficient and effective for clinical information extraction tasks, such as disorder normalization, gout flares extraction, cancer identification, or thromboembolic disease identification. Despite their impressive improvements, large amounts of manually labeled training data are a crucial building block and a key enabler for a successful machine learning model. However, such large labeled training data is not always readily

available and usually expensive to create due to human annotation. This problem becomes more significant in the clinical domain, mainly due to i) the lack of publicly available clinical corpora because of privacy concerns and ii) the annotation of clinical text requiring medical knowledge. Therefore, popular methods for creating labeled training data for clinical information extraction tasks, such as crowdsourcing, are not applicable. The distant supervision strategy has been utilized to create large training data in the literature quickly. It uses existing resources or heuristics to generate weakly labeled training data, which has been applied in common NLP tasks, including relation extraction, knowledge base completion, sentiment analysis, and information retrieval. In addition to the need for large labeled training data, machine learning requires feature engineering, where each training instance needs to be transformed into a feature vector representation. Recently, deep representation learning has become popular due to its capability of representing the raw data (such as the pixel values of an image or words in a textual document) in a high-level representation or feature vector. In NLP, word embeddings are one of the most successful deep learning technologies to capture high-level semantic and syntactic properties of words. Word embeddings have been utilized in various clinical NLP applications, such as clinical abbreviation disambiguation named entity recognition and information retrieval [Yanshan Wang et al.].

The rest of the paper is organized as follows. Bootstrapped Pattern Learning algorithm and Artificial Intelligence are presented in sections II and III, respectively. A literature survey of work done in information extraction of clinical data is presented in section IV. The proposed system is described in section V. Finally, and the conclusion has discoursed in Section VI.

II. BOOTSTRAPPED PATTERN LEARNING ALGORITHM

Bootstrapped Pattern Learning is used for learning patterns to learn entities of given entity types from unlabeled text, starting with seed sets of entities. Though there is significant variation between implementations, bootstrapping approaches in natural language processing all follow the same general format:

1. Start with an empty list of things. Usually words or phrases, but it can be any representation of language (such as regular expressions, tuples, etc.)
2. Initialize this list with carefully chosen seeds (the initial set of data).
3. Leverage the things in the list to find more items from a training corpus.
4. Score those newly found things; add the bestones to the list.
5. Repeat step 3. Stop after a set number of iterations or some other stop condition. The intuition for bootstrapping stems from the observation that words from the semantic category tend to appear in similar patterns and similar contexts. For example, the words “water” and “soda” are both in the semantic BEVERAGE category, and in a largetext, they will likely both be found in phrasescontaining drank or imbibed. The core intuition is that by searching

The core feature of a bootstrapping algorithm isthat each iteration is fed the same type of data as its input that it produces as its output. The output of the first iteration is used as the input of the second iteration, and so on. It should not be surprising that choosing the initial set of data (the seeds)from which all other data is “grown” is a critical factor (arguably the most critical factor) in the performance of the algorithm.[Daniel Waegel] At their core, bootstrapping algorithms are not limited to learning general semantic categories. They help understand any variety or relationship for which words appear in similar linguistic phrases or contexts.

Weaknesses of Bootstrapping include Semantic Drift and Stop conditions. A very well-known flaw in bootstrapping is a phenomenon known as semantic drift or creep. This occurs when, after multiple iterations, the bootstrapper wanders away from the original semanticmeaning of the seeds and begins to accept incorrect or undesirable entities. This can occur if entities have more than one semantic word sense, it can also happen if words similar meaning and usage buthave different undertones. Another major drawback that was bootstrapping suffers isthe difficulty in knowing when to stop the iterativecycles. Some quit after a set and seemingly arbitrary number of iterations, while others [Lin et al., 2003] stopped when theiralgorithm had exhaustively added all candidates or rejected any remaining candidates. Finally, the choosing of seeds is arguably the most critical step in bootstrapping. Most authors chose the most frequently occurring words in their corpus that they have quickly identified belong to the category they are interested in. While this ensures that the most significant amount of contextual information will be available to learn from, it does nothing to ensure the quality of the contexts. It is easy to imagine improperly chosen seeds that would pick up tons of extraction patterns that, in turn, extract feeble additional words, ultimately producing poor results.

III. ARTIFICIAL INTELLIGENCE

AI is a branch of computer science by which we can create intelligent machines that can behave like humans, think like humans, and make decisions on their own. Artificial Intelligence is composed of two artificial intelligence, which defines something made by humans, and intelligence, which refers to the ability to think independently. Hence, this makes artificial intelligence “thinking power made by humans.” Natural language processing (NLP) is a significant area of artificial intelligence research, which in turn serves as a field of application and interaction of several other traditional AI areas. Until recently, AI applications in NLP focused on knowledge representation, logical reasoning, and constraint satisfaction - first applied to semantics and later to grammar. However, in the last decade, a dramatic shift in NLP research has led to the prevalence of very large-scale applications of statistical methods, such as machine learning and data mining. Naturally, this also opened the way to the learning and optimization methods that constitute the core of modern AI, most notably genetic algorithms and neural networks. In this paper, we overview the current trends in NLP and discuss the possible applications of traditional AI techniques and their combination in this fascinating area.

IV. LITERATURE SURVEY

[Nidhin Nandhakumar et al., 2017] proposed a system that performs extraction of medical phrases and their criticality level from free-text radiology reports and classification of the complete information as being critical or not. As radiology reports are dictated by the radiologists and transformed into text, spelling and joined-word errors appear in the text, which they automatically correct, aiming to improve phrase extraction and classification accuracy. Information extraction from the radiology reports, in the form of medical phrases, is complex but provides valuable data, which can be further used in populating structured databases for data mining tasks. The complexity of their job is due to assigning the criticality level based on the textual context of the extracted phrases. Based on conditional random fields, the information extraction model extracts medical phrases and the associated

criticality level (high-critical, critical and non-critical). The model is trained on a small corpus of reports labeled by two emergency physicians. They demonstrated that their approach achieves performance that is comparable to the inter-annotator agreement. Furthermore, using the extracted medical phrases as features, they address the report classification task that classifies entire radiology reports as critical or non-critical (i.e., whether an emergency physician needs to take immediate action on them).To allow the emergency physician user to efficiently inspect the extracted medical phrases and correct them if needed, they have built an adaptive active learning interface.

Searching for a cure for cancer is one of the most vital quests in modern medicine. In that aspect, microRNA research plays a key role. Therefore, keeping track of the shifts and changes in established knowledge in the microRNA domain is very important. [Nisansa de Silva et al., 2017] introduced an Ontology-Based Information Extraction method to detect occurrences of inconsistencies in microRNA research paper abstracts. The primary research contribution of this study was to use ontology-based information extraction to observe how inconsistencies rise in the literature about previously established knowledge in a scientific field. .is study successfully proposed a method to do that observation and succeeded in finding 503 such inconsistencies in a corpus of 39149 research paper abstracts. Since these inconsistencies are rooted in very domain-specific medical jargon, medical experts need to be analyzed before getting incorporated into future studies. They proposed a method to use Ontology for microRNA Targets (OMIT) to extract triples from the abstracts. Then they introduce a new algorithm to calculate the oppositeness of these candidate relationships. Finally, they presented the discovered inconsistencies in an easy-to-read manner to be used by medical professionals. Their study is the first ontology-based information extraction model introduced to find shifts in the established knowledge in the medical domain using research paper abstracts. However, their study had to face the problem of the ontology that was being used not having the relationship rules that most of the established OBIE systems use. Their study came up with a novel way to solve this problem by involving open information extraction systems to extract the relationships and then using the conventional OBIE systems to do the information extraction. This methodology can be considered a new way of doing OBIE and the traditional and established methods. They downloaded 36877 abstracts from the PubMed database. From those, They found 102 inconsistencies relevant to the microRNA domain.

The objective of [Yanshan Wang et al.] research work is to automatically create sizeable labeled training datasets and reduce feature engineering efforts for training accurate machine learning models for clinical information extraction. To achieve this, they proposed a deep representation empowered distant supervision paradigm for information extraction from free-text EHRs. This paradigm utilizes rule-based NLP algorithms as weak labels to curate a large set of training data and leverage pre-training word embedding features to represent the training data for machine learning methods. Although the training data is weakly labeled, they theoretically show that machine learning models trained from these weak labels can achieve similar training performance to that trained from accurate labels. They validated the effectiveness of the proposed paradigm using two clinical natural language processing tasks: a smoking status extraction task and a fracture extraction task and tested three prevalent machine learning models, i.e., Support Vector Machine (SVM), Random Forrest (RF), and Convolutional Neural Network (CNN). Both experiments show that CNN is the best fit in the proposed paradigm that could outperform the rule-based NLP algorithms. They also verified the advantage of word embedding features in the proposed paradigm over two widely adopted features: tf-idf and topic modeling. Moreover, They found that the CNN captures different extraction patterns compared with the rule-based NLP but is more sensitive to training data size.

For epidemiological research, the usage of standard electronic health records may be regarded as a convenient way to obtain large amounts of medical data. Unfortunately, large parts of clinical reports are in written text form and cannot be used for statistical evaluations without appropriate pre-processing. This functionality is one of the main tasks in medical language processing. So [Georg Fette et al.] presented an approach to extract information from medical texts and a workflow to integrate this information into a clinical data warehouse. Their technique for information extraction is based on Conditional Random Fields and keyword matching with terminology-based disambiguation. They developed a structured approach for the homogeneous integration of different data domains used in clinical routine into a DW. They described the general architecture of the system and the workflow for extracting information from unstructured text domains and their integration into the DW. The IE from the domain of echocardiography reports already shows satisfying f1 score results. Thus, the extracted information can already be reliably integrated into clinical studies or applied to clinical research questions. The other domains show promising results but still have to be improved to use them for clinical studies with the same reliability. They compared two IE methods: CRF and keyword matching with terminology-based disambiguation. They evaluated both ways on a selected set of text domains and yielded very encouraging results.

Artificial intelligence (AI) has been developing rapidly in recent years in software algorithms, hardware implementation, and applications in a vast number of areas. In this review, we summarize the latest developments of applications of AI in biomedicine, including disease diagnostics, living assistance, biomedical information processing, and biomedical research. This review aims to keep track of new scientific accomplishments, understand the availability of technologies, appreciate the tremendous potential of AI in biomedicine, and provide researchers in related fields with inspiration. It can be asserted that, just like AI itself, the application of AI in biomedicine is still in its early stage. New progress and breakthroughs will continue to push the frontier, widen AI applications' scope, and fast developments are envisioned shortly. Two case studies are provided to illustrate the prediction of epileptic seizure occurrences and the filling of a dysfunctional urinary bladder. [Guoguang Rong et al., 2020] reviewed the latest developments in the application of AI in biomedicine, including disease diagnostics and prediction, living assistance, biomedical information processing, and biomedical research. AI has exciting applications in many other biomedical areas as well. It can be seen that AI plays an increasingly important role in biomedicine, not only because of the continuous progress of AI itself but also because of the innate complex nature of biomedical problems and the suitability of AI to solve such problems. New AI capabilities provide novel solutions for biomedicine, and the development of biomedicine demands new levels of ability from AI. This match of supply and demand and coupled results will enable both fields to advance significantly in the foreseeable future, ultimately benefiting the quality of life of people in need.

V. PROPOSED SYSTEM

We have proposed a system that will extract the medically significant information in structured form and the level of importance and classify the overall report into critical or non-critical categories. This helps doctors to identify potential high-priority reports. As shown in Fig. 1, our proposed system consists of five steps. In the first step, medical reports in the unstructured form will be

provided as input. Input reports will be in unstructured form. So they need to undergo pre-processing step. So in the second step, pre-processing will be done for document preparation. In the third step, feature extraction will be done on Word and Sentence level to retrieve medical phrases. In the fourth step, information extraction will be done using the Bootstrapped Pattern Learning method. Artificial Intelligence Techniques will be used for classifying data into critical and non-critical categories in the last and fifth steps. This classification result will help medical practitioners in emergencies. Also, there are many uses of these information extractions.

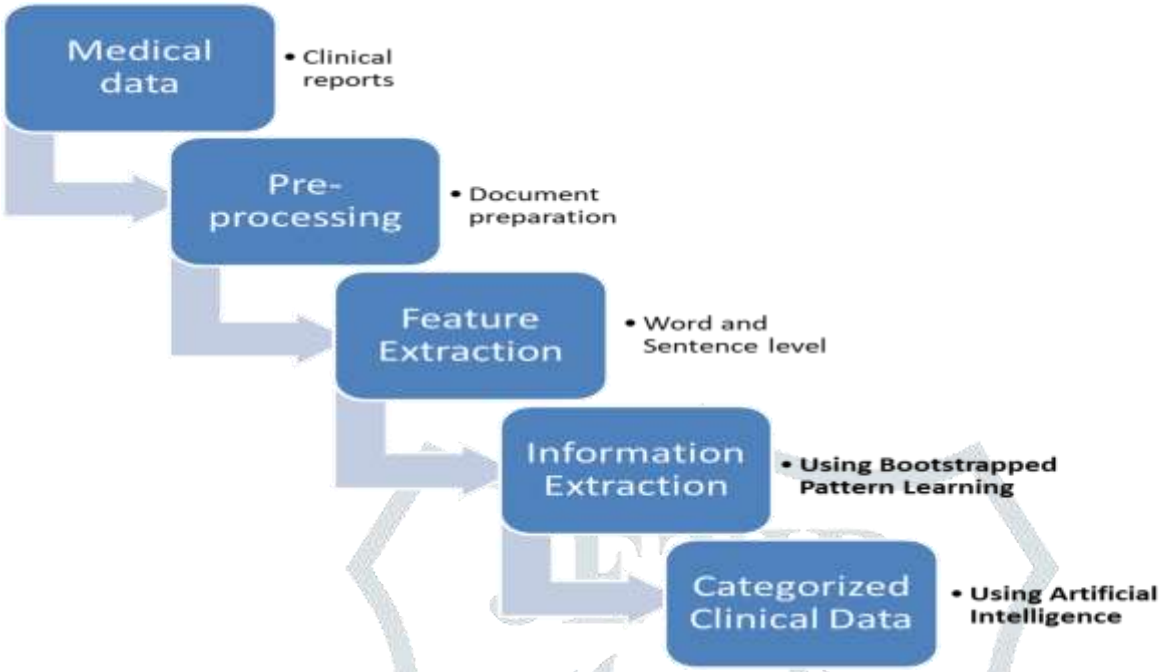

Fig. 1: Proposed System Architecture

For our system, we are going to consider medical reports which can belong to the following categories :

- Disease-Related Group (DRG) Diagnoses
- Sonography Reports
- Radiology Reports
- Patients medical history
- Diagnoses From Discharge Letters
- Pathology Laboratory Reports
- Electrocardiogram Reports
- 24-Hour Holter ECG Reports
- 24-Hour Blood Pressure Reports
- Bicycle Stress Test Reports
- Ergospirometry Reports
- 6-Minute Walk Test Reports
- Lung Function Reports
- Coronary Angiography Reports
- Transthoracic Echocardiography Reports
- X-Ray Reports
- Endoscopy Reports
- Medical Treatment Reports
- Magnet Resonance Tomography Reports
- Therapy Reports
- Anamnesis And Physical Examination Reports

VI. CONCLUSION

In today's world, colossal information is available in digital form. However, most of the information is present in unstructured form. To extract meaningful information from this data, we need to process this to be transformed into a structured one. This paper focuses on a Significant Information Extraction of Medical Data through Bootstrapped Pattern Learning and Artificial Intelligence Techniques. This paper proposed a system that will extract medical phrases and clinically significant information like their criticality level from various clinical reports and categorize data into critical and non-critical categories. We believe from our rigorous literature survey that our proposed system will produce promising results, as we will use well-proven techniques like Bootstrapped Pattern Learning for information extraction and Artificial Intelligence Techniques for classification.

REFERENCES

[1] Daniel Waegel, "A Survey of Bootstrapping Techniques in Natural Language Processing"

[2] Guoguang Rong, Arnaldo Mendez, Elie Bou Assi, Bo Zhao, Mohamad Sawan, "Artificial Intelligence in Healthcare: Review and Prediction Case Studies," https://doi.org/10.1016/j.eng.2019.08.015, 2020

[3] Georg Fette, Maximilian Ertl, Anja Wörner, Peter Kluegl, Stefan Störk, Frank Puppe, "Information Extraction from Unstructured Electronic Health Records and Integration into a Data Warehouse"

[4] Lin, Winston, Yangarber, Roman, and Grishman, Ralph. 2003. Bootstrapped learning of semantic classes from positive and negative examples. In Proceedings of ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data, Volume 4, No. 4.

[5] Nidhin Nandhakumar, Ehsan Sherkat, Evangelos E. Milios, Hong Gu, Michael Butler, "Clinically Significant Information Extraction from Radiology Reports"

[6] Nisansa de Silva, Dejing Dou, Jingshan Huang, "Discovering Inconsistencies in PubMed Abstracts through Ontology-Based Information Extraction," ACM-BCB'17, August 20–23, 2017, Boston, MA, USA

[7] Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J. Atkinson, Shreyasee Amin, Hongfang Liu, "A Deep Representation Empowered Distant Supervision Paradigm for Clinical Information Extraction"