Information Extraction from Clinical Text using NLP and Machine Learning: Issues and Opportunities

M. Sridevi
Assistant Professor
Dept. of MCA, BMSIT&M
Bengaluru, India

Arunkumar B.R.
Professor & Head
Dept. of MCA, BMSIT&M
Bengaluru, India

ABSTRACT

Natural Language Processing (NLP) and Machine Learning concepts are gaining rapid importance in the era of digitalization of data. The value of data keeps changing over time and makes it important to harness that value for performing in depth research in various domains. Extracting information from clinical text helps in automated terminology management, data mining, de-identification of clinical text, research subject identification and studying effect of research on them, predicting the onset and progress of various chronic diseases, disease-treatment-side effect analysis etc. Methods based on NLP and Machine Learning tends to perform better in this area but more experience is required to analyse clinical text than the biomedical literature. The issues and opportunities in information extraction from the clinical text need to be intensively reviewed to find new avenues in this domain of research.

General Terms

Clinical Text Mining, Natural Language Processing, Machine Learning.

Keywords

Natural Language Processing, Machine Learning, Clinical Text, Information Extraction, Electronic Health Records, Viterbi Algorithm.

1. INTRODUCTION

Healthcare related NLP (Natural Language Processing) laid doors open to Medical Language Processing. The data that is available in health care domain is mostly available in narrative form which is a culmination of dictated transcriptions, direct entry, or usage of speech recognition applications. This kind of data constituting free text expresses events and concepts in a convenient way, but is not friendly for summarization, searching, statistical analysis, or decision-support. In order to extract information, some pre-processing is required. Pre-processing includes document structure analysis, tokenization, part-of-speech tagging, spell checking, sentence splitting, word sense disambiguation (WSD), and some form of parsing. Situation dependent features like event subject identification, temporality, and negation play a crucial role for appropriate interpretation of the information that is extracted. Techniques like simple pattern matching, processing methods based on symbolic information and rules, or based on machine learning and statistical methods can be used for information extraction. The information thus extracted can then be related to concepts in the standard terminologies and can be used for analysis. This information can be used for enriching the EHR (Electronic Health Record) and for further decision support.

This paper focuses on the issues that transform into hurdles, and opportunities that lead to new avenues in information extraction from clinical text. Issues with information extraction become more intricate with respect to life style diseases like Alzheimer's and Cancer and open up more opportunities as these diseases do not have standard symptoms, defined progression, common diagnosis, and standard treatment. The symptoms, diagnosis and treatment patterns vary from person to person, one geographic region to another, and from one's lifestyle to another lifestyle. This kind of information can better be extracted from clinical text than from biomedical text which is standard and static. Clinical texts are texts that are written by clinicians in the clinical environment. These texts explain about subjects (patients), the pathologies, and their social, personal and medical histories, findings made during procedures or during interviews, and so on. Using NLP techniques to extract information from clinical text available for the life style diseases poses many challenges which will be discussed in the further sections of this paper. Section 2 provides a review on the literature, section 3 about information extraction through pattern matching, section 4 about restrictions on shareable clinical text, section 5 about contextual analysis, section 6 about opportunities in clinical text mining based on Hidden Markov Model, section 7 about application of Viterbi algorithm in machine learning for clinical text with an example thus reaching the conclusion section.

2. LITERATURE REVIEW

EHR mining or Electronic Health Record mining has a potential to establish new patient classification principles and discovering unknown correlations in diseases. But, a broad range of legal, ethical, and technical reasons currently create hurdles for the systematic deposition of the data in Electronic Health Records and their mining [1]. EHRs clinical text mining provides assistance for Clinical Decision Support (CDS). The goal of CDS is to "help clinicians make decisions, manage medical data about the patients or with the knowledge of medicine required to analyse and interpret such data [2]". NLP plays a crucial role in using clinical text information to drive CDS, representing clinical knowledge and CDS interventions in standardized formats, and leveraging clinical narrative [3].

Even highly developed NLP systems are constructed on the foundation of identifying words or phrases as medical terms that illustrate the domain concepts like named entity recognition and understanding correlations between the identified concepts. [4] discussed about a large scale project, the Linguistic String Project Medical Language Processor (LSP-MLP) at the University of New York, which enabled the extraction and summarization of symptoms and the drug information, and identification of the potential medication side effects.

Special Purpose Radiology Understanding System (SPRUS) discussed in [5] was the first NLP application which was developed by the University of Utah – Medical Informatics

Group. It was only a semantically driven system. The SymText (Symbolic Text Processor) that was developed later and was discussed in [6] was equipped with probabilistic and syntactic semantic analysis. SymText used semantic analysis that used Bayesian Networks.

Any biomedical NLP systems that are used to extract information from clinical narrative reports show drop in the performance when applied to another institution different than the one where it was developed. But some adjustments will make the system perform as well as in the original institution. The NLP systems discussed above required considerable resources for development and implementation. To overcome this issue, several researchers experimented progressively with simpler systems that focused on specific information extraction tasks and on a limited set of information to extract. These more focused systems presented good performance statistics and now form the majority of the systems used for information extraction.

A variety of methods have been adapted in general and biomedical literature domains for fact extraction from free text and to fill template slots. A typical information extraction system consists of a combination of the following components as described in [7], namely tokenizer, part-of-speech tagger, sentence boundary detector, morphological analyser, gazetteer, shallow and deep parser, NER, template extractor, discourse module and template combiner. In the hierarchy, the performance of the lower level components, most of the time, determine the performance of the higher level components like discourse module, template extractor, and template combiner.

The context dependency of chronic diseases like Alzheimer's and Cancer pose greater challenges to NLP to process the clinical text due to high variability in the symptoms, their progress, demographic factors, geographic factors, diagnosis methods, treatment patterns etc. This paper discusses about the challenges and opportunities that would be encountered at different stages of clinical text mining.

3. INFORMATION EXTRACTION THROUGH PATTERN MATCHING

Pattern matching technique exploits the basic patterns over various structures – text strings, semantic pairs, part-of-speech tags, and dictionary entries [8]. But the pattern matching approaches lack generalizability that limits their applicability to new domains.

Another approach is to use shallow and full syntactic parsing. The non-robust performance of the parser is a prominent issue as clinical text has different features than the general English. This difference between medical and general English has led to the improvements in sub-language driven approaches, which frame and exploit a particular set of constraints of the sub language [9]. But these sub language approaches are not easily transferrable to new domains. Machine learning techniques hold promising results in clinical domain also, but they require huge annotated corpora for training, which are not only expensive but also time consuming to generate.

4. LIMITED ACCESS TO SHAREABLE CLINICAL TEXT

The information extraction from clinical and medical domain has lagged behind due to limited access to clinical data that is shareable. The constraints imposed to protect the patient confidentiality are the main obstacle. The major challenge lies in creating a large vibrant community around the shared data, annotation guidelines, tasks, annotations and techniques of evaluation.

The HIPAA (Health Insurance Portability and Accountability Act) of the United States protects the confidentiality of the patient data. The Common Rule protects the research subject's confidentiality. Similar type of confidentiality is provided by the European Union Data Protection Directive. These laws require the informed consent of the patient to use their data for the research purposes. But these requirements can be waived off if de-identification of the data is performed. De-identification means removal or hiding of explicit identifiers. In [10], the author have evaluated the time cost to de-identify narrative text notes manually, and came to a conclusion that it was time-consuming and complex to exclude all the Protected Health Information (PHI) required by HIPAA.

To resolve the above discussed issue, many systems were developed for automated de-identification of documents of narrative text from the EHRs. The Scrub System discussed in [11] hides personal identifying information like names, contact information, age, etc. A specific algorithm was used to detect each specific entity using a list of all possible values. A system built for disambiguation is illustrated in [12] which detects and replaces all instances of titles and names. This MEDTAG system used a lexicon to tag semantic types, and disambiguation rules that are manually written.

An open source system was developed by Beckwith et al. [13] which removes PHI from pathology reports and named it as HMS Scrubber. This system initially removed all the identification information from the report headers that were also included in the body of the report. Then 50 regular expressions were used to detect and remove addresses, dates, names cited with markers such as MD, PhD, etc. and accession numbers. This systemhas been evaluated to have removed 98.3% of the patient information present in nearly 1800 pathology reports.

In [14], the authors have first created an 889 "de-identified" and "re-identified" discharge summaries corpus. They have used realistic surrogates for re-identification of the discharge summaries. The identifying information was first tagged by using statistical NER (Named Entity Recognition) techniques. This system was based on Simple Vector Machines using local context and few dictionaries. Overall, methods that are based on dictionaries performed better with PHI but are difficult to generalize.

5. EXTRACTION OF CONTEXT OF THE CONCEPTS (CONTEXTUAL ANALYSIS)

The contexts of the concepts that are being extracted from narrative text documents play a critical role. The contextual information may include temporality (eg. "_ _ stroke 2 years ago _ _"), negation (eg. "denies any knee pain"), and the event's subject identification (eg. "his father has blood pressure"). NegExpander [15] was a program detecting negation terms and later expanding the concepts related to it. Much more complex system, NegFinder [16] used indexed concepts and used UMLS (Unified Medical Language System) and regular expressions. It added a parser instead, which uses a Look-Ahead Left-Recursive (LALR) grammar to identify the negations. It was used to analyse discharge summaries and surgical notes and achieved 95.3% sensitivity and 97.7% specificity. The negation detection algorithm that was published recently used hybrid approach that is based on grammatical parsing and regular expression [17]. Initially, using the regular expressions, negation terms were detected for achieving high sensitivity, and then the pos (part of speech) parse tree was traversed to identify negated phrases with greater specificity.

In [18], the authors have developed the temporal analysis concept in the CLEF context (Clinical eScience Framework) information extraction component. Patient chronicle was built from the extracted information which was an overview of the important events in the medical history of the patient. Some researchers used techniques in machine learning for automatic temporal segmentation and ordering. They use syntactic, topical, lexical and positional features for temporal segmentation. Integer Linear Programming framework discussed in [19] served to obtain best results for ordering of the segments.

In [20], the authors proposed an algorithm for the identification of contextual features. The algorithm named ConText determines three contextual feature values: Negation [affirmed, negated], Temporality [hypothetical, recent, historical], and Experiencer [patient or other]. Regular expressions are used in the algorithm to detect scope termination terms (important for context analysis), trigger and pseudo-trigger terms, and then maps the contexts detected to concepts between trigger terms and end of the sentence or the scope termination term. Separate algorithms specialized in analysing the contextual features are easier for the implementation. Major part of these algorithms is based on lexical information though some algorithms even add the part-of-speech information also.

6. OPPORTUNITIES IN CLINICAL TEXT MINING BASED ON HIDDEN MARKOV MODEL

When the information is being extracted from the clinical text, at any point of time with any level of seriousness, we have only partial information about any non-trivial situation. So we need to work with levels/ layers. The models which are used for uncertainty processing are fuzzy logic, probability, information theory and non-monotonic logic. All these are important models each having its own characteristics.

The kinds of uncertainties can be categorized into the following:

a) Part-of-speech ambiguity which can be resolved with surface analysis and deeper analysis with respect to semantics

b) Sense ambiguity which can be resolved with clue words

c) Lexical loss ambiguity

d) Scope ambiguity, and

e) Co-referencing or anaphora ambiguity.

Many artificial intelligence tasks are sequence labelling tasks. For this, probabilistic framework and Markov process are the best choice to process textual information to perform NLP.

The Markov assumption states that the probability of a state being the state of the machine depends only on the previous state. This is Order-1 Markov assumption. The best possible tag sequence, i.e., the highest probability path from head symbol to the dot symbol of a tagged sentence can be found by the *argmax* computation:

Best tag sequence = T^*

T^* = argmax P (T|W)

= argmaxP(T).P(W|T) (by Baye’s Theorem)

where T is the tag sequence

W is the word sequence

The prior probability P(T) helps us as a filter to eliminate bad possibilities and is a representation for highly likely tag sequences as learned from the corpora.

By applying Baye’s Theorem, Chain Rule, and Markovian assumption, we get the expression as

P(T) = $\prod_{i=1}^{N+1} P(t_i|t_{i-1})$

This is Bigram Assumption which states to disregard anything which is very distant from the current tag.

This assumption will be quiet useful in clinical text mining as the disease onset, severity etc. can be expressed with mostly two words like {high, confusion}, {severe, memory loss} etc. Some string structures for which there is further ambiguity can be resolved by furthering the Markov assumption to trigrams where the scope is increased and the context analysis can be intensified.

7. APPLICATION OF VITERBI ALGORITHM IN MACHINE LEARNING FOR CLINICAL TEXT

The key element of the Viterbi Algorithm is Markov assumption. Viterbi algorithm is used to find most likely sequence of states (tags) which is called as the Viterbi Path which results into a sequence of the observed events.

Consider the patient conditions from the clinical text be either mild or acute degree of the memory loss. Based on the later diagnosis made over time, the recorded health condition of the patient be any of the three states – normal, confusion, and memory loss. There are two states “mild” and “acute”, but the machine can’t observe these directly as they are hidden from it. On each visit, there is a certain chance that the patient will tell the clinician that he/she is “normal”, “confusion”, or “memory loss”, depending on his/her health condition.

The observations (normal, confusion, memory loss) along with hidden states (mild, acute) form a Markov model and can be graphically represented as shown in Figure 1.

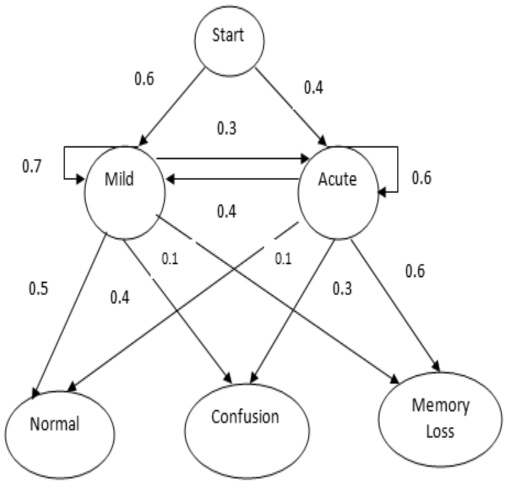

Fig 1: Graphical representation of the observations and hidden states

The given transition probabilities represent the change in the health condition in the underlying Markov Chain. From the given data, there is only 30% chance for the patient to be diagnosed with acute health condition in next scheduled visit if his illness is mild in the initial stage. The emission probabilities represent how likely the patient is to feel on each next visit. If his health condition is mild, there is a 50% chance that he feels normal; if his health condition is acute, there is a 60% chance that he encounters memory loss.

The patient visits the clinician in three episodes and the clinician discovers that on the first visit, the patient feels normal state of the health condition, on the second visit, he feels that he encounters a confused state, and on the third visit, he encounters complete memory loss. In this scenario, the most likely sequence of health conditions of the patient that would explain these observations can be answered by Viterbi algorithm.

7.1 The Viterbi Algorithm

Consider that we are given a HMM with a state space S, initial probabilities as π_i (being in state i) and transition probabilities $a_{i,j}$ (transitioning from state i to j). Say, we observe the outputs y_1,----,y_Q. The likely state sequence x_1,----,x_Q that produces the observations is given by the recurrence relations as:

$V_{1,r} \;= P(y_1|r).\pi_r$-------(1)

$V_{q,r} = \;max_{x\in S}\;(P(y_q|r).a_{x,r}.V_{q-1,x}$-------(2)

Here $V_{q,r}$ is the probability of the most likely state sequence P(x_1,----x_Q, y_1,----y_Q) responsible for the first *q* observations that have *r* as its final state. By saving the back pointers which remember the state *x* that was used in the eq.(2), Viterbi Path can be retrieved.

Let *f(r,q)* be the function that returns the value of *x* used to compute $V_{q,r}$ if *q*>1, or *r* if *q*=1.

Then $\quad x_Q = argmax_{x\in S}(V_{Q,x})$

$\quad\quad x_{q-1} = f(x_q, Q)$

The computed complexity of the algorithm is

O (Q X $|S|^2$)

Applying the Viterbi Algorithm to the given problem, obtaining the most likely sequence of health conditions of the patient can be depicted with the following steps and figures:

Step 1: Calculate P_Start(State).P_Obs(“Normal”) as shown in Figure 2.

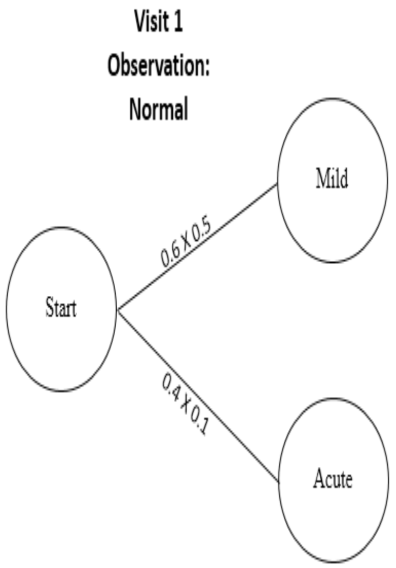

Fig 2: Graphical representation of Step 1

Step 2: Calculate P(OldState).

P_trans(OldState->NewState).P(“Confusion”|NewState) as shown in Figure 3.

Eg. - P(Mild).P(Mild->Mild).P(“Confusion”|Mild)

= 0.3 x 0.7 x 0.4 = 0.084

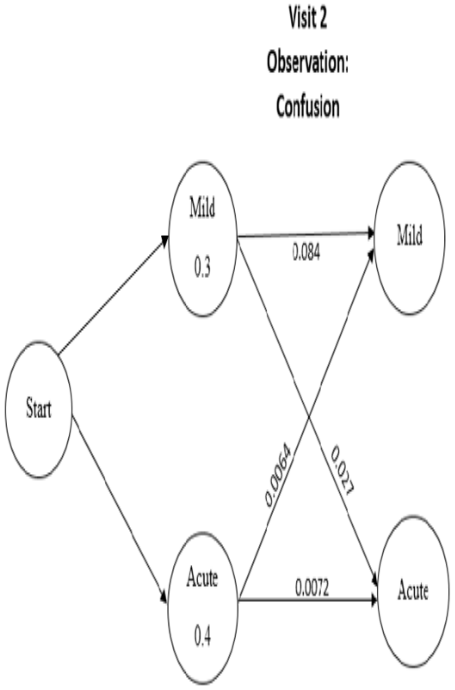

Fig 3: Graphical representation of Step 2

Step 3: For each state transition, select the path with highest probability as shown in Figure 4.

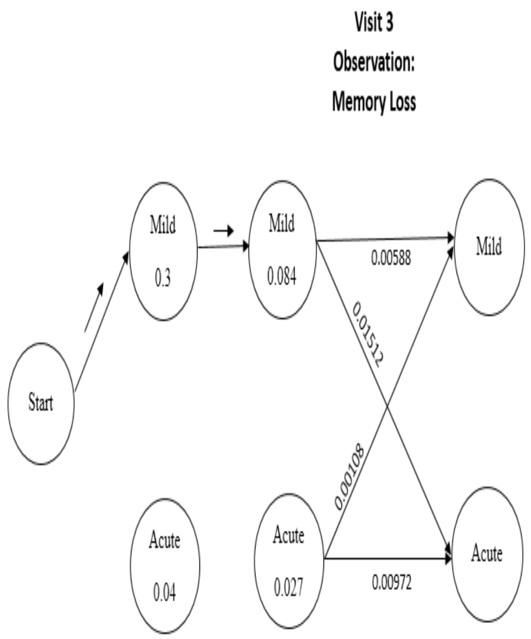

Fig 4: Graphical representation of Step 3

Step 4: Repeat steps 2 and 3 for each observation to complete, as shown in Figure 5.

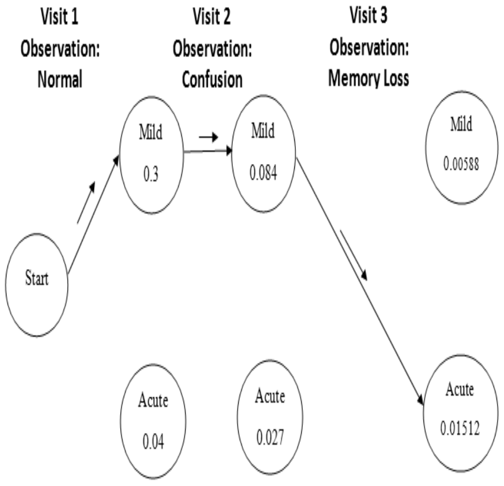

Fig 5: Graphical representation of Step 4

By applying the Viterbi Algorithm, the most likely sequence of health conditions of the patient that would explain the observations are {'Mild','Mild','Acute'}

This discussion indicates the potential for Viterbi Algorithm in information extraction from clinical text, in medical corpus building, and in machine learning in healthcare domain.

8. CONCLUSION

This paper discusses about the information extraction from clinical text, and the issues and opportunities with respect to this task using NLP and Machine Learning. A brief review of the work has been provided that was already done in this domain and issues related to them like limited access to shareable clinical text and importance of contextual analysis to draw conclusions from the clinical text. The usage of Viterbi algorithm has been proposed for clinical text mining in machine learning. A scenario has been discussed to exemplify the usage and invoke further research in this area. This would have a good scope in future as the Viterbi algorithm may play a crucial role in medical corpus building and context dependent information extraction from clinical text that would help the clinicians take better and quicker decisions. This would also help in identifying most probable disease-treatment-side effect associations that would increase the quality of treatment.

9. REFERENCES

[1] Peter B. Jensen, Lars J. Jensen &SorenBrunak, "Mining EHRs towards better research applications and clinical care", Nature Reviews Genetics, June 2012.

[2] Shortliffe EH, 1987. Computer programs to support clinical decision making.

[3] DunaDemner-Fushman, Wendy W. Chapman, Clement J. McDonald, "What can NLP do for Clinical Decision Support?", Journal of Biomedical Informatics, August 2009, Elsevier Inc.

[4] Sager N, Chi E, Friedman C, "The analysis and processing of clinical narrative", Medinfo;1986, Elsevier.

[5] Haug PJ, Ranum DL, Frederick PR, "Computerized extraction of coded findings from from free-text radio-logic reports", Radiology, February 1990.

[6] Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM, "Experience with a mixed semantic/ syntactic parser", Proceedings of Annual Symposium of Computational Appl. Med Care, 1995.

[7] Hobbs JR, "Information extraction from biomedical text", Journal of Biomed Information, August 2002.

[8] Pakhomov S, Buntrock J, Duffy PH, "High throughput modularized NLP system for clinical text", 43rd Annual Meeting of the Association for Computational Linguistics, 2005.

[9] Liu K, Mitchell KJ, Chapman WW, Crowley RS, "Automating tissue bank annotation from pathology reports – comparison to a gold standard expert annotation set", AMIA Annual Symposium Proceedings, 2005.

[10] Dorr DA, Phillips WF, Phansalkar S, Sims S A, Hurdle JF, "Assessing the difficulty and time cost of de-identification in clinical narratives", Methods Inf Med, 2006.

[11] Sweeney L, "Replacing personally-identifying information in medical records, the Scrub system, Proceedings of AMIA Annual Fall Symposium 1996.

[12] Ruch P, Baud RH, Rassinoux AM, BouillonnP, Robert G, "Medical document anonymization with a semantic lexicon", Proceedings of AMIA Symposium, 2000.

[13] Beckwith BA, Mahaadevan R, Balis UJ, Kuo F, "Development and evaluation of an open source software tool for de-identification of pathology reports", BMC Medical Informatics & Decision Making, 2006.

[14] Uzuner O, Luo Y, Szolovits P, "Evaluating the state-of-the-art in automatic de-identification", JAMIA 2007.

[15] Aronow DB, Fangfang F, Croft WB, "Ad hoc classification of radiology reports", JAMIA 1999.

[16] Mutalik PG, Deshpande A, Nadkarni PM, "Use of general purpose negation detection to augment concept indexing of medical documents: a quantitative study using UMLS, JAMIA 2001.

[17] Huang Y, Lowe HJ, "A novel hybrid approach to automated negation detection in clinical radiology reports", JAMIA 2007.

[18] Harkema H, Setzer A, Gaizauskas R, Hepple M, "Mining and modelling temporal clinical data", Proceedings of the UK e-Science All Hands Meeting 2005.

[19] Bramsen P, Deshpande P, Lee YK, Barzilay R, "Finding temporal order in discharge summaries", Proceedings of AMIA Annual Symposium 2006.

[20] Chapman W, Chu D, Dowing JN, "ConText: An algorithm for identifying contextual features from clinical text", BioNLP 2007: Biological, translational, and clinical language processing, Prague, CZ.