Comparison of Current Algorithms for Extracting information in Electronic Medical Record

Yiwen Meng
PhD in Bioengineering Department / UCLA
ywmeng369@gmail.com

Abstract

Electronic Medical Record (EMR) has becoming a wide spread way for physicians to keep track of patients' health record.However, there is not an efficient way to extract useful information for clinics or diagnosis because of the lengthy and narrative nature of EMR.With the development of Natural Language Processing (NLP), several tools have been invented to identify and extract information and negation terms from EMR.This paper summarizes and compares the performance of the recent NLP tools as HITEx,MetaMap, MedLEE and re-implement the negation algorithm in NegEx.

1 Introduction

Clinical records contain much potentially useful information in free text form. In electronic medical records (EMR), information such as family history, signs and symptoms and personal history of drinking and smoking are typically embedded in text descriptions provided by clinicians in the form of progress notes and in more formalized discharge summaries. Even when coded data are available (e.g.billing codes for principal diagnoses and co-morbidities), they may not always be accurately assigned or widely utilized and may be subtly influenced by financial incentives.

Natural Language Processing (NLP) is a successful application of machine learning algorithms to resolve real-life problems.Apart from the traditional aspect of linguistics and training the machine to understand different corpus. There is another promising application for NLP to facilitate classifying medical information and critical terms like negation in Electronic Medical Record (EMR) which is also called Electronic Health Record (EHR). After adequate training and learning,the machine can even predict the health condition for patients. Like other corpus which has a dictionary to explain every composed word and phrases.The one in medical field is called the Unified Medical Language System (UMLS), which is designed and maintained by the US National Library of Medicine. Many terminologies in EMR or EHR are from the database of UMLS, which constructs a huge number in clinics. However, due the narrative nature of many EMR or EHR written by physicians,the meaning of the context is not accurate, concise and timely.Therefore,many researchers and medical doctors put effort to investigate classifier or system to make EMR or HER as an efficient tool for record of disease progression and detection.

In this paper, comparison are made bewteen several recent NLP tools to extract information from EMR for medical and clinics like HITEx,MetaMap,MedLEE. In addition, re-implementing one negation algorithm called NegEx is presented here in order to discuss the current status and challenge. One challenge with many previously described medical NLP tools is that they are not easy to adapt, generalize and reuse.

2 UMLS and NLP Tools

The Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical science, which provides a mapping structure among these vocabularies and thus allow one to translate among various terminology systems. It is designed and maintained by the US National Library of Medicine.The purpose of the National Library of Medicine Unified Medical Language System (UMLS) is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. The UMLS provides data for system developers as well as search and

report functions for less technical users. There are three UMLS Knowledge Sources:

- The Metathesaurus, which contains over one million biomedical concepts from over 100 source vocabularies
- The Semantic Network, which defines 133 broad categories and fifty-four relationships between categories for labeling the biomedical domain
- The SPECIALIST Lexicon and Lexical Tools, which provide lexical information and programs for language processing

2.1 Health Information Text Extraction(HITEx)

HITEx uses GATE (General Architecture for Text Engineering) as the development platform. GATE is an open-source natural language processing framework which includes a set of NLP modules, collectively known as CREOLE:a Collection of Reusable Objects for Language Engineering.CREOLE contains NLP modules that perform some common tasks, such as tokenizing, part-of speech (POS) tagging, and noun phrases parsing.The GATE framework can be viewed as a backplane for plugging in CREOLE components. The framework provides various services to the components, such as component discovery, bootstrapping, loading and reloading, management and visualization of data structures, and data storage and process execution.HITEx uses 11 GATE modules (components),the most critical one are listed here:

- The POS tagger: assigns part-of-speech tags to each word (token) in the sentence.This module is based on the Brill-style, rule based POS tagger,as a plug-in for the GATE framework
- UMLS concept mapper: maps the strings of text to UMLS concepts.The module first attempts exact match; when exact matches are not found, it stems, normalizes and truncates the string
- Negation finder: assigns the negation modifier to the existing UMLS concepts. Currently, this module is an implementation of NexEx-2 negation algorithm.
- Regular expression-based concept finder: finds all occurrences of the concepts defined as a regular expression in the input chunk of text.

2.2 MetaMap

MetaMap is a widely available program providing access to the concepts in the (UMLS) Metathesaurus from biomedical text.

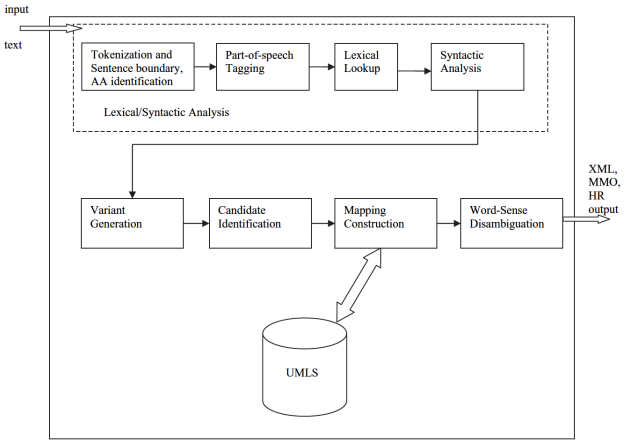

Figure 1: Structure diagram of MetaMap

As shown in Fig.1, the structure of MetaMap is very similar to HITEx except there is no negation algorithm in MetaMap while HITEx lacks word sense disambiguation (WSD),in which mappings involving concepts that are semantically consistent with surrounding text are favored.MetaMap possesses a number of strengths and weaknesses. Among its strengths are its thoroughness, characterized by its aggressive generation of word variants, and its linguistically principled approach to its lexical and syntactic analyses as well as its evaluation metric for scoring and ranking concepts. It is also adept at constructing partial, compound mappings when a single concept is insufficient to characterize input text phrases.As evidenced by the above description of some of its options,it is highly configurable; its behavior can be easily customized depending on the task being addressed. Finally, because its lexicon and target vocabulary can be replaced with others from another domain, it has the property of domain independence. One of MetaMap's weaknesses is that it can be applied only to English text.MetaMap's English-centric nature is evident throughout its implementation, not just in its lexical and syntactic algorithms. Also, a negative consequence of its thoroughness is that it is relatively slow. In its current

implementation, it is not appropriate for real-time use; and it would require a major fine-grained parallel re-implementation in order to overcome this weakness.

2.3 MedLEE

The MedLEE (Medical Language Extraction and Encoding) system has been deployed to extract and encode information in clinical narratives for a large number of different applications and studies.For a given report, MedLEE generates a set of structured findings,such as problem (headache),or medication (ibuprofen),along with associated modifiers,such as certainty (no, high certainty), status (previous, recent),body location (chest), and section (Hospital Course). The output of MedLEE is consistent with frames, and has the format. Type-Value-Modifiers, where Type is the type of information in the frame, Value is the value and Modifiers are a sequence of frames containing the same format where each modifier frame denotes a certain type of qualifying information. Therefore, the result is more meaningful than previous tools by linking several medical information together.

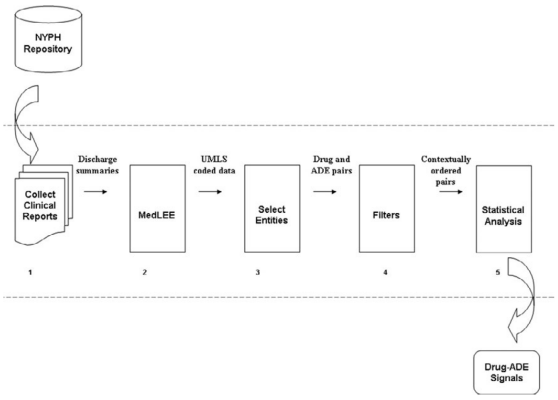

Figure 2: Structure diagram of MedLEE

3 NegEx:The Negation Algorithm

In clinical reports, the presence of a term does not necessarily indicate the presence of the clinical condition represented by that term. In fact, many of the most frequently described findings and diseases in discharge summaries, radiology reports, history and physical exams, and other transcribed reports are denied in the patient. Physicians often note that a particular disease can be ruled out or that a finding consistent with a suspected disease is absent.We use the term“pertinent negatives” to refer to findings and diseases explicitly or implicitly described as absent in a patient.Differentiating pertinent negatives from positive conditions in a clinical report is crucial to accurate indexing of the report. Therefore, it was first implemented as a computationally simple algorithm that could be quickly and easily applied to any medical concept indexing or feature extraction system.

The input to NegEx is a sentence with indexed findings and diseases. The output is whether an indexed phrase is negated in the sentence.It first identifies a UMLS terms or phrases by a simple string matching algorithm.After that, search for negative phrases in the sentence to match the 35 pre-defined ones in NegEx, which are divided into two groups.The first group is called “pseudo-negation” phrases, consisting of phrases that appear to indicate negation but instead identify double negatives (“not ruled out”), modified meanings (“gram-negative”), and ambiguous phrasing (“unremarkable”).The second group consists of phrases we believed are used to deny findings and diseases when used in one of two regular expressions.In the first regular expression, the negation phrase precedes the UMLS term.In the second exepression,the negation phrase follows the UMLS term.

4 Results and Discussion

All of the information extraction tools have been applied to different types of patient discharge summaries to investigate their sensitivity and accuracy.

HITEx has been tested on patient with history of asthma and Chronic Obstructive Pulmonary Disease (OCPD), to extract principal diagnoses, comorbidities and smoking status information for 150 reports.Compared to the results from expert physicians,it achieved the accuracy of HITEx for principal diagnosis extraction was 82% and for comorbidities was 87%. The sensitivity and specificity of HITEx were 77% and 87% for principal diagnosis and 70% and 89% for comorbidity extraction.The accuracy of smoking status extraction was 90% and the sensitivities and specificities range from 60% to 100% and 93 to 99% respectively.

MetaMap is more versatile for input, so the current experiment was a task for large-scale indexing of online biomedical resources on four document sets ranging in size from 2 K documents to 99 K documents and for two entity types,

biological processes and diseases.Overall, the precision and recall is 85% and 82%,respectively. Fig. 4 shows an example output by the input of sentence “obstructive sleep apnea”. The output first displays the confident score of mapping the phrases to UMLS Metathesaurus. In this example, the sentence gets the perfect score of 1000.In addition, showing the a list of intermediate results in the descend order by score consisting of Metathesaurus strings matching some or all of the input text.Moreover, the preferred name of each candidate is displayed in parentheses if it differs from the candidate, and the semantic type of the candidate is also shown.Finally, the mapping combinations of candidates matching as much of the phrase as possible. There is an web version of MetaMap in this link:`https://ii.nlm.nih.gov/Interactive/UTS_Required/metamap.shtml`, but it requires to register an account approved UMLS.

```
Phrase: "obstructive sleep apnea"
Meta Candidates (11):
  1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or
       Syndrome]
   901 Apnea, Sleep (Sleep Apnea Syndromes) [Disease or Syndrome]
   827 APNOEA (Apnea) [Pathologic Function]
   827 Sleep [Organism Function]
   827 Obstructive (Obstructed) [Functional Concept]
   827 Apnea (Apnea Adverse Event) [Finding]
   793 Sleeping (Asleep) [Finding]
   755 Obstruction [Individual Behavior,Pathologic Function]
   755 Sleepy [Finding]
   755 Sleeplessness [Sign or Symptom]
   755 Obstruction (Obstruction within Medical Device) [Phenomenon or
       Process]
Meta Mapping (1000):
  1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or
       Syndrome]
```

Figure 3: An example output of MetaMap

Finally, NegEx was tested on 1000 sentences contained 1235 occurrences of UMLS terms, 245 of which were unique strings.The gold standard is by the reliability of physicians judgments of the UMLS terms. In total, the sensitivity and specificity are 77.84% and 94.51% respectively, while the positive predictive value(PPV) and negative predictive value(NPV) are 84.49% and 91.73%, respectively. Fig. 4 shows the 15 negation phrases actually identified by NegEx in the test set, along with their respective PPVs Of the 15 negation phrases found in the test set, three (no, without, and no evidence of) accounted for 82% of the correctly identified pertinent negatives. Three of the 15 phrases had very poor PPV: “versus”(0%), not (58%), and doubt (50%) and are therefore candidates for further examination.

Negation phrase	PPV (%)	Number of times identified by NegEx in test set
no signs of	100.00	2
ruled out	100.00	3
unlikely	100.00	2
absence of	100.00	1
not demonstrated	100.00	1
denies	100.00	7
no sign of	100.00	2
no evidence of	94.59	37
no	92.36	157
denied	90.00	10
without	88.10	42
negative for	80.00	10
not	58.33	24
doubt	50.00	2
versus	0.00	16

Figure 4: Positive Predictive Value (PPV) of NegEx Negation Phrases Found in the Test Set

The re-implementation of NegEx is done in python notebook in which there are 20 sentences from patient discharge summary. The overall accuracy is 97.22%.The code and the data file is through this link:`https://github.com/lanyexiaosa/NegEx-demo`