Topic Segmentation and Medical Named Entities Recognition for Pictorially Visualizing Health Record Summary System

Wei Ruan

A thesis submitted in partial fulfillment of the requirements for the
Master of Applied Science in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

© Wei Ruan, Ottawa, Canada, 2019

Abstract

Medical Information Visualization makes optimized use of digitized data of medical records, e.g. Electronic Medical Record. This thesis is an extended work of Pictorial Information Visualization System (PIVS) developed by Yongji Jin (Jin, 2016) Jiaren Suo (Suo, 2017) which is a graphical visualization system by picturizing patient's medical history summary depicting patients' medical information in order to help patients and doctors to easily capture patients' past and present conditions. The summary information has been manually entered into the interface where the information can be taken from clinical notes.

This study proposes a methodology of automatically extracting medical information from patients' clinical notes by using the techniques of Natural Language Processing in order to produce medical history summarization from past medical records. We develop a Named Entities Recognition system to extract the information of the medical imaging procedure (performance date, human body location, imaging results and so on) and medications (medication names, frequency and quantities) by applying the model of conditional random fields with three main features and others: word-based, part-of-speech, Metamap semantic features. Adding Metamap semantic features is a novel idea which raised the accuracy compared to previous studies. Our evaluation shows that our model has higher accuracy than others on medication extraction as a case study.

For enhancing the accuracy of entities extraction, we also propose a methodology of Topic Segmentation to clinical notes using boundary detection by determining the difference of classification probabilities of subsequence sequences, which is different from the traditional Topic Segmentation approaches such as TextTiling, TopicTiling and Beeferman Statistical Model. With Topic Segmentation combined for Named Entities

Extraction, we observed higher accuracy for medication extraction compared to the case without the segmentation.

Finally, we also present a prototype of integrating our information extraction system with PIVS by simply building the database of interface coordinates and the terms of human body parts.

Acknowledge

First of all, I would like to express my gratitude and appreciation to my supervisor, Prof. WonSook Lee, for her excellent guidance in all steps of my research preparation. She provided me with many insightful suggestions and offered me with many precious academic opportunities. Without her guidance and persistent help, I would not have all the successes of my research.

The memory in CG++ group is always unforgotten. I am grateful to my lab mates: Naveenkumar Appasani, Shuangyue Wen, Nan Wang, Kang Wang and Ph.D. Candidate Mr. Hamed Mozaffari, for their unfailing support and assistance in my graduate studies. In addition, I would also like to give warm thanks to Gingmi Lee, Soonmo Kwon and Andy Wang, my best friends in Canada, for their love, encouragement and constant help.

Thank you to the entire department at the University of Ottawa; all of the faculty and staff who I met throughout my program were wonderful people. You have given me more or less help in my postgraduate career.

Finally, I would like to thank my parents and grandmother, the persons who love me the most in the world. It was them who let me come to this colourful world and spared no effort to support me grow up. Many of my successes today are indispensable to my parents' support. I love them.

Table of Contents

Abbreviations

EMR	Electronic Medical Record
HER	Health Electronic Record
PIVS	Pictorial Information Visualization System
NLP	Natural Language Processing
LDA	Latent Dirichlet Allocation
InfoVis	Information Visualization
NER	Named Entity Recognition
HMM	Hidden Markov Model
MEMM	Maximum-entropy Markov Model
CRF	Conditional Random Fields
TS	Topic Segmentation/ Text Segmentation
2D	2-Dimensional
3D	3-Dimensional
CT	Computed Tomography scan
MRI	Magnetic Resonance Imaging
LSA	Latent Semantic Analysis
WSD	word sense disambiguation
UMLS	Unified Medical Language System

UIMA	Unstructured Information Management Architecture
NB	Naïve Bayes
SVM	Support Vector Machine

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Motivation

This study is an extending work of Pictorial Visualization System with Patient Portal for Problem-based Electronic Medical Record (EMR), a master thesis work, proposed by Suo (Suo, 2017) which followed Jin's thesis work (Jin, 2016) in University of Ottawa. They proposed a web-based pictorial visualization system to which patients and doctors both have access. The system they proposed could allow spatial interactivity and temporal interactivity by representing human body images and interconnecting time axes respectively. The functions of visualization and interaction between the doctors and patients enable physicians to rapidly know about patients' health conditions and accordingly make medical decisions, which greatly improve the efficiency of the process of diagnosis. The idea of Jiaren's system could to some extent avoid some shortcomings of current EMR systems. EMR, also known as EHR, the abbreviation of Electronic Medical Record and Electronic Health Record respectively, is a computer-based patient record specific to a single clinical practice, such as family health team or group practice. Figure 1 shows a workflow in an integrated health information (Suo, 2017).

Current EMR systems as digital information sources do not fully realize their potential, where medical information can be structured in order to maximize comfortable extraction of target information. EMR systems should enable doctors to quickly integrate or recuperate different types of medical information, such as symptoms, diagnosis, laboratory tests, treatments and medicines. Visualization of information (InfoVis) is one

method that can maximize the value of electronically available medical data (Chittaro, 2001). Because InfoVis offers medical information in intuitive, understandable, recognizable, navigable and manageable formats, users can quickly understand and extract useful information from large quantities of documents.

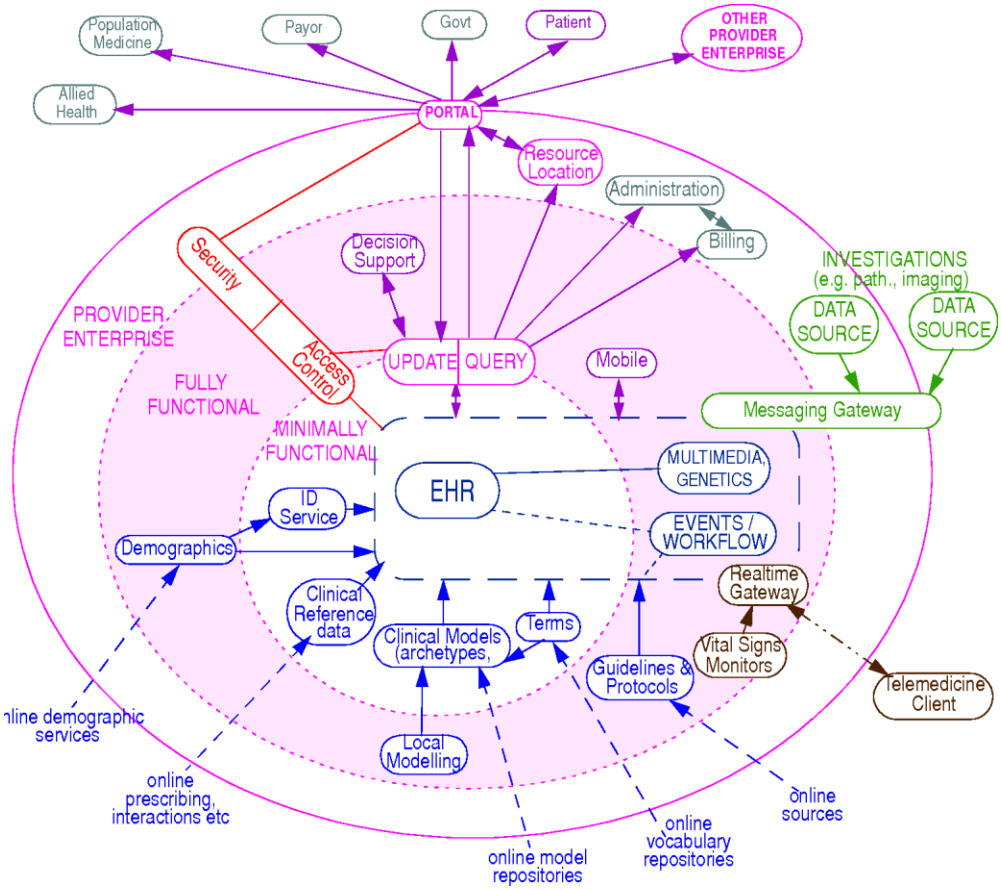

Figure 1: A health information environment (Suo, 2017). This figure should be read from inside out, starting at the level of EHR environment, from which some basic information on patient health and identification, drug data and e-prescriptions, among other information are available.

Commercial popular EMR systems like EpicCare [1], Allscripts [2], and eClinicalWorks [3], offer problem lists, electronic prescriptions, allergy inspections, medicines, order management and billing functions. Consequently, these systems contain

[1] Epic: https://www.epic.com/software#PatientEngagement
[2] Allscripts: https://ca.allscripts.com/
[3] eClinicalWorks: https://www.eclinicalworks.com/

different texts including documents, orders and notes. A frequent mode of visualization is the Flowsheet, which is widely used in ICUs in commercial EMR systems. For a patient, it contains key medical variables over a given time and therefore emphasizes trends and abnormal values. Many of these systems include automated analysis techniques based on machine training and information mining, such as IBM Watson Health and Oracle Health Sciences. Substantial medical information is classified in ways that doctors do not meet the requirements for a fast overview of the medical history of the patient. Current EMR systems make it easy to incorporate and collect medical data quickly but lack the necessary functions for temporary queries and exploratory analysis tasks.

The pictorial visualization system proposed by Suo (Suo, 2017) and Jin (Jin, 2016) could solve the above current EMR problem if lacking the necessarily functionality. Firstly, the medical information is spatially structured in the system. Secondly, 11 individual physiological systems are built for classifying different disease and illnesses. Finally, the system could visualize historical records on a horizontal time axis. This study is based on these three features of the Pictorial Visualization System.

1.2. Objectives

The core goal of this study is to help the Pictorial Medical Information Visualization System extract medical data from plain texts. It should be clear which medical data would be extracted. Diseases and their locations should be extracted for spatial visualization on the six physiological systems: Musculoskeletal/Dermatological, Cardiovascular/Respiratory, Gastrointestinal, Nervous, Immune/Endocrine and Genitourinary. Medications and medical imaging procedures are also crucial for temporal visualization. Therefore, diseases and their body part locations, imaging procedures, medications and dates would be the information to be extracted for visualization. We

applied the techniques of Topic Segmentation and Named Entities Extraction for information retrieval.

The final goal of this thesis study is to:

- Visualizing the desired medical information in clinical notes on the Pictorial Information Visualization System.

For realizing the goal presented above, there are some sub-goals as follows:

- Extracting the desired medical named entities, such as medications and medical imaging procedures, from clinical notes.
- For enhancing the accuracy of entities extraction, the other objective is to apply Topic Segmentation to clinical notes for separating and categorizing the text into short parts each of which is one of the five topics: Medications, History, Physical Examinations, Laboratories and Hospital Course.

Extracted information are finally visualized on the interface spatially and temporally.

1.3. Contributions

This study is an extending work of previous Pictorial Medical Information Visualization System. The goal of this study is to realize desired medical information extraction from clinical notes or medical documents in text-plain-version. What we have achieved through this thesis work is as follows:

- We developed methods of Topic Segmentation experimented on the following topics:
  - ✓ History: describing patient's past illnesses, surgeries and other medical information;
  - ✓ Medications: illustrating clinical drugs that patients need to take.
  - ✓ Physical Examinations: a routine test to check your overall health, such as head, eyes and ears.

- ✓ Laboratories: a routine test on same tissues or substances taken from patients' body, such as blood and urine, to help diagnose disease or other conditions;
- ✓ Hospital Course: containing information about of the sequence of events from admission to discharge in a hospital facility.

- We implemented a system of named entity recognition for extracting desired information from clinical notes in English language. The entities we experimented with are as follows:
  - ✓ Medication - names, medication quantity, drug administration, frequency, clinical note date
  - ✓ Medical imaging – related human body part, performance date, medical image results, medical imaging procedure
- We experimented the effect of Topic Segmentation for the named entity extraction on medication names.
- We evaluated Topic Segmentation model using Windiff, P_k , Boundary and Segmentation similarity while $\boldsymbol{F_1}$ score and SegEval were employed for evaluation of Named Entities Extraction.
- We presented the idea of integrating our information extraction framework proposed in this study with the pictorial information visualization system developed by Suo (Suo, 2017) and Jin (Jin, 2016)

1.3.1. Contributions

The research contributions are as follows:

- We introduced Naïve Bayes and Support Vector Machine models to perform Topic Segmentations.
- In the Topic Segmentation a new method to detect boundary has been proposed. It is to use score difference between different topics where the scores are accumulated

probability scores for each topic. Taking the difference between different topic probability scores shows boundary clearly.

- In the Topic Segmentation the process of score assignment and segmentation are done simultaneously while these two steps are generally separated in other algorithms. Consequently, our algorithm could potentially be more efficient when dealing with a large size of dataset.
- In the Named Entities Extraction, we applied Metamap, a UMLS concepts recognizing tool, for tagging medical named entities as the semantic types and semantic groups for enhancing the accuracy of medical named entities extraction. Metamap offers the identification of medical terminologies using "semantic types" and "semantic groups". We utilize this feature to enrich the features for CRF model training for improving the extraction performance, especially, medication entities extraction comparing with (Yefeng Wang, 2009), (Kumar, Alam, Kumar, & Sheel, 2014) and (Patrick & Li, 2010).

1.4. Thesis Structure

The remainder of the thesis is organized as follow:

- Chapter 2 serves good understandings of related work of our research. We mainly discuss some famous Text and Topic Segmentation models: TextTiling, TopicTiling and Beeferman model. Meanwhile, several current researches on text or Topic Segmentation applied for clinical notes are introduced in detail. In addition, the technique of Named Entity Recognition (NER) will be introduced. Three well-known NER models: Hidden Markov Model (HMM), Maximum-entropy Markov Model (MEMM) and Conditional Random Fields (CRF), will also be presented. In the end, we would discuss Medical Information Visualization in which InfoVis technique and examples, novel proposed InfoVis system and current research on medical InfoVis are introduced.

- Chapter 3 provides details of our Topic Segmentation algorithm. This chapter first offers an overview of the segmentation methodology by illustrating a diagram. In the next part, Topic Score Predictor will be primarily demonstrated with dataset, feature and model selection detailed out. Finally and most importantly, the steps of boundaries detection would be mentioned in this section, which is the core technique of our Topic Segmentation.
- Chapter 4 presents the methodology of information extraction using Named Entities Extraction based on Conditional Random Fields model. In this chapter, we mainly demonstrate four features (word-based feature, POS feature, semantic feature and other feature) for Conditional Random Fields model training.
- Chapter 5 gives the idea of how we visualize extracted information into the Pictorial Information Visualization System spatially. A database of the pixels and medical names would be illustrated in this section.
- Chapter 6 serves the most important part of this research- Evaluation and Results. In this chapter, several evaluation metrics for Text Segmentation and Information Retrieval are introduced. We then show all the results of segmentation and entities extraction. We also discuss and analyse possible errors.
- Chapter 7 concludes this thesis study. It summarizes the components of our research and systems and briefly presents potential future work.

Chapter 2. Literature Review

This Chapter will serve good understandings of related work of our research. We firstly discuss about Natural Language Processing for clinical notes with free-text version. For helping better understand the importance and role of NLP in clinical notes, we detailed out some researches related to applications of NLP in biomedical informatics. Secondly, we discuss about text and Topic Segmentation which plays an important role in our study. In this section, some famous text segmentation models: TextTiling, TopicTiling, Beeferman model and et al, and several current research on text or Topic Segmentation applied for clinical notes, are introduced in detail. In addition, this chapter also provides an introduction of Named Entity Recognition (NER), a subtask of Information Extraction in the process of Natural Language Processing. Three well-known NER models: Hidden Markov Model (HMM), Maximum-entropy Markov Model (MEMM) and Conditional Random Fields (CRF), are presented. Besides, we compared some current research on Medical Entity Extraction for offering a better understanding of the techniques applied in this research. Finally, Medical Information Visualization is discussed, in which InfoVis technique and examples, novel proposed InfoVis system and current research on medical InfoVis are introduced.

2.1. Natural Language Processing for Clinical Notes

Clinical notes contain crucial information about patients. There are many researchers who have worked on applying the techniques of Natural Language Processing to clinical notes for achieving desired tasks. Byrd et al. (Byrd, Steinhubl, Sun, Ebadollahi, & Stewart, 2014) used NLP techniques to identify signs and symptoms of Framingham HF among patients with primary care. Pakhomov et al. (Pakhomov, Buntrock, & Duffy, 2005) presented a text analysis and information

retrieval system, which is high throughput, real-time modularized, to identify clinically relevant entities in clinical note. Meanwhile, the entities are mapped to several standardized nomenclatures and available for subsequent information retrieval and data mining. Savova et al. (Savova et al., 2008) employed a supervised machine learning technique for exploring the word sense disambiguation (WSD) problem across two biomedical domains (biomedical literature and clinical notes. Table 1 illustrates more details of their research. In 2012, a system which could characterize empirical instances of Unified Medical Language System (UMLS) Metathesaurus term strings in a large clinical corpus was proposed by Wu et al. (Wu et al., 2012) for illustrating what types of term characteristics are generalizable across data sources. Meanwhile, LePendu et al. (LePendu, Iyer, Fairon, & Shah, 2012) described the application of simple clinical text annotation tools and the mining of the resulting annotations to calculate the patients ' risk of developing a myocardial infarction.

2.1.1. Unified Medical Language System

As we can see from above research, it is clear to see that UMLS[4] is well-known and widely used in the field of biomedical informatics. The UMLS is a collection of many controlled biomedical vocabularies (created in 1986). The Metathesaurus, a very wide, multifunctional and multiple language vocabulary database, contains information on the concepts related to biomedicine and health, its various names and the relations between them. It transcends the thesauri, vocabulary and classifications it contains. The Metathesaurus is organized according to concept or meaning. It connects alternative names and perspectives with the same concept and identifies useful relations between different concepts. The Metathesaurus is also connected to the other UMLS sources of knowledge, the Semantic Network and the SPECIALIST Lexicon. The Semantic Network assigns all Metathesaurus concepts to at least one Semantic type. This ensures

[4] UMLS: https://www.nlm.nih.gov/research/umls/

a consistent classification of all concepts in the Metathesaurus at a relatively general level represented in the Semantic Network.

Table 1: Comparison of 5 researches associated with applications of techniques of NLP in clinical notes

	(Savova et al., 2008)	(LePendu et al., 2012)	(Wu et al., 2012)	(Pakhomov et al., 2005)	(Byrd et al., 2014)
Objective	Explore WSD problem across biomedical literature and clinical notes;	Compute the risk of getting a myocardial infarction	Characterize empirical instances of (UMLS) Metathesaurus term strings; Illustrate what types of term characteristics are generalizable;	Text Analysis and Information Retrieval	Identify Framingham HF signs and symptoms
Dataset	NLM WSD test set; Mayo Clinic;	Stanford Clinical Data Warehouse (STRIDE);	Mayo's Enterprise Data Trust (EDT); I2B2/VA 2010 NLP Challenge data;	351 Documents	Geisinger Clinical (GC) primary care practice EHRs.
Tools	UMLS	National Center for Biomedical Ontology (NCBO) Annotator Web Service; UMLS; RxNORM terminology	UMLS	UIMA; UMLS Lexical Variant Generator (LVG) tool; Shallow Parser; SNOMED-CT. MeSH, RxNorm and Mayo Synonym Clusters (MSC)	LanguageWare; IBM Language Ware Resource Workbench (LRW); UIMA; Concordance program;
Techniques	WSD algorithm;	Data Annotation; Normalizing and Aggregating terms;	Aho-Corasick algorithm;	Part-of-Speech tagging; Maximum Entropy Classifier; Naïve Bayes Classifier;	Basic text processing; Recognize words and phrases using dictionaries and grammars; Build Text Analysis Engines (TAEs);

2.1.2. Metamap

Dr. Alan Aronson's Metamap at the National Medical Library (NLM) is a powerful tool to recognize UMLS concepts in clinical notes[5].. It is able to detect Metathesaurus conceptions in text by mapping biomedical text to or equivalently to UMLS Metathesaurus.

2.2. Topic Segmentation

For extracting desired medical information, we applied Topic Segmentation for automatically splitting clinical notes into topically coherent shorter segments. Topic Segmentation contains two main tasks: text segmentation and topic identification (Li & Yamanishi, 2000). Text segmentation, as its name suggests, aims to automatically divide a single document into coherent passages or subtopics. In this section, we mainly focus on discussing some famous text segmentation models: TextTiling, Beeferman model and TopicTiling, and (ii) application of Text/Topic Segmentation on clinical notes with free-text version.

2.2.1. Algorithms

2.2.1.1. TextTiling

Hearst proposed an approach to the text segmentation named as TextTiling (Hearst, 1993, 1994, 1997), which has three parts: tokenization into terms and sentence-sized units, determination of a score for each sentence-sized unit, and detection of the subtopic boundaries. Tokenization refers to the division of the input text into individual lexical units (Hearst, 1997). In the algorithm of TextTiling, all the tokens in the body of the text are all converted into lower-case characters and stop-words are removed from text in the step of tokenization. Two methods for calculating the score to be assigned at each token-sequence gap are explored by Hearst: blocks and vocabulary introduction. Block refers to a group of token-sentences while block size, labelled ***k***, is the number

[5] https://www.nlm.nih.gov/

of grouped token-sentences compared against an adjacent group of token-sequences. The following equation shows how the lexical score for similarity between blocks is calculated:

$$score(i) = \frac{\sum_t w_{t,b_1} w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}}$$

where

$b_1 = \{\, tokensequence_{i-k}, \ldots, tokensequence_i \,\}$,

$b_2 = \{tokensequence_{i+1}, \ldots, tokensequence_{i+k+1}\}$,

t ranges over all the terms that have been registered during the tokenization step excluding stop words and

$w_{t,b}$ is the weight assigned to term t in block b.

In the version of the vocabulary introduction, the lexical score is calculated by determining the ratio of new words in an interval divided by the length of that interval. See the following equation:

$$score(i) = \frac{NumNewTerms(b_1) + NumNewTerms(b_2)}{w * 2}$$

where $NumNewTerms(b)$ returns the number of terms in interval ***b*** seen for the first time in the text, parameters ***i*** and ***w*** represent token-sequence gap and the length of the token-sequences respectively and $b_1 = \{tokens_{i-w}, \ldots, tokens_i\}$ and $b_2 = \{tokens_{i+1}, \ldots, tokens_{i+w+1}\}$.

Figure 2 illustrates how the boundary is identified by assigning a depth score, the depth of the valley, to each token-sequence gap. In Figure 2(a), the depth score at gap a_2 is $\left(\mathrm{y}_{\mathrm{a}_1} - \mathrm{y}_{\mathrm{a}_2}\right) + \left(\mathrm{y}_{\mathrm{a}_3} - \mathrm{y}_{\mathrm{a}_2}\right)$, while there is a slight difference in Figure 2(b), in which a small valley at gap b_4 can be said to "interrupt" the score for b_2. For avoiding this problem, an algorithm, named as Smoothing the Plot (Hearst, 1993), uses smoothing to help eliminate this small perturbation.

However, another potentially problematic case is shown in Figure 2(c), in which it seems either gap c_2 or gap c_3 or both are supposed to be assigned a boundary. Such "plateaus" occur mainly because vocabulary changes very gradually and reflects a poor fit of the document's corresponding portion to the model TextTiling assumes. Usually, when the plateaus occur over a longer stretch, both bordering gaps are assigned as boundary. If such a plateau takes place over a very short stretch of text, however, the system would make an arbitrary choice.

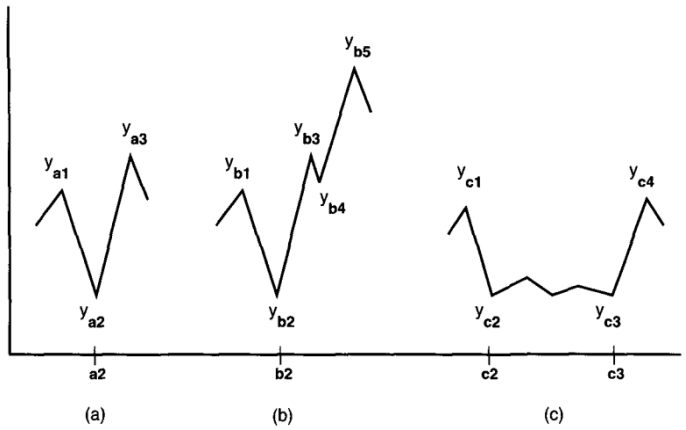

Figure 2: A sketch that illustrates the depth scores calculation in three different situations. The x-axis shows the number of token sequence gaps and the y-axis shows the lexical score (Hearst, 1997)).

TextTiling does not require extensive training on the labelled data. It is designed to identify the subtopics instead of segmenting consecutive documents. Furthermore, TextTiling does not assume the presence of explicit paragraph boundaries since the algorithm is at paragraph level (Beeferman, Berger, & Lafferty, 1999).

2.2.1.2. Beeferman Statistical Models

Statistical models for text segmentation introduced by Beeferman et al. (Beeferman et al., 1999) are well-known and widely used in the field of NLP by many researchers. This approach they proposed is based on feature selection. A set of informative features is collected into a model which can be used to predict location of boundaries in text. The approach they propose is actually

based on the statistical framework for random fields and exponential models for feature selection. This idea is to assign every sentence a probability value by the pre-trained model to detect the probability that there is a boundary between one sentence and the next. Two classes of features are extracted for building model. One is the features of topicality that use adaptive language models in a new way to detect broad subject changes. The other is cue-word characteristics that detect occurrences of specific words that may be domain-specific, which tend to be used near segment boundaries.

Riedl and Chris (Riedl & Chris, 2012) proposed a Text Segmentation algorithm called TopicTiling which is based on TextTiling and leads to significant improvement. TopicTiling model considers the smallest basic unit as a sentence s_i and calculates a coherence score c_p between each position p between two adjacent sentences. They exclusively used the topic IDs assigned to the words by inference to calculate the coherence score. Each block is represented as a T-dimensional vector. The coherence score for each adjacent "topic vector" is calculated by cosine similarity. The value of the coherence score ranges from 0 to 1: close to 0 refers to the marginal connection between two adjacent blocks and close to 1 indicates a substantial connectivity. Figure 3 shows the coherence scores plotted to trace the local minima.

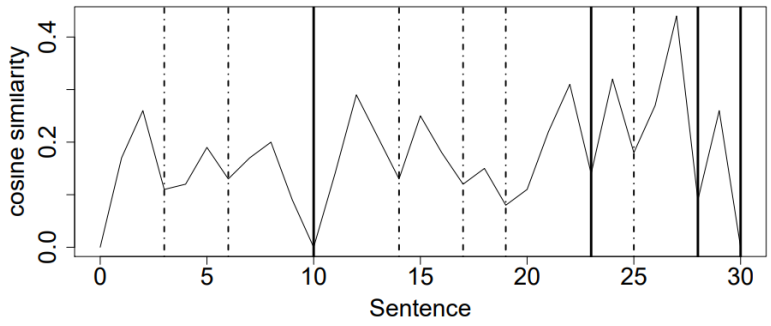

Figure 3 : Similarity scores for a document plotted. The vertical lines show all possible boundaries of segments. The solid lines indicate segments selected according to the threshold criterion if the number of segments is not given in advance (Riedl & Chris, 2012).

To detect the boundaries, rather than using the c_p values in TextTiling, a depth score d_p is calculated for each minimum. The depth score is measured by looking at the highest coherence scores on the left and on the right. Equation 3 shows how the depth score is calculated.

$$d_p = \frac{1}{2} * \left(hl(p) - c_p + hr(p) - c_p\right)$$

where hl represents the highest peak on the left side; hr indicates the highest peak on the right side (see Figure 4).

The n highest depth scores are used as segment boundaries if the number of segments, n, is given. Otherwise, a threshold predicts a segmentation if the depth score is greater than $\mu - \sigma/2$, where μ and σ represent the mean and the standard variation calculated on the depth score respectively.

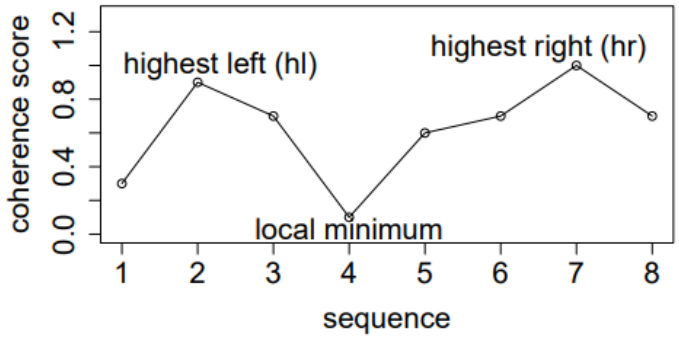

Figure 4: Illustration of the highest left and the highest right peak according to a local minimum (Riedl & Chris, 2012). $hl(4) = 0.93$*, the score value at position 2, and* $hr(4) = 0.99$ *from the value at position 7.*

2.2.1.3. Topic Segmentation algorithms with LDA

In the same paper, Riedl and Chris (Riedl & Chris, 2012) also presented a general method to use Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) as topic model for Text Segmentation. They used two well-known Text Segmentation algorithms, TextTiling and C99, for topic assignments instead of word assignments. LDA introduced by Blei, Ng and Jordan (2003) is a generative model which assigns topics based on a training dataset

The topic-based version of the C99 algorithm (Choi, 2000), called C99LDA, divides the input text into minimal units on sentence boundaries. In this algorithm, there is a similarity matrix $S_{m\times m}$ computed, where *m* represents the number of sentences or units. The cosine similarity is used to determine every element s_{ij} between unit *i* and *j*. For these determinations, a T-dimensional vector represents each unit *i*, where *T* means the number of topics selected for the topic model. The numbers of times topic *ID k* occurring in unit *i* are denoted in each element t_k of this vector. Next, to improve the contrast of S, a rank matrix R must be calculated: each r_{ij} element contains the number of s_{ij} neighbors with lower similarity scores than s_{ij} itself. This step increases the contrasts between regions compared to matrix S. A top-down hierarchical clustering algorithm to divide the document into m segments is then performed in a final step. The entire document is considered at the beginning as one segment in this algorithm. It is then divided until the criteria for the stop are met, e.g. the number of segments or the threshold for similarity. At this, the ranking matrix is split at indices *i, j* that maximize the inside density function D.

$$D = \sum_{k=1}^{m} \frac{sum\ of\ ranks\ within\ sement\ k}{area\ within\ segment\ k}$$

2.2.1.4. LCSeg

Galley et al's LCSeg proposal (Galley, McKeown, Fosler-Lussier, & Jing, 2003) based on the lexical approach to cohesion has been particularly influential in the segmentation of dialogues. Lexical chains (Morris & Hirst, 1991) have been implemented as repetitions of simple terms. Chains are identified in the text for all repeated terms and weighted according to their term frequency (when they are more frequent, the terms would be weighted higher) and chain length (when weighted higher if shorter chains). The cosine distance is then used as the key metric between each pair of windows lexical chain vectors, and the sharpest local minima is taken as the

hypothesized boundaries. It is shown that this methodology has good performance on difficult data (Purver, 2011).

2.2.2. Topic Segmentation for Clinical Notes

In the field of medical informatics, there are several researchers working on Topic Segmentation applied on clinical notes with free-text version. Ginter et al. (Ginter, Suominen, & Pyysalo, 2009) proposed an unsupervised approach of Topic Segmentation and labelling system by combining Hidden Markov models and Latent Semantic Analysis. Theoretical and methodological Latent Semantic Analysis (LSA) is used to extract and represent the contextual meaning of words using statistical computations applied to a large corpus of text (Landauer, Foltz, & Laham, 1998). The introduced method could allow the topic of interest to be defined freely, without data annotation, for identifying short segments. The conditional probabilities $P(w(t)|q(t))$, represented as emission probabilities, and $P(q(t)|q(t-1))$, typically referred to as transition probabilities, are defined and obtained from training data as maximum-likelihood estimates.

Tepper et al. (Tepper, Capurro, Xia, Vanderwende, & Yetisgen-Yildiz, 2012) also worked on section segmentation in free-text clinical notes. They proposed a fully statistical system for section segmentation and classification. The basic proposed methodology of text segmentation is to classify each line, instead of each sentence, into one document for indicating its classification. They attempted two approaches: a joint approach, named as one-step approach, and a pipeline approach, named as two-step approach. They also compared One-step approach and Two-step approach after segmentation. One-step approach is based on a section segmentation model which has already been labelled with section categories. Two-step approach relies on two separate model for section segmentation and classification. Figure 5 illustrates the stages in the one-step and two-step approach for training and testing.

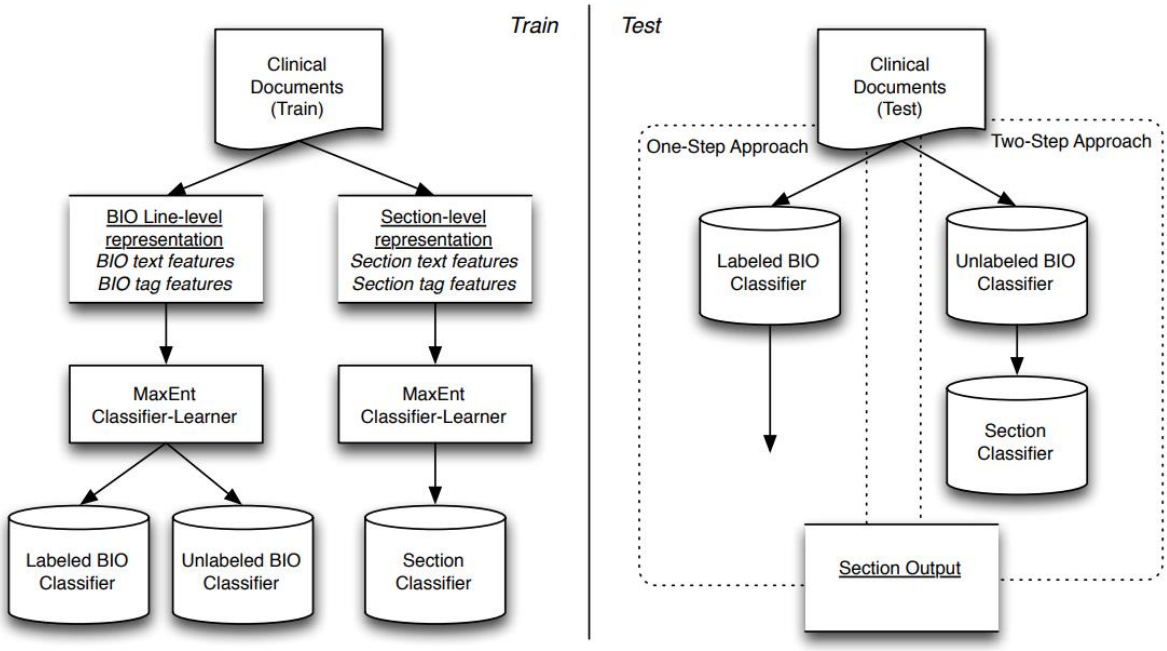

Figure 5: Training and testing stage in the one-step and two-step approach.(Tepper et al., 2012)

Edinger et al. (Edinger, Demner-Fushman, Cohen, Bedrick, & Hersh, 2017) developed search rules in 2017 to identify sections of clinical papers that may be indexed using NLM-facing search engines like Lucene or Essie. In segmenting documents, a subset of each document type was examined, to identify the most common section headings and record all variations in terminology, orthography and punctuation.

Table 2: Comparison of recent researches of Topic Segmentation on clinical notes

	(Ginter et al., 2009)	(Tepper et al., 2012)	(Edinger et al., 2017)
Dataset	Nursing notes of 516 adult ICU patients (breathing, hemodynamics, consciousness, relatives, and diuresis.)	Discharge summaries (General Patient Info, Provider Info, Condition Before Admission, Condition at Discharge, Medical History, Hospital Course, Discharge Instructions, Addenda, Other) and radiology reports (Clinical Info, Exam Details, Findings, Impression and Other)	Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) data (discharge summaries, MD notes, radiology reports, and nursing notes)
Method	Combination of LSA and HMMs: modeling the problem of segmenting the clinical texts and assigning a topic to each resulting segment as a sequence labeling task.	Classify each line in a document to indicate its membership to a section - (Maximum Entropy for classification and Beam search for finding a good tag sequence)	Searching-Headings-Based
Advantages	Unsupervised: allows the topics of interest to be easily changed by specifying key words	Better performance than others on three different datasets	Fast and Easy to change topics
Drawbacks	Applicable to short segments (average length of a topic segment is 18 tokens.)	Intensive annotation	Limited to the format of data (spelling, abbreviations and etc.)
Evaluation Metric	Windiff (around 0.2)	F-1 score (around 0.9)	No Evaluation shown

Table 2 compares the researches mentioned before. In summary, clinical notes has been attempted by these researchers whose methodologies could be used to segment short notes into sentences. However, our Topic Segmentation algorithm can deal with long clinical notes written by professional doctors. In addition, traditional methods, such as regular expression and heading-based searching, cannot be suitable for a variety of format pattern of medical records. Our algorithm, to some extent, could solve this problem easily.

2.3. Medical Information Extraction

With the rise of digital age, news, articles, social media, and so on are an explosion of information. Much of this data lies in unstructured form and it is tedious, boring and labor intensive to manage it manually and to make effective use of it. This explosion of information and the need for more sophisticated and efficient information handling tools give rise to IE (Information Extraction) and IR (Information Retrieval) technology (Singh, 2018). Information Extraction, or Information Retrieval, is a subtask of automatically extracting desired data from unstructured and/or semi-structured documents in natural language processing. Natural Language Processing (NLP) aims to process spoken or written form of such free text which acts as a mode of communication commonly by using computational methods (Assal, Seng, Kurfess, Schwarz, & Pohl, 2011). Information extraction plays an important role in Natural Language Processing for helping human obtain desired information from various documents in an instant.

Information Extraction is often an early stage in pipeline for various high-level tasks such as Question Answering Systems, Machine Translation, event extraction, user profile extraction, and so on (Assal et al., 2011). In this study, we aim to extract medical information from clinical notes, which means information extraction is our final goal in this task. There are a wide variety of subtasks in Information Extraction, such as: Named Entity Recognition, Conference Resolution, Named Entity Linking, Relation Extraction and so on. In this study, we applied the technique of Named Entity Recognition to extract medical entities from clinical notes to achieve the goal of medical information extraction.

2.3.1. Named Entity Recognition

Name-Entity Recognition is a process used to obtain information, which seeks to determine and identify a text string in predefined categories, known also as entity identification, entity chunking and entity extraction. There are several and different definitions of Named Entity.

However, these definitions can be categorized in terms of four criteria according to an analysis(Marrero, Urbano, Sánchez-Cuadrado, Morato, & Gómez-Berbís, 2013): grammatical category, rigid designation, unique identification and domain of application. Here is an example[6]:

*At least **26** people have been killed as a result of the storm in **North Carolina**, including in **Union County**.*

In the above example, the bolded tokens hold the key information and are helpful for further language processing applications.

Figure 6 also shows a short paragraph with several Named Entities in different colour tags.

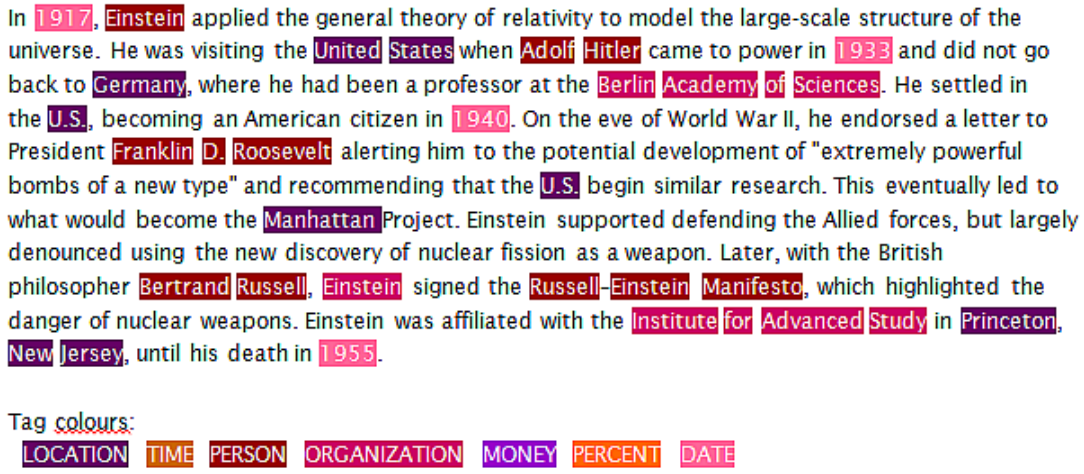

Figure 6: A paragraph showing examples of Named Entities (Source: http://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/)

Named Entity Recognition is a significant process in Natural Language Processing for advanced applications, such as Information Extraction, Question Answering and Machine Translation (Mohit, 2014). There are two challenges in Named Entity Recognition: 1) recognition of named entity boundaries; 2) recognition of named entity categories. Like most problems in other natural language processing tasks, the ambiguities are the most challenging in the process of NER.

[6] BBC News: https://www.newsobserver.com/news/local/article218984940.html

Ambiguity could exist between named entities and common words, such as the word "May", which could be "month", "verb" and even "surname". Besides, there are ambiguities between named entity types as well, for example, "Washington" might be classified as Location or Person.

2.3.1.1. Hidden Markov Model (HMM):

Morwal et al. (Morwal, Jahan, & Chopra, 2012) proposed to achieve the task of NER by using Hidden Markov Model (HMM) which is a popular statistical Markov model. A hidden Markov model is a tool for representing probability distributions over sequences of observations (Ghahramani, 2001). In this model, there are two distributions (Dietterich, 2002): transition distribution $P(y_t|y_t-1)$, which tells how adjacent y value is related, and the observation distribution $P(x|y)$, which tells how observed x valued related to hidden y values (see Figure 7).

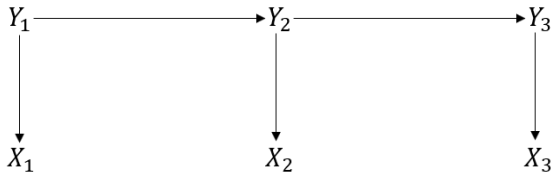

Figure 7: The model of HMM. Y refers to states sequence while X represents observation sequence.

Figure 8 shows a concrete example of HMM, in which there is a 30% probability that tomorrow will be sunny if today is rainy, which means transition probability is 0.3. The emission probability represents the distribution of the observed variable at a time given the state of the hidden variable at that time. In this example, the guy has a 50% chance to do cleaning if it is a rainy day.

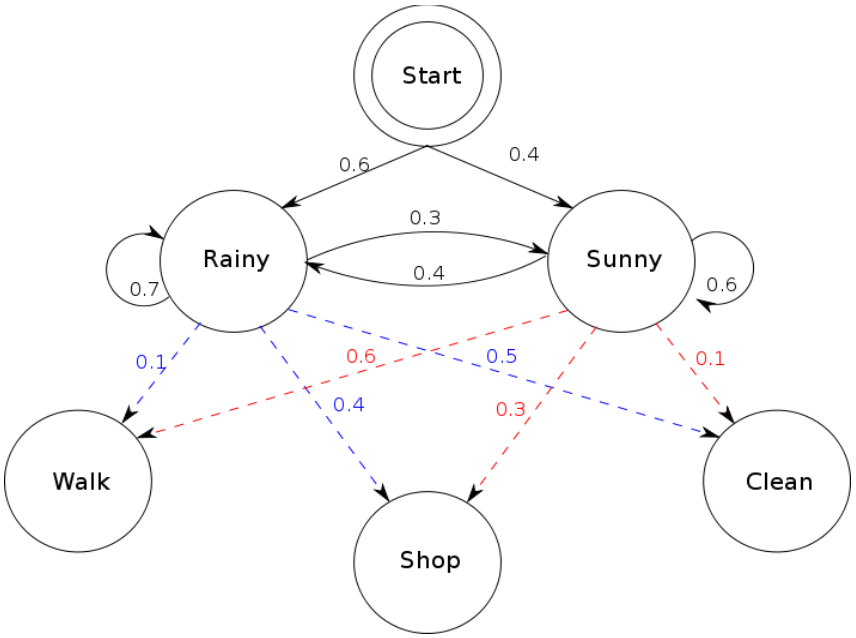

Figure 8: A concrete example of Hidden Markov model (source: https://en.wikipedia.org/wiki/Hidden_Markov_model)

There are three key problems of interest that must be solved for the HMM model to be useful in real-world applications (Rabiner & Juang, 1986):

1). Given the observation sequence $O = O_1, O_2, \ldots \ldots O_T$, and the model $\gamma = (A, B, \pi)$, how we compute $P_r(O|\gamma)$, the probability of the observation sequence. Where:

$A = \{a_{ij}\}, a_{ij} = P_r(q_j \ at \ t+1 | q_i \ at \ t)$, state transition probability distribution;

$B = \{b_j(k)\}, b_j(k) = P_r(v_k \ at \ t \ | q_j at \ t)$, observation symbol probability distribution in state j;

$\pi = \{\pi_i\}, \pi_i = P_r(q_i \ at \ t = 1)$, initial state distribution;

- Evaluation problem: how can we compute the probability that the observed sequence was produced by the model given a model and a sequence of observations

2). Given the observation sequence $O = O_1, O_2, \ldots \ldots O_T$, how we choose a state sequence $I = i_1, i_2, \ldots \ldots, i_T$ which is optimal in some meaningful sense.

- Estimation Problem: how can we estimate the optimal state sequence given an observation sequence and a model.

3). How we adjust the model parameters $\gamma = (\mathrm{A}, \mathrm{B}, \pi)$ to maximize $\mathrm{P_r}(O|\gamma)$.

- Optimize the model parameters to best describe how the observed sequence comes about.

Rabiner and Juang (Rabiner & Juang 1986) presented solutions to the three above-mentioned HMM problems. With efficient learning algorithms, HMM has a strong statistical foundation. It depends, however, only on each state and its respective object observed.

Chieu and Ng (Chieu & Ng, 2002) have presented, using global information, a maximum enterropy named entity recognizer. The trainings features of the model consist of two classes: local and international. Local characteristics are based on nearby token, including token itself, while international characteristics are extracted from other occurrences of the same token in the whole document.

2.3.1.2. Maximum-Entropy Markov Models (MEMM):

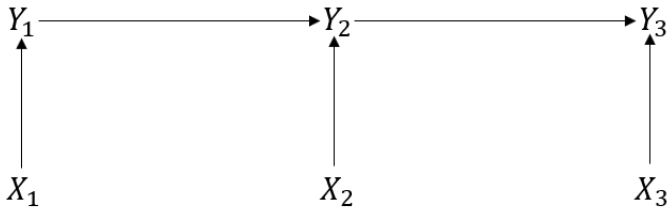

Figure 9: The model of MEMM. Y refers to states sequence while X represents observation sequence.

Maximum-entropy Markov Models, short for MEMM, are a variation on the traditional Hidden Markov Models (HMM). These models are a powerful probabilistic tool for modelling sequential data (Lafferty, McCallum, & Pereira, 2001). Instead of transition function and observation function in HMM, a single function $P(s|s', o)$ is used to predict the probability of the

current state s given the previous state s' and the current observation o in the models of MEMM. In HMMs, the current observation only depends on the current state. In contrast, previous state may also affect the current observation (see Figure 9).

Model: Suppose there is a sequence of observations $O_1, O_2, \dots\dots, O_n$ with the labels of $S_1, S_2, \dots\dots, S_n$. The goal of MEMM model is to maximize the conditional probability $P(S_1, S_2, \dots\dots S_n | O_1, O_2, \dots\dots O_n)$.

$$P(S_1, S_2, \dots\dots S_n | O_1, O_2, \dots\dots O_n) = \prod_{t=1}^{n} P(S_t | S_{t-1}, O_t)$$

Each of these transition probabilities comes from the same general distribution $P(s|s', o)$.

$$P(s|s', o) = P_{s'}(s|o) = \frac{1}{Z(o, s')} exp(\sum_{a} \lambda_a f_a(o, s))$$

Here, the λ_a are parameters to be learned and $Z(o, s')$ is the normalizing factor which makes the distribution sum to one across all next state s. Each feature a gives a function $f_a(o, s)$ which is a real-valued or categorical feature-function. In the study of (Lafferty et al., 2001), fuction $f_a(o, s)$ is defined as following:

$$f_a(o_t, s_t) = \begin{cases} 1 & if\ b(o_t) is\ true\ and\ s = s_t \\ 0 & otherwise \end{cases}$$

2.3.1.3. Condition Random Fields (CRF):

CRF is another probabilistic model for sequence segmentation and labeling data (Lafferty et al., 2001). CRF offers several advantages over HMM and also prevents the basic restriction on Markov Maximum Entropy Models (MEMM). The advantages and limitations will be detailed out in the next section 2.3.1.4. This is why CRF is widely used to recognize named entities for the extraction of information.

Definition (Lafferty et al., 2001): Let $G = (V, E)$ be a graph such that $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field in case, when conditioned on X, the random variables $\mathbf{Y}_v$ obey the Markov property with respect to the graph:

$p(\boldsymbol{Y}_v \mid \boldsymbol{X}, \boldsymbol{Y}_w, w \neq v) = p(\boldsymbol{Y}_v \mid \boldsymbol{X}, \boldsymbol{Y}_w, w \sim v)$, where w ~ v means that w and v are neighbors in G (see Figure 10).

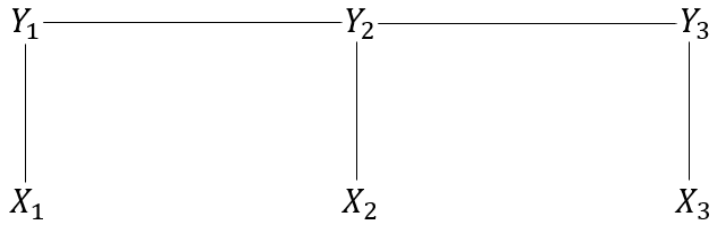

Figure 10: The model of CRF. Y refers to states sequence while X represents observation sequence.

CRF model is a type of discriminative model. Given observation sequence **O**, the state sequences **I** could be predicted by the following formula:

$$P(I|O) = \frac{1}{Z(O)} \prod_i \varphi_i(I|O) = \frac{1}{Z(O)} \prod_i e^{\sum_k \lambda_k f_k(O, I_{i-1}, I_i, i)} = \frac{1}{Z(O)} \prod_i e^{\sum_i \sum_k \lambda_k f_k(O, I_{i-1}, I_i, i)}$$

Where i refers to the location of a token; k represents the kth feature function with a weight value λ_k; Each $token_i$ is assigned with M features. $Z(O)$ is normalizing function.

CRF model could be used for completing many NLP tasks, such as Part-of-speech tagging and Named Entity Recognition. McCallum and Li (McCallum & Li, 2003) used the CRF algorithm to extract named bodies in the joint task competition for coNLL2003. Sarawagi and Cohen propose a semi Markov CRF recognition algorithm (Sarawagi & Cohen, 2005). The Markov semi model was further extended with a dictionary and the concept of the similarity function.

2.3.1.4. Comparison of HMM, MEMM and CRF

HMM, MEMM and CRF are three popular statistical sequence modelling methods, often applied to achieve NLP tasks such as Part-of-Speech Tagging, Named Entity Recognition and

other machine-learning-related problems. Figure 11 intuitively shows the difference between the models of HMM, MEMM and CRF.

HMM has a strong statistical foundation with efficient learning algorithms where learning can take place directly from raw sequence data. This model assumes that each observation is independent, and current state is only related to previous state. In the task of Part-of-Speech or NER, however, every state is not only associated with current token but the context, the length of token. MEMM model could perfectly solve HMM's limitation. Unfortunately, MEMM has labelling bias problem.

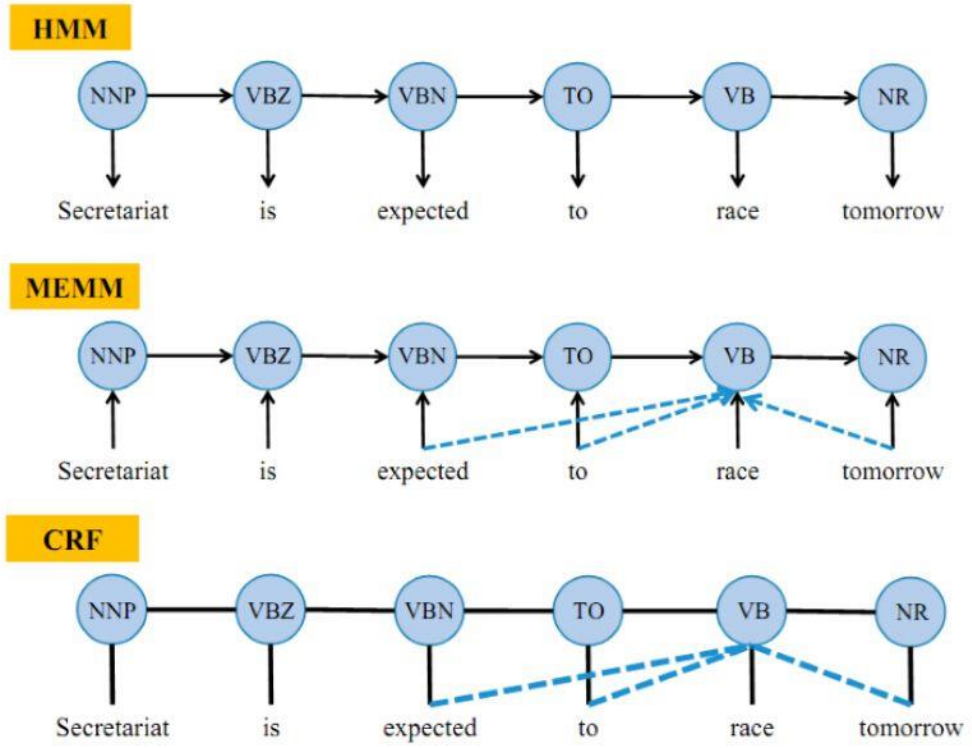

Figure 11: Graphic Structure of HMMs, MEMMs and CRFs for a sequence - 'Secretariat is expected to race tomorrow. (Source: Wikipedia)

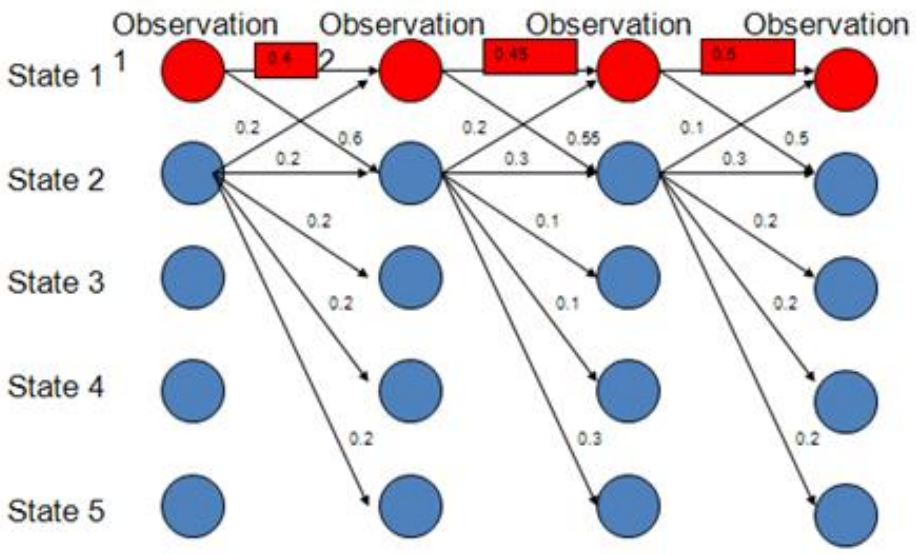

Figure 12: An example to MEMM labelling bias problem. State 1 tends to transfer to state 2 while state 2 tends to stay at state 2. (Source: Wikipedia)

Figure 12 shows why there is a labelling bias problem in MEMM. As shown on Figure 12, state 1 tends to transfer to state 2 while state 2 tends to stay at state 2. However, the optimal path is state1-state1-state1-state1 ($0.4 \times 0.45 \times 0.5 = 0.09$ the highest score). The reason why this situation happens is because state 2 has more convertible states than state 1 dose which reduces the probability for state 2 to convert to next state. CRF successfully avoids this limitation of MEMM since CRFs computes probability of global optimal output nodes. Meanwhile, CRF does not have as independence assumptions as HMM, it has an ability to adopt any context information with designing features flexibly.

The section mainly details out a comparative analysis between HMMs, MEMMs and CRFs. MEMMs and CRFs are primarily discriminative sequence models while HMMs are generative sequence models. CRF could overcome all the drawbacks in HMMs and MEMMs.

2.3.2. Medical Information Retrieval from Clinical Notes

As we discussed in 2.1, the adoption of electronical health records (EHRs) has been increasing rapidly in the hospitals and clinics, which is desirable to harvest information and knowledge from EHRs in order to support automated systems and care and enable the secondary use of EHRs in clinical and translation research (Wang et al., 2018). However, most of the EHRs data is in free-text form (Jensen et al., 2017) and (Hearst, 1993).

As noted in the introductory section, the extraction of clinical notes is essential for the health care provider in the diagnosis process. Medical recognition is still a popular topic in the field of health and computer technology. Kumar et al. (Kumar et al., 2014) presented an integrated approach in the extraction of medical entities through the modeling of the systems using Conditional Random Field (CRF) from patient discharge (see Figure 13). They have added Boolean characteristics aside from basic word-based features to define if the present token is a medication or symptom or a particular medical condition with a broad dictionary of medical terminology (medical conditions, symptoms and medicines) using SNOMED-CT.

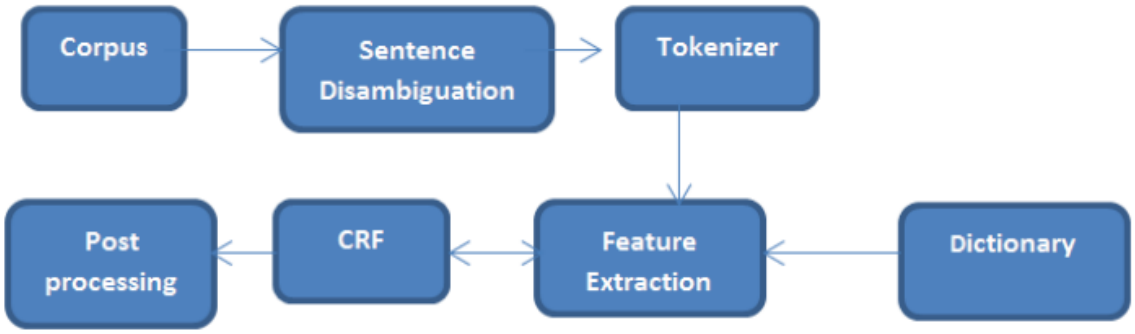

Figure 13: Architecture of system proposed by Kumar et al. (2014)

Finally, the $\boldsymbol{F_1}$ score they obtained of each medical named entity is around 75% compared with Patrick and Li (Patrick & Li, 2010) who developed a medication entity recognizer based on CRF as well with the results of 84.44%. Patrick and Li (Patrick & Li, 2010) developed a novel

supervised learning model which has the ability to incorporate several rule-based engines and two machine learning algorithms. These two machine learning algorithms were conditional random fields (CRF), which was used to extract named entities, and support vector machines (SVM), which was utilized to classify the relationship between two entities. Figure 14 shows the structure of this system. Similarly, Wang (Wang, 2009) introduced a method to identify clinically named persons with orthographic, lexical and semantic characteristics using CRF. The author compares an F 1 score of 64.12 percent and 81.48 percent, respectively, based upon a rule and a CRF system. In the same year, Wang and Patrick improved their algorithm by adding more features for CRF model and obtained an overall accuracy of 83.3% (see Table 3).

Table 3: Comparison of the research of Medical Named Entities Recognition using CRF model

	(Wang & Patrick, 2009)	**(Patrick & Li, 2010)**	**(Kumar et al., 2014)**
Dataset	Admission summaries from an ICU	I2B2 Discharge Summaries	I2B2 corpus
Entities	Body; Finding; Behavior; Object; Observable; Organism; Procedure; Qualifier; Occupation; Substance;	Medication, Dosage, Mode, Frequency, Duration and Reason entities	Symptoms; Medications; Generic medical named entity
Features	Word features; Orthographic features; Affixes; Context Features; Dictionary Features; Abbreviations and Acronyms; POS features;	Drug feature set; Dosage feature set; Mode feature set; Frequency/duration feature set; Reason feature set; Morphology feature set; Five-word context window;	Word based features; Semantic Knowledge; Orthographic features; Parse tree features; Dictionary based features; To further reduce noise related; Extensive regex-based features; Character level features; A spelling dictionary check;
Evaluation Metrics	F-1 Score (0.799)	F score (0.849)	F score (0.798)

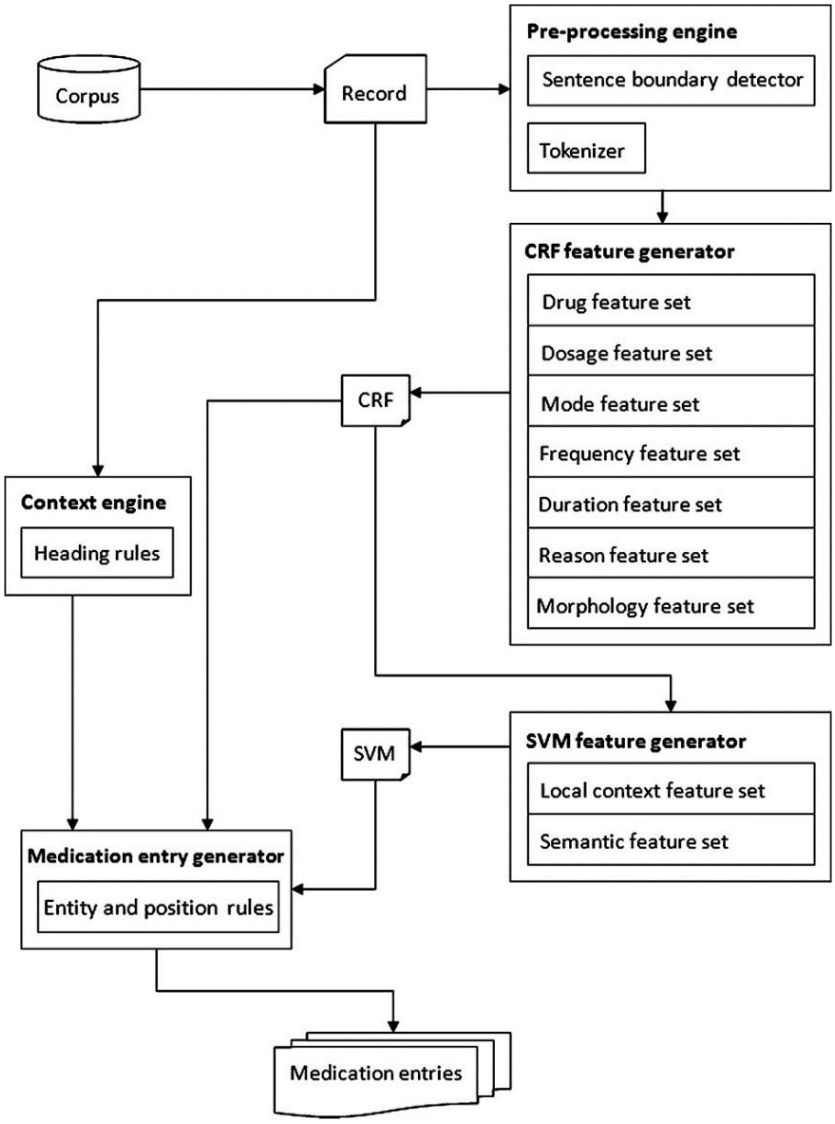

Figure 14: Medication extraction system architecture diagram. CRF, conditional random fields; SVM, support vector machines (Patrick & Li, 2010).

2.4. Medical Information Visualization

This section discusses current techniques and research on Medical Information Visualization. Medical Information Visualization aims to visualize medical data, such as medicines, diseases, human body parts and so on, over a specific interface for interaction with human beings to help them obtain the medical knowledge in a short time. Information visualization

techniques and examples are firstly discussed for demonstrating a brief introduction of InfoVis. Finally, current research of medical InfoVis are presented.

2.4.1. Information Visualization Techniques and Examples

Information visualization is the study of visual representations of abstract data to reinforce human cognition. Figure 15 shows three simple examples of everyday information visualization. According to the article[7], multiple techniques for information visualization could be split into several groups according to different needs. 2D-dimensional area, multi-dimensional data visualizations, hierarchical data visualizations, network data models and temporal visualizations are widely used data visualization models. Figure 16 shows four well-known data visualization techniques. Node-link diagram (a) is a kind of network data models, in which a circular image with dots represents the data nodes and lines refer to the links between nodes. This model helps users understand the relationship between the data sources rapidly and efficiently. In temporal visualization, polar area diagram (see b) is a complex model, which looks like a standard pie chart, yet the distance from the center in addition to the arc length and angle is used to evaluate the size of the sector. Thus said, a blunt sector stretched near away from the center might be less important than a sharp sector which does not reach near. Hierarchical data could be visualized by dendrogram (c) model, which helps understanding relations between data in an instant.

[7] https://itsvit.com/blog/big-data-information-visualization-techniques/

Figure 15: Examples of Everyday Information Visualization. a) Top left, Subway map of Toronto, Canada. b) Top right, Plant hardiness zone map of Canada. c) Bottom, map of direction from Ottawa to Montreal, Canada. Author/Copyright holder: Google, Inc.

Kapler and Wright (Kapler & Wright, 2005) developed an effective technique to visualize the spatial information over time and geography within a single, highly interactive three-dimensional view. It is effective when applied to analysis of complex past and future events within a geographic context since it could demonstrate data spatially and temporally at a time.

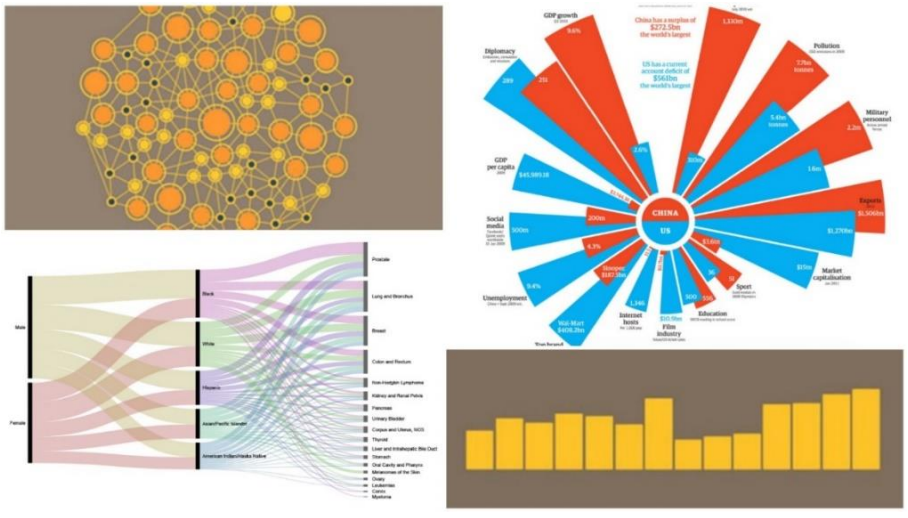

Figure 16: Example of Information Visualization Technique. a) top left – node-link diagram; b) top right – polar area diagram; c) bottom left – dendrogram; d) bottom right – histogram.

2.4.2. Medical Information Visualization

Information visualization in medical area has been around, but it has not been reached the full potential of information visualization. In this section we discuss about two types of medical information visualization systems: non-commercial systems, in which part research related to medical information visualization would be introduced, especially the pictorial visualization system proposed by Suo (Suo, 2017) and Jin (Jin, 2016); commercial systems, where some popular EMR systems will be discussed.

Assal et al. (Assal et al., 2011) proposed a prototype Image Overlay system in which the images are transformed in real-time so that the user could be an integral part of the surrounding environment. Image Overlay, a computer display technique, combines computer images and user's direct view of the real world together. For example, reconstructed 3D CT medical image of a bone can be displayed to a surgeon inside the patient's anatomy at the exact location of the real bone, no matter where the surgeon locates, to help the performance of surgery. Their research has shown that Image Overlay, the information visualization technique, could be applied for a wide variety of medical applications, for instance: intraoperative guidance and surgical education.

Another medical data visualization system based on secure cloud proposed by Mohanty et al. (Mohanty, Atrey, & Ooi, 2012) has an ability to protects the security of data at the cloud centers. They integrated a cryptographic secret sharing approach with pre-classification volume ray-casting. Their results showed that their framework has a high safety factor.

2.4.2.1. Pictorial Visualization System

Pictorial visualization system for clinical notes have been studied by Jin and then Suo who proposed a prototype web-based pictorial visualization system that could be operated by both patients and doctors. Ruan et al. (Ruan et al., 2018) further developed this system by including

natural language processing, which is related to this thesis. So here we introduce the system more in detail.

This pictorial visualization system they developed is, to some degree, an interactive visualization system in which medical information are navigated over pictorial-based user interface. Medical records are divided into multiple classes to be visualized in two interfaces: one for the position - based representation of a space interface and one for a time-based medical information representation.

In a spatial medical record, illness or diseases location is organized in accordance with its physiological systems. 11 physiological systems (integrated, skeletal, muscular, cardiovascular, digestive, nervous, immune, respiratory, reproductive, endocrine, urinary and skin systems) are categorized in this system. Figure 17 shows its relevant images with font and back views. The medical information would be presented at a particular location with a circle on the corresponding image of the physiological system (see Figure 18). The colors indicate the physiological system to which it belongs. Black circles mean a symptom that is still unknown to a physiological system. Moreover, an entire body problem image would be used if the disease does not have a certain location, such as Myasthenia Gravis.

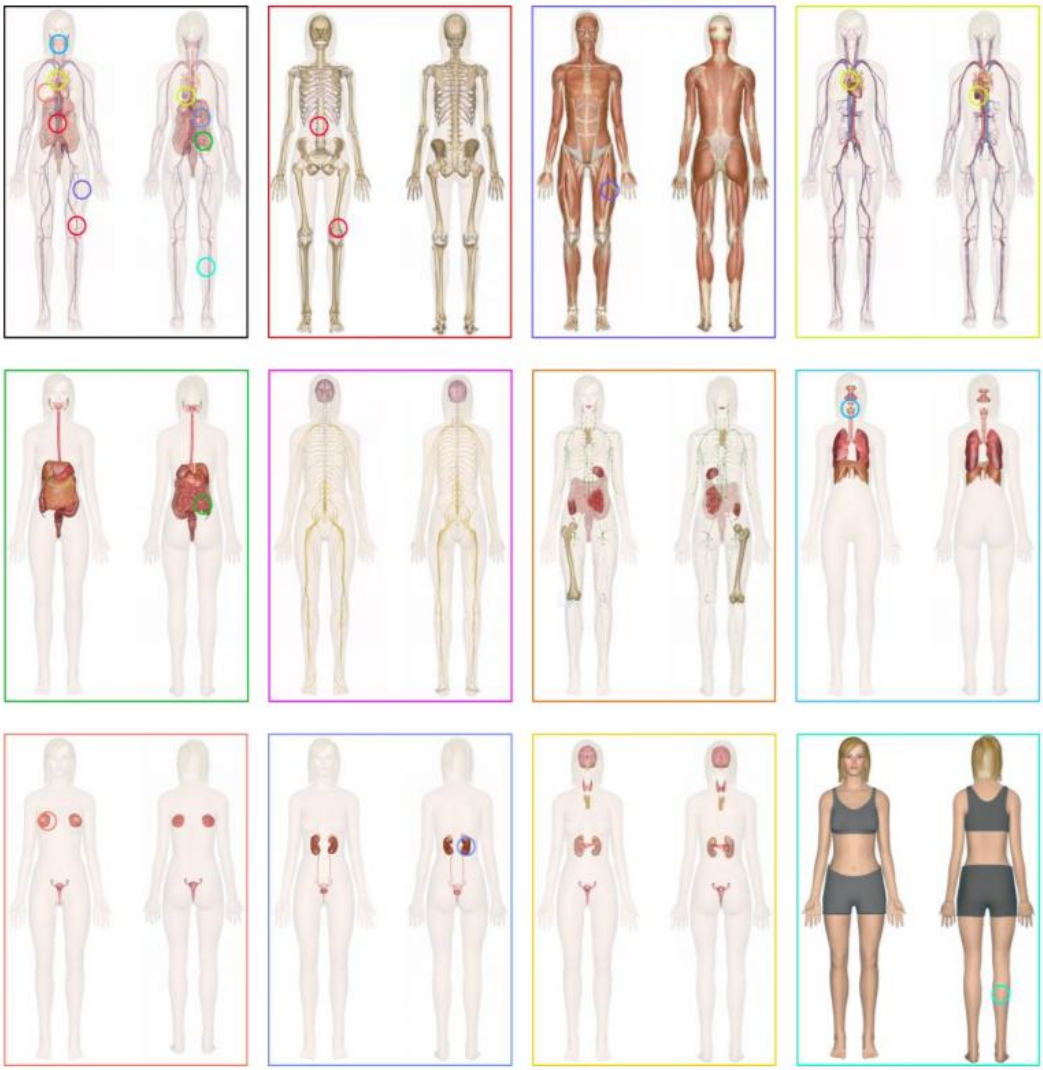

Figure 17: Display of physiological systems in front and back views. The diseases are partitioned into 11 physiological systems so that the information can be easily accessed on the basis of disease groups. Left to right: integrated, skeletal, muscular, cardiovas. (Ruan et al., 2018)

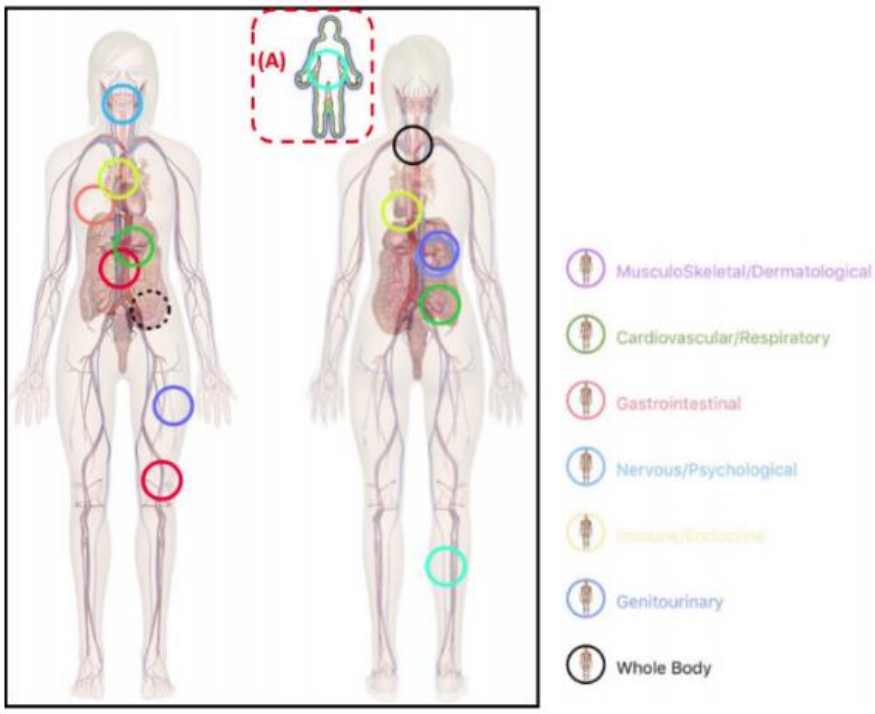

Figure 18: Spatial Interface - Position-based medical information on a 2D representation of the human body. The coloured circle shows the location of symptom and diagnosis of specific physiological system. e.g. red circle means the diagnosis belongs to musculoskeletal system. Whole-body problem is indicated separated in (A).(Ruan et al., 2018)

Temporary medical records deal with time-based data, which display medical stories like a data sheet enabling information visualization and health evolution on time. A temporary summary can be used as the whole medical record, but only separate details can be obtained. The categories include visits, medical imaging, medication, laboratory testing and treatment. They are presented with different icons on the timeline. Figure 19 shows a time-based representation sample.

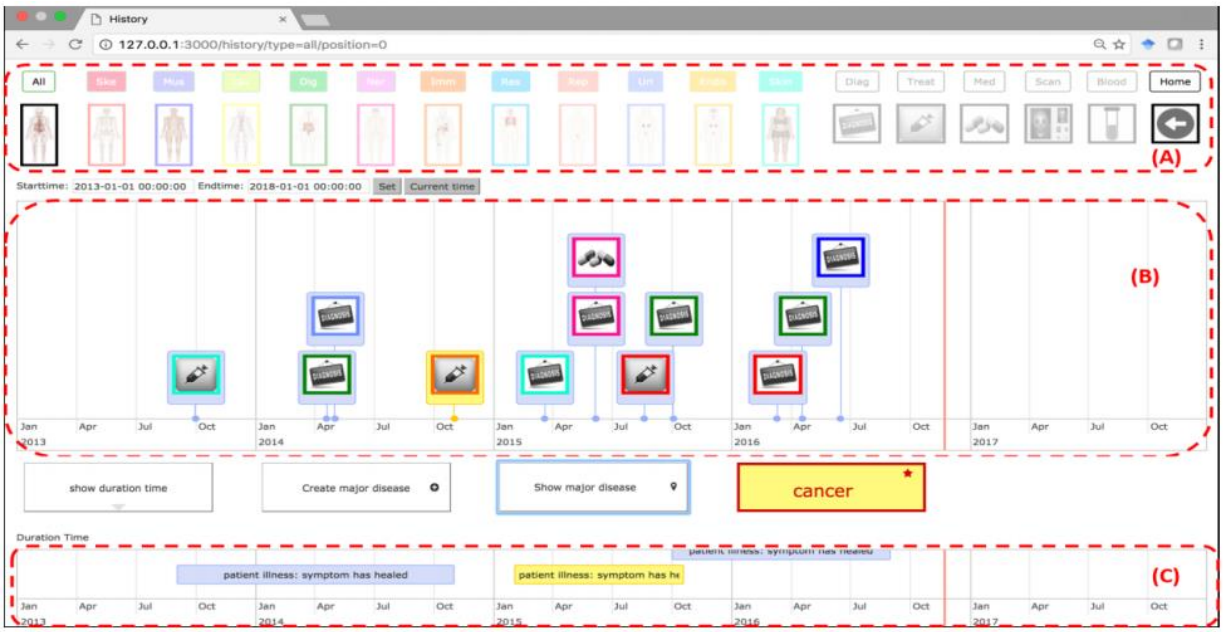

Figure 19: Information timeline display. The horizontal line denotes time information, and the icons represent different medical events. The button at the bottom show duration timelines and major diseases (Ruan et al., 2018).

2.4.2.2. EMR - Commercial systems

Electronic medical records (EMRs) are gaining increasing prominence in the delivery of healthcare, although the focus is primarily on deploying EMRs (Kaelber, Greco, & Cebul, 2005). In this section, we discuss two famous commercial EMRs.

EpicCare: a solution for health recording with robust features and features, complex workflow design, user-friendly interface, mix of chart check, documentation and order management. In order to make doctors productive through simplifying workflows, patient-facing health care elements and tracking bills, invoices and payments, EpicCare integrates clinical and income cycle management systems into an EMR solution. The EpicCare Ambulatory-Core EMR uses predictive analytics and decision support tools to ensure the delivery of safe and quality care. This capacity is demonstrated and applied to the various software modules.

eClinicalWorks: an Integrated Electronic Health Record (EHR) and solution for practice management. EClinicalWorks provides technology through each step of the delivery process of patient care. This solution includes patient engagement, population health, coordination of care and financial analytics. Integration of eClinicalWorks enables structured data capture and trend analysis, while customizable documentation options support clinicians across multiple specialties. All patient chart components including billing, patient demographics, previous and upcoming appointments, and outstanding items are permitted to be accessed by users. EClinicalWorks could be used on various devices

Chapter 3. Topic Segmentation

This chapter will detail out our Topic Segmentation Algorithm. For better presenting the algorithm, we split this chapter into three parts. We firstly provide an overview of the segmentation algorithm primarily using a diagram which could clearly shows our idea of Topic Segmentation. In the next section, we mainly illustrate how we train and build the Topic Score Predictor that is the most important part in our segmentation algorithm. Datasets used, features and model selection are respectively presented in detail. The last section in this chapter describes how the documents are segmented by using Topic Score Predictor. The steps of boundaries detection would be mentioned in this section.

3.1. Overview

In this study, we applied Topic Segmentation for automatically dividing a clinical note with free-text and English version into segments with 5 categories: History, Medications, Physical Examinations, Laboratories and Hospital Course. We also name this study as Boundary Detection by Determining the Difference of Classification Probabilities of Sequences (Ruan & Lee, 2018a). Segmenting the clinical notes is necessary for the further process of information extraction and visualization. Figure 20 illustrates the technical process of Topic Segmentation in our study. We firstly collected 1127 text plain clinical notes in English language version from I2B2 for Text Classification model training to build Topic Score Predictor. As for the part of Text Classification, we experimented it based on Naïve Bayes and Support Vector Machine models with Bag of Words (BOW), Word2vec and their corresponding versions with TFIDF features.

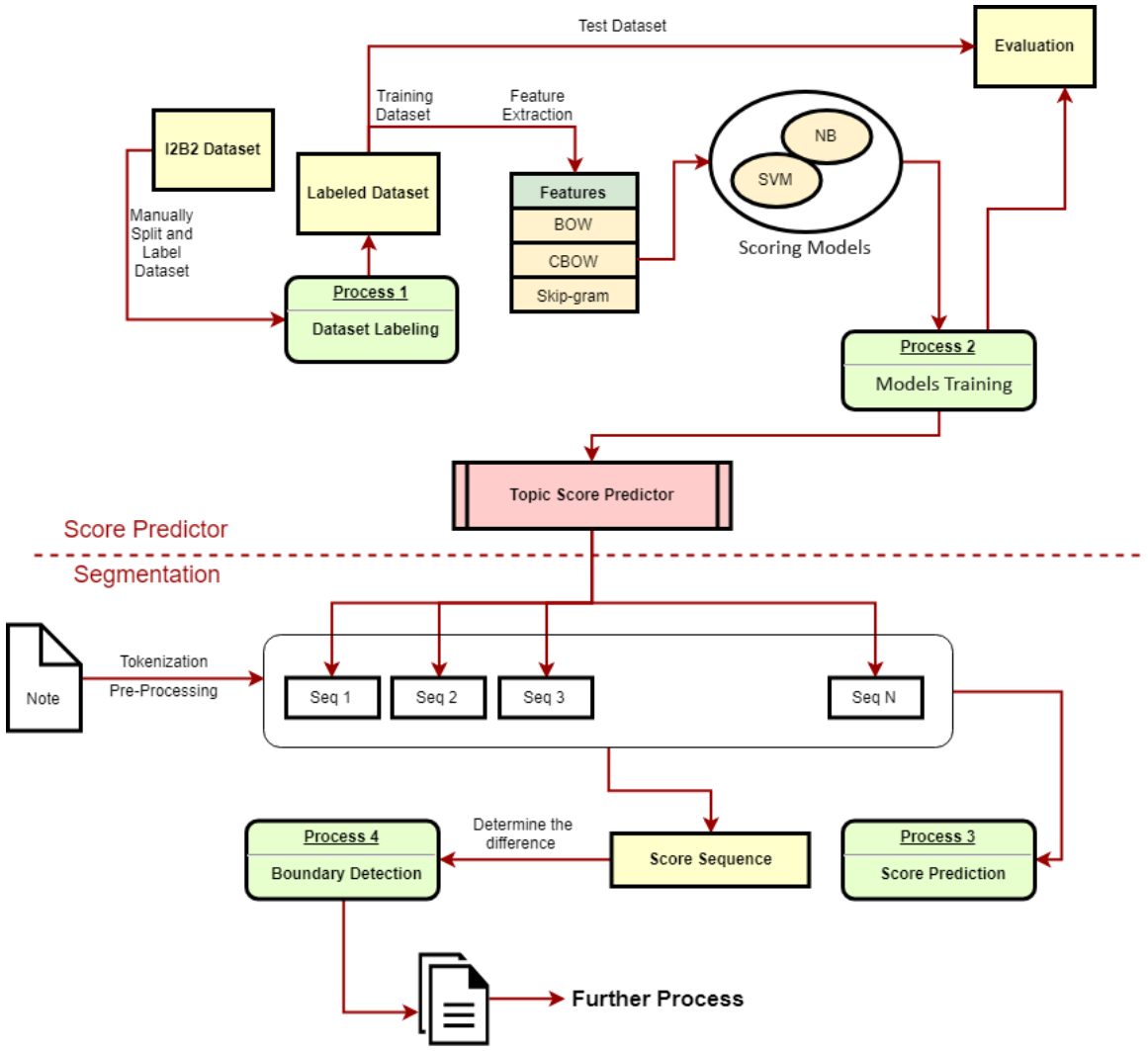

Figure 20. The framework of Topic Segmentation for clinical notes in this study

The core technique of Topic Semination is boundary detection. Topic score predictor could give each sequence or sentence scores five score each of which is the probability of belonging to corresponding topic. Finally, the segmenter could predict tokenized sequences' (S_i) probability $P_{C_k}(S_i)$ of belonging to each category C_k for boundary detection by determining the difference of $P_{c_k}(S_i)$ with respect to S_i, while S_i represents i^{th} tokenized sequences. Next sections will detail out these processes.

3.2. Why Topic Segmentation

Topic Segmentation plays an important role in the process of information extraction. It could divide the text into several small segments each of which has specific topic, which enhances the accuracy of information extraction. Currently, text-plain format with patients' medical information documents are clinical notes, also named as progress notes, which contain crucial medical information, such as medications, allergies and family history.

Table 4 : A sample of History of Present Illness. In UMLS, the concepts ID of Aspirin, Dopamine, Atropine, Lidocaine and Dyazide are C0004057, C0805940, C0004259, C0023660 and C0058829 respectively belonging to Organic Chemical, Pharmacologic Substance

HISTORY OF PRESENT ILLNESS: The patient is a 69-year-old woman from Maine without prior history of overt cardiac disease *...* *She was given Streptokinase, aspirin, Dopamine for hypotension and Atropine with external Zoll pacemaker after bradycardia into the 30 's developed* *...* *she was given a Lidocaine bolus and brought to the FIH by helicopter. Her only medication was Dayside*

Over 1500 clinical notes have been released from Informatics for Integrating Biology and the Bedside (I2B2) as research data sets for general research purposes, which sufficiently embodies the value of clinical records in the field of clinical medicine and NLP. Many researchers are working on the data mining, also named as information retrieval, of free-text clinical records. Patrick and Li (Patrick & Li, 2009) proposed a string-map-based system called as Intelligent Clinical Notes System (ICNS) that could extract needed information for doctors from clinical notes. Since string-map-based methodology of data mining requires a large number of databases, they employed SNOMED CT (Donnelly, 2006), Systematized Nomenclature of Medicine – Clinical Terms, for medical terminology data retrieval, which however, ignored the semantic relationships

expressed between paragraphs and even sentences, just simply matching the required string with the string in the database.

Apparently, ignoring semantics easily causes over-matching. Table 4 is part of the history of current disease in the clinical notes. A compendium of many controlled biomedical science vocabularies, UMLS (Unified Medical Language System) (Bodenreider, 2004), classifies organic chemicals (C0004057) as a pharmacologic substance, commonly considered as a medication. Assuming that a doctor needs the current medicines of a patient and that data set method for extracting information about medicines would be used, aspirin, dopamine, atropine, and others not supposed to be present.

In addition, doctors usually mention patients' admission medications and discharge medications (Table 5) when they make a clinical progress. Admission medications and discharge medications are different. When the patients are discharged by the hospital there can be many changes to their home medicine or admission medicines. These changes are very important and can help prevent a readmission to the hospital. Just simple string map extraction does not have the ability to connect the context to correctly retrieve data. The topical segmentation of clinical notes (text segmentation and theme identification) is needed to enhance medical data mining.

Since most of the clinical notes are structured following a regular and general pattern in forms with titles and subtitles and Ruan et al. (Ruan et al., 2018) employed regular expression to detect the titles and subtitle for segmentation. The titles and subtitles, however, are not consistent varying quite depending on the source of clinical notes. Firstly, subsection structure is not consistent with each other having various titles even for a similar topic. Secondly, lowercase and uppercase are both used in these clinical notes as titles. Thirdly, there are many abbreviations of subtitles. For convenience, doctors always tend to simplify the words or phrases when making a clinical note. For instance, HPI is popularly used as History of Present Illness. Same as Family History, Physical Examinations and Hospital Course abbreviated as FH, PE and HC respectively.

More seriously, subtitles are omitted in some short notes. Lastly, there are some notes not organized in a fixed format, named as unstructured data, in which the information is commonly described in general terms. Regular expression almost lost its role in segmentation.

Table 5: A sample of admission medications and discharge medications. q.day, or qd and q.d, means one a day; b.i.d, or BID and bid, means twice a day; t.i,d, or TID and tid, means three times a day;

MEDICATIONS ON ADMISSION: Vasotec 40 mg q.day , Soma 1 tablet q.day , Demerolprn , Clonidine . *...* *MEDICATIONS ON DISCHARGE: Vasotec 20 mg PO b.i.d. , Clonidine 0.2 mg PO b.i.d. , Nifedipine 20 mg PO t.i.d. , Flexeril 10 mg PO t.i.d. , Valium 5 mg PO t.i.d. , Micronase 1.25 mg q.day .*

3.3. Topic Score Predictor

The idea of the segmenter in our study is to assign each sequence with five topic scores each of which is the value of probability of belonging to corresponding topic. We employee Naive Bayes and Linear SVM models with features of BOW and Word2vec with/without TFIDF for training Topic Score Predictor.

3.3.1. Dataset Used

All the datasets in our study are collected from I2B2[8], which is a biomedical informatics research data warehouse. I2B2 is short for Informatics for Integrating Biology and the Besides, which was an NIH-funded National Center for Biomedical Computing based at Partners HealthCare System. For better understanding the genetic bases of complex disease, a scalable

[8] "Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY."

informatics framework developed by the i2b2 NCBC is designed to connect the vast data banks and clinical research data. The main source of I2B2 data for research data warehouse is the Epic EHR. A variety of discrete information from EHR, such as demographics (age, gender, race, etc.), diagnoses (ICD-9), medication orders and allergies, have been loaded into the data warehouse. They have also added additional information such as, history (surgical, medical, social, family) and other condition-specific variables. Other than these, they are in process of adding billing information to the warehouse.

The i2b2 framework consists of facts and dimensions, which employs a simple, yet powerful data model. Figure 21 shows a star schema utilized by the I2B2 database. This schema consists of one fact table surrounded by numerous dimension tables.

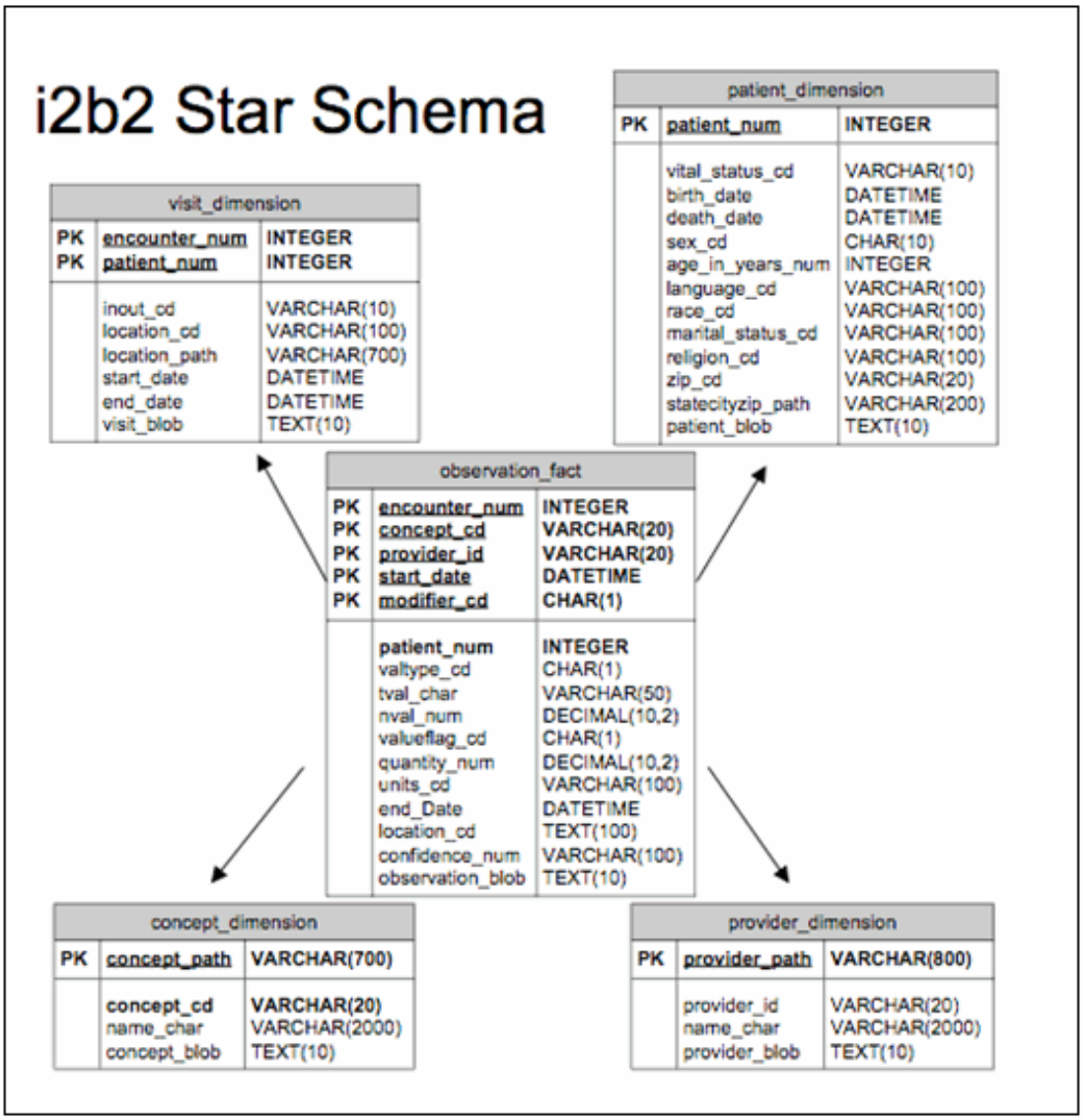

Figure 21: I2B2 start schema (Source: I2B2)

Facts in i2b2 are observations about a patient, including information such as history, diagnoses, demographics, laboratory results, etc. The advantage of using this star schema model is that any data can be easily added and integrated into database from various sources which do not need to change the underlying architecture or redesign the system. All new observations are simply added to the fact table.

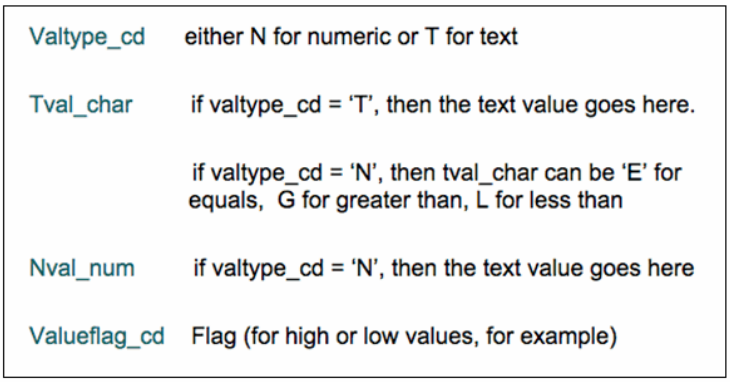

Figure 22: Information stored about a laboratory result "fact." (Source: I2B2)

Users can create a hierarchical categorization of the different concepts in the dataset by using metadata employed by i2b2 which helps accurately describe and navigate through facts. Query terms are selected from these categorizations including history, medications, demographics, diagnoses, laboratory tests and procedures. In instances where standardized terminologies are used to capture the data, such as ICD-9 for diagnoses and CPT for procedures, those hierarchies can be directly ported into the metadata table (FAQ | i2b2 Research Data Warehouse). An example is shown on Figure 22.

There are over 1500 notes released from the first four i2b2 Challenges as i2b2 NLP Research Data Sets. To access these notes, user need to sign standard Data Use Agreement[9]. In this study, we collected 1253 text plain clinical notes in English language from i2b2 with 126 notes abandoned due to existence of a plenty of unknown characters. The remaining 1127 clinical notes are split into two parts (see Table 6). One part which contains sequences from 1027 notes is used for Topic Score Predictor model training and test while the other 100 notes are for segmentation evaluation.

We manually extract and divide each note into five parts: History, Hospital Course, Medications, Physical Examinations and Laboratories, as a new dataset (see Figure 23) for training

[9] https://www.i2b2.org/NLP/DataSets/AgreementAR.php

Topic Score Predictor. This new dataset has 5123 segments each of which is labelled as its corresponding category. Five labels A, B, C, D, E respectively represent History, Hospital Course, Medications, Physical Examinations and Laboratories.

Table 6. The use of i2b2 dataset for Topic Segmentation

<table>
<tr><th colspan="2" rowspan="2">Number of Notes</th><th rowspan="2">Use</th><th colspan="4">Segments</th></tr>
<tr><th>Topic</th><th>Number</th><th colspan="2">Possible Contents</th><th>Label</th></tr>
<tr><td rowspan="8">1127</td><td rowspan="7">1027</td><td rowspan="7">Topic Score Predictor Training and Testing</td><td>History</td><td>2242</td><td colspan="2">Medical, Medication, Family and Social History</td><td>A</td></tr>
<tr><td>Hospital Course</td><td>755</td><td colspan="2">Hospital Course, Medical imaging</td><td>B</td></tr>
<tr><td>Medications</td><td>907</td><td colspan="2">Admission, Discharge Medications</td><td>C</td></tr>
<tr><td>Physical Examinations</td><td>766</td><td colspan="2">Physical Examinations</td><td>D</td></tr>
<tr><td>Laboratories</td><td>433</td><td colspan="2">Laboratories</td><td>E</td></tr>
<tr><td rowspan="2">TOTAL</td><td rowspan="2">5123</td><td>3586</td><td colspan="2">Train</td></tr>
<tr><td>1537</td><td colspan="2">Test</td></tr>
<tr><td>100</td><td colspan="5">Segmentation Test</td></tr>
</table>

3.3.2. Feature Selection

Feature selection is the process of selecting a subset of the terms in the training set and only using this subset as text classification features (Manning, Raghavan, & Schütze, 2010). There are two reasons why we must do feature selection in text classification. Firstly, feature selection could enhance the accuracy of classification by avoiding noise features. A noise feature is one that increases the classification error on new data when added to the document representation. Suppose there is a rare term, "abcd", which has no information about a class, but all instances of "abcd" happen to occur in "history" documents in the training set. In this case, the trained models might mis-assign test document containing "abcd" to "history". Sometimes such an incorrect generalization from an accidental property of the training set is named as overfitting. The other reason is that feature selection could make training and applying a model more efficiently by decreasing the size of the effective vocabulary. This is of importance for models, such as SVM, that are expensive to train.

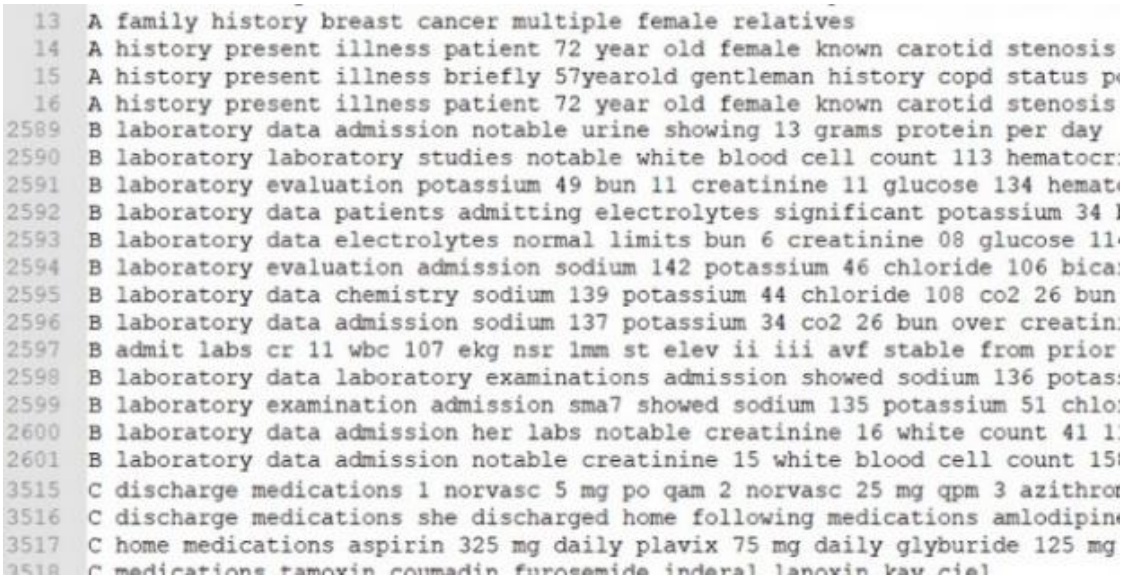

Figure 23: The format of the dataset for training Topic Score Predictor. Each line represents a segment with its corresponding topic label at the beginning of each line. A, B, C, D, E respectively represent History, Hospital Course, Medications, Physical Examinations and Laboratories.

3.3.2.1. Bag of Words (BOW)

In natural language processing, the bag-of-words model is the most popular and simplest representation of documents (Zhang, Jin, & Zhou, 2010). In this model, a document is considered as the bag of its words disregarding word order and grammar but keeping multiplicity.

Suppose there is a training set in which there are 2 documents:

> *1). The patient is recently diagnosed abdominal carcinomatosis.*
>
> *2). The patient presented with abdominal pain and abdominal bloating.*

Based on this tiny training set, a dictionary could be obtained as following:

> *{"the", "patient", "is", "recently", "diagnosed", "abdominal", "carcinomatosis", "presented", "with", "pain", "and", "bloating"};*

This dictionary contains every token occurring in the training set, which is the idea of Bag of Words. After the text is transformed into a "bag of words", each document in the training set could be represented as a list to record the token frequencies of all the distinct words:

1). [1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]

2). [1, 1, 0, 0, 0, 2, 0, 1, 1, 1, 1, 1]

In our training set for topic score predictors, the dictionary is 4079-dimension.

3.3.2.2. Word2Vec

Word2Vec is a group of related models which are used to produce word embeddings. The idea of Word Embedding is to transfer a word vector with high dimension into a relatively lower dimension. Input a large corpus of text, Word2vec could produce a vector space, usually just a hundred dimensions, with each unique word in the corpus being assigned a corresponding vector

in the space. In this section, we applied two well-known word embedding models in Word2vec, including continuous bag-of-word model and the continuous Skip-Gram model, to produce word embedding.

3.3.2.3. CBOW

The continuous bag-of-word model, commonly abbreviated to CBOW, predicts the probability of a word given a context. Figure 24 (left) shows how CBOW model works. In this model, the input layer is made of the one-hot encoded input context words $\{x_1, \dots\dots, x_c\}$ with vocabulary of size V and a word window of size C. A N-dimensional vector h is the hidden layer. Finally, in the training example, the output layer is output word y which is also encoded one hot. $V \times N$ weight matrix W connects one-hot encoded input vectors to the hidden layer, and the hidden layer is connected via a $N \times V$ weight matrix W' to the output layer.

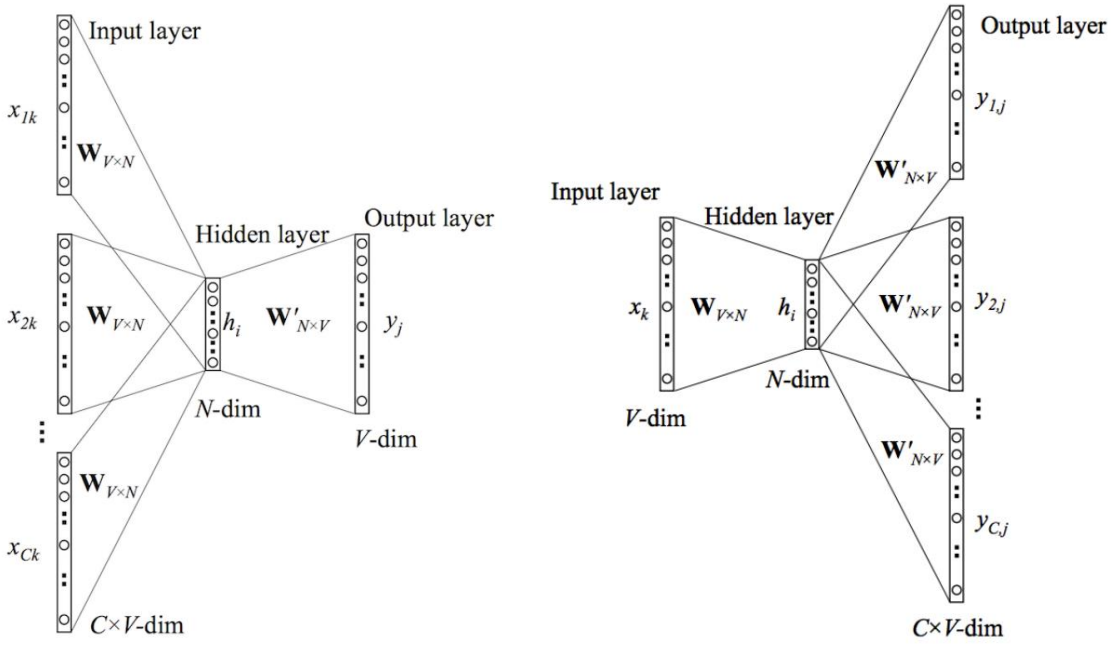

Figure 24 : This figure demonstrates how CBOW (left) and skip-gram (right) works. W and W' must be learnt. (Source: http://elgibborsms.com/blog/intuitive-understanding-of-word-embeddings-count-vectors-to-word2vec/)

3.3.2.4. Skip-gram

Contrary to CBOW, the Skip-Gram model aims to use central word as input to predict the surrounding words in a fixed window. Figure 24 (right) illustrates how Skip-gram works. As seen from the figure, the input word in the training instance represented as the one-hot encoded vector x and $\{y_1, \dots \dots y_c\}$ are the one-hot encoded vectors corresponding to the output words in the training instance. The weight matrix W between input layer and hidden layer is $V \times N$ dimension. The i^{th} row in hidden layer refers to the weights corresponding to the i^{th} word in the vocabulary. As we mentioned before, word embedding aims to reduce word vectors from a high dimension to a lower dimension. Therefore, the weight matrix W is what we are interested in learning because it contains the vector encodings of all of the words in the vocabulary. As for output layer, each output word has a matrix W' with $N \times V$ dimensions. In our study, the vocabulary is transformed into 100-dimension word vectors.

3.3.3. Scoring Algorithms

To find the best classifiers, we primarily employed two algorithms: Naïve Bayes (NB) and Support Vector Machine. The former one is generative model while the latter one is discriminative model. This section will detail out these two classifiers.

3.3.3.1. Naïve Bayes Model

The classifier Naive Bayes (NB) is known to be simple and very efficient, which is supervised learning method based on the theorem of Bayes which is a valuable tool for dealing with uncertain information. Building classifiers is a simple technique. Bayesian networks have been widely used for the processing of uncertain information in intelligent systems and have been used successfully in medical diagnosis, statistical decision making, expert systems and other fields. Abstractly, Naïve Bayes is a conditional probability (a measure of the probability of an event given that another event has occurred) model. Given an event represented by a vector $\boldsymbol{x} = (x_1, \dots \dots x_n)$

in which each x_i represents a feature (each feature is independent variable), the NB model assigns to this event probabilities

$$p(C_k|x_1, \dots \dots x_n) = p(C_k|\boldsymbol{x})$$

For each if K possible outcomes or classes C_k. Based on Bayes' theorem, the above formulation could be decomposed as

$$p(C_k|\boldsymbol{x}) = \frac{p(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})}$$

In practice, we are only interested in the numerator of above faction since the denominator does not depend on C and feature x_i. So that

$$p(C_k|x_1, \dots \dots x_n) \propto p(C_k, x_1, \dots \dots x_n) = p(C_k)p(x_1|C_k)p(x_2|C_k)p(x_3|C_k) \dots \dots$$
$$= p(C_k)\prod_{i=1} p(x_i|C_k)$$

Finally, a Naïve Bayes classifier is the function that assigns a class label $\mathcal{Y} = C_k$ for some *k* as follows:

$$\mathcal{Y} = \underset{k\in\{1,2,..K\}}{argmax}\ p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

There are two popular distribution in Naïve Bayes: Multinomial NB and Bernoulli NB. Multinomial NB, as implied in the name, implements the NB algorithm for multinomial distributed data, while Bernoulli NB, however, deals with data that is distributed according to multivariate Bernoulli distributions (Manning et al., 2010; Mccallum & Nigam, 1998; Metsis, Androutsopoulos, & Paliouras, 2006). In a model of a multinomial event, feature vectors represent the frequencies with which a multinomial $(p_1, \dots, p_n)$ generates certain events where p_i is the probability that event i will occur. However, features are independent binary variables representing inputs with a Bernoulli NB model. We use Multinomial Naive Bayes and Bernoulli Naive Bayes model in this study to compare model perfection

3.3.3.2. Support Vector Machine Model

Support Vector Machines (SVM) are also well-known supervised learning methods which can analyze data, identify patterns, and use for classification and regression analysis. The linear SVM algorithms try their best to find a hyperplane to split the sample with maximum margin between positive and negative data.

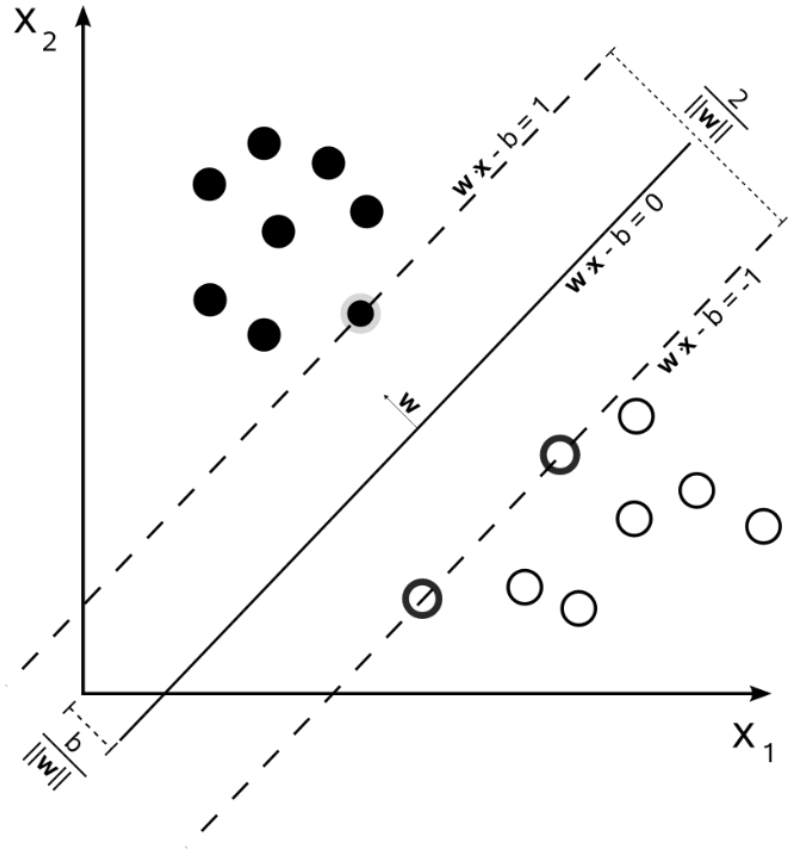

Figure 25: Maximum margin and margins for a SVM with samples from two classes trained. Samples on the margin are called vectors of support (source: Wikipedia).

Suppose there is a training dataset of n points of the form $(\boldsymbol{x_1}, y_1), \ldots\ldots (\boldsymbol{x_n}, y_n)$, where each $\boldsymbol{x_i}$ labeled with y_i which are either 1 or -1, indicating the class to which the point $\boldsymbol{x_i}$ belongs. SVM model wants to find the "maximum-margin hyperplane" which could split the group of points $\boldsymbol{x_i}$ for which for which $y_i = 1$ from the group of points for which $y_i = -1$, which is defined

so that the distance between the hyperplane and the nearest point $\boldsymbol{x_i}$ from either group is maximized (see figure 25).

We employed Linear Support Vector Machine algorithm to train and test classification model and compared it with Multinomial NB and Bernoulli NB model. A Linear Lin-SVM scores the best for text classification task and Multi-NB does the next best. Note that it is not a score for Topic Segmentation task. Support Vector Classifier is trained based on one-vs-the-rest (Nasrabadi, 2007) scheme which involves training a single classifier per class, with positive samples and negatives.

3.3.4. Training

Firstly, pre-processing steps, such as tokenization, removing stopwords and punctuations, are applied to the training and test document. We employee the default English stopwords list on RANKS NL[10] for removing stopwords in the clinical notes. We used "models.word2vec" in gensim[11] for feature selection to achieve modelling CBOW and Skip-gram feature. Sklearn[12], an efficient data mining and data analysis tool, is employed to train NB and SVM models. Figure 26 shows a small part of pre-trained dictionary based on NB model with BOW feature. Each word is assigned with a 5-dimension vector in which each element is the probability of belonging to corresponding topic.

[10] https://www.ranks.nl/stopwords
[11] Gensim: https://radimrehurek.com/gensim/models/word2vec.html
[12] http://scikit-learn.org/stable/

```
sublingually    [6.4250835260858394e-06, 2.21837703536093e-05, 0.00010723860589812332, 1.2764219340345144e-05, 5.213546880213547e-06]
prn [0.00012850167052171679, 2.21837703536093e-05, 0.009360398314821907, 3.829265802103543e-05, 0.0003597347347347476]
chest   [0.004735286558725263, 0.004702959314965172, 0.0004442742244350823, 0.0037654447054018174, 0.002591132799466133]
pain    [0.007369570804420457, 0.000443675407072186, 0.0031711987744159325, 0.0009445522311855407, 0.002648481815148482]
surgical    [0.0013685427910562838, 4.436754070721 86e-05, 1.5319800842589045e-05, 0.0002297559481262126, 0.0006412662662662662]
history [0.023663582626574146, 0.0001109188517680465, 9.191880505553428e-05, 0.0002169917287858 6747, 0.0017204704704704705]
significant [0.0013813929581084554, 0.001375393761923 7766, 7.659900421294524e-05, 0.0008041458184417441, 0.0010635635635635637]
hypertension    [0.003514520688768954, 8.873508141443 72e-05, 4.595940252776714e-05, 7.658531604207086e-05, 0.0007768184851518184]
basal   [5.782575173477255e-05, 0.0001109188517680465, 1.5319800842589045e-05, 2.5528438680690288e-05, 9.384384384384384e-05]
cell    [0.0004754561809303521, 0.0024845822796042416, 3.063960168517809e-05, 0.0001531706320841 4173, 0.0006047714381047715]
carcinoma   [0.0005461320997172963, 2.21837703536093e-05, 3.063960168517809e-05, 1.2764219340345144e-05, 0.0001303386720053 3866]
nose    [5.782575173477255e-05, 2.21837703536093e-05, 6.127920337035618e-05, 0.0007403247217400183, 3.128128128128128e-05]
six [0.0004304805962477512, 2.21837703536093e-05, 0.000658751436231329, 6.382109670172573e-05, 0.0004587921254587914]
years   [0.0023451554870213315, 4.436754070721 86e-05, 4.595940252776714e-05, 3.829265802103543e-05, 9.90573907240574e-05]
ago [0.0017347725520431765, 4.436754070721 86e-05, 0.0001072386058981 2332, 3.829265802103543e-05, 6.256256256256256e-05]
cataracts   [9.637625289128758e-05, 2.21837703536093e-05, 1.5319800842589045e-05, 3.829265802103543e-05, 1.0427093760427094e-05]
hysterectomy    [0.0002891287586 7386276, 2.21837703536093e-05, 1.5319800842589045e-05, 3.829265802103543e-05, 9.384384384384384e-05]
social  [0.0028206116679516833, 4.436754070721 86e-05, 3.063960168517809e-05, 1.2764219340345144e-05, 0.0001720470470 4704705]
lives   [0.0012914417887432536, 2.21837703536093e-05, 1.5319800842589045e-05, 1.2764219340345144e-05, 3.128128128128128e-05]
his [0.00796710357234644, 0.003549403256577488, 0.0010111106855610877, 0.006165117941386705, 0.012517726059392727]
wife    [0.0006746337702390131, 2.21837703536093e-05, 3.063960168517809e-05, 1.2764219340345144e-05, 6.777610944277611e-05]
he  [0.017302749935749165, 0.001042637206619637, 0.0009191880505553428, 0.005169508832839784, 0.013033867200533867]
cigar   [2.5700334104343358e-05, 2.21837703536093e-05, 1.5319800842589045e-05, 1.2764219340345144e-05, 5.213546880213547e-06]
smoker  [0.0003533795939 3472114, 2.21837703536093e-05, 4.595940252776714e-05, 1.2764219340345144e-05, 5.213546880213547e-06]
rarely  [6.425083526085839e-05, 2.21837703536093e-05, 1.5319800842589045e-05, 1.2764219340345144e-05, 5.213546880213547e-06]
drinks  [0.0002955538421999486, 2.21837703536093e-05, 4.595940252776714e-05, 2.5528438680690288e-05, 1.0427093760427094e-05]
alcohol [0.0013878180416345412, 6.655131106082 79e-05, 0.0001072386058981 2332, 3.829265802103543e-05, 0.0001251251251251512]
family  [0.0028591621691081985, 4.436754070721 86e-05, 9.191880505553428e-05, 0.0001148779740631063, 0.0007351101101101101]
cardiac [0.0019082498072474942, 0.001331026221216558, 0.0001378782075833 0143, 0.002144388849177984, 0.0020176426426426427]
disease [0.004568234387047032, 0.000665513110608279, 0.0001225584067407 1236, 0.0002169917287858 6747, 0.0020854187520854186]
present [0.0050115651503469544, 0.0002883890145969209, 3.063960168517809e-05, 0.0010083733278872665, 0.0002398231564898231 5]
illness [0.004953739398612182, 4.436754070721 86e-05, 1.5319800842589045e-05, 1.2764219340345144e-05, 5.213546880213 5466e-05]
mr  [0.0008673862760215882, 0.0001996539331824837, 4.595940252776714e-05, 0.0001148779740631063, 0.0007716049382716049]
marsh   [1.9275250578257516e-05, 2.21837703536093e-05, 1.5319800842589045e-05, 1.2764219340345144e-05, 5.213546880213547e-06]
```

Figure 26: A small part of trained dictionary based on Naïve Bayes Model with BOW feature. Numeral vector represents the probability of belonging to corresponding topics.

For the further process of segmentation, the performance of each classifier with various features has been compared by calculating the traditional evaluation F_1 metric. From Table 7 it is clear to see that Linear Support Vector Machine with BOW feature performs best while Multinomial does the next best. Interestingly, with the same BOW feature, the performance of Multinomial NB classifier and linear SVM classifier is close with the difference of 0.02 and 0.07 using BOW and BOW-TFIDF respectively. However, these two classifiers offer a big difference accuracy using word2vec features. Linear SVM classifier is much better with around 0.70 F_1 score, while NB only has 0.50. This is mainly because the probability distribution of each feature in NB model can be independently estimated as one-dimensional distribution. However, word2vec transformed the distribution into 100-dimensional vector, which easily lose some feature information to some extent. As for Support Vector Machine, it is considerably memory efficient due to its own advantage of kernel mapping to high-dimensional feature spaces, which makes it perform better than NB-based classifier.

We select BOW as the feature to train the Topic Score Predictors based on MNB (Multinomial Naïve Bayes) and Linear SVM for building topic segmenters. The performances of these two segmenters are compared in Chapter 5.

Table 7: F_1 score of the text classification with different algorithms and features; Lin-SVM does the best for text classification task and Multi-NB does the next best. Note that it is not a score for Topic Segmentation task.

Feature		Classifiers		
		Multi_NB	**Bern_NB**	**Lin_SVM**
BOW		0.96745	0.88737	0.98046
BOW_TFIDF		0.91549	0.88346	0.98242
Word2vec	CBOW	0.51497	0.44076	0.74674
	CBOW_TFIDF	0.53581	0.43880	0.76693
	Skip-Gram	0.52279	0.42578	0.78971
	Skip-Gram_TFIDF	0.52279	0.42643	0.81185

3.4. Boundary Detection and Segmentation

The Topic Score Predictor might predict and assign 5-dimensional vector sequences in which each element is the likely to belong to the respective topic. The next step is the segmentation of topics in this study. Figure 27 demonstrates these steps by using diagram which is clearer to understand our idea.

1. Tokenize text T into n (number of tokenized sentences) sentences s_i;
2. Let $t = 1$;
3. Topic score prediction respectively assign $s_1, s_2, s_3, \dots\dots, s_t$ with a 5-dimensional vector $\boldsymbol{v} = [score_{history}, score_{labs}, score_{meds}, score_{PE}, score_{course}]$ and obtain a $t-dimensional$ accumulated score vector $\boldsymbol{\rho} = [\boldsymbol{as_1}, \boldsymbol{as_2}, \dots\dots, \boldsymbol{as_t}]$. Each *element* in vector $\boldsymbol{\rho}$ could be obtained by using: $\boldsymbol{as_i} = \boldsymbol{v_1} + \boldsymbol{v_2} + \cdots + \boldsymbol{v_i}$ which means that each value in

vector as represents the accumulated probabilities of belonging to each class: $P(s_1) + P(s_2) + \cdots + P(s_t) = \sum_{i=1}^{t} P(C_k|s_i)\ i < n$ where: C_k : class $k \subset$ *{history, labs, medications, physical exams, hospital courses} and* s_i: the i^{th} tokenized sentence or sequence s;

4. Analyzing vector $\boldsymbol{\rho}$**,**

   If return 0, let $t = t + 1$,

   If $t > i$, $s_1 + \cdots + s_{t-1}$ would be a segment;

   Else, go back to step 3;

   If return 1 and $topic_{index}$, $boundary_{index} = t - 1$. In other words, $s_1 + \cdots + s_{boundary_{index}}$ is a segment which discusses about $topic_{index}$. Simultaneously, let $T = s_t + \cdots + s_i$ and go back to step 1 to segment the rest text;

The boundary gets detected and the topic assignment is done.

3.4.1. Analyzing vector $\boldsymbol{\rho}$

By analyzing vector $\boldsymbol{\rho}$, our idea is to detect the variety of each topic score with the detected sequences location changing to detect the boundary. Let us take a brief note[13] as an example to better understand this section, illustrating how to detect the border.

[13] This note is quoted from I2B2: https://www.i2b2.org/

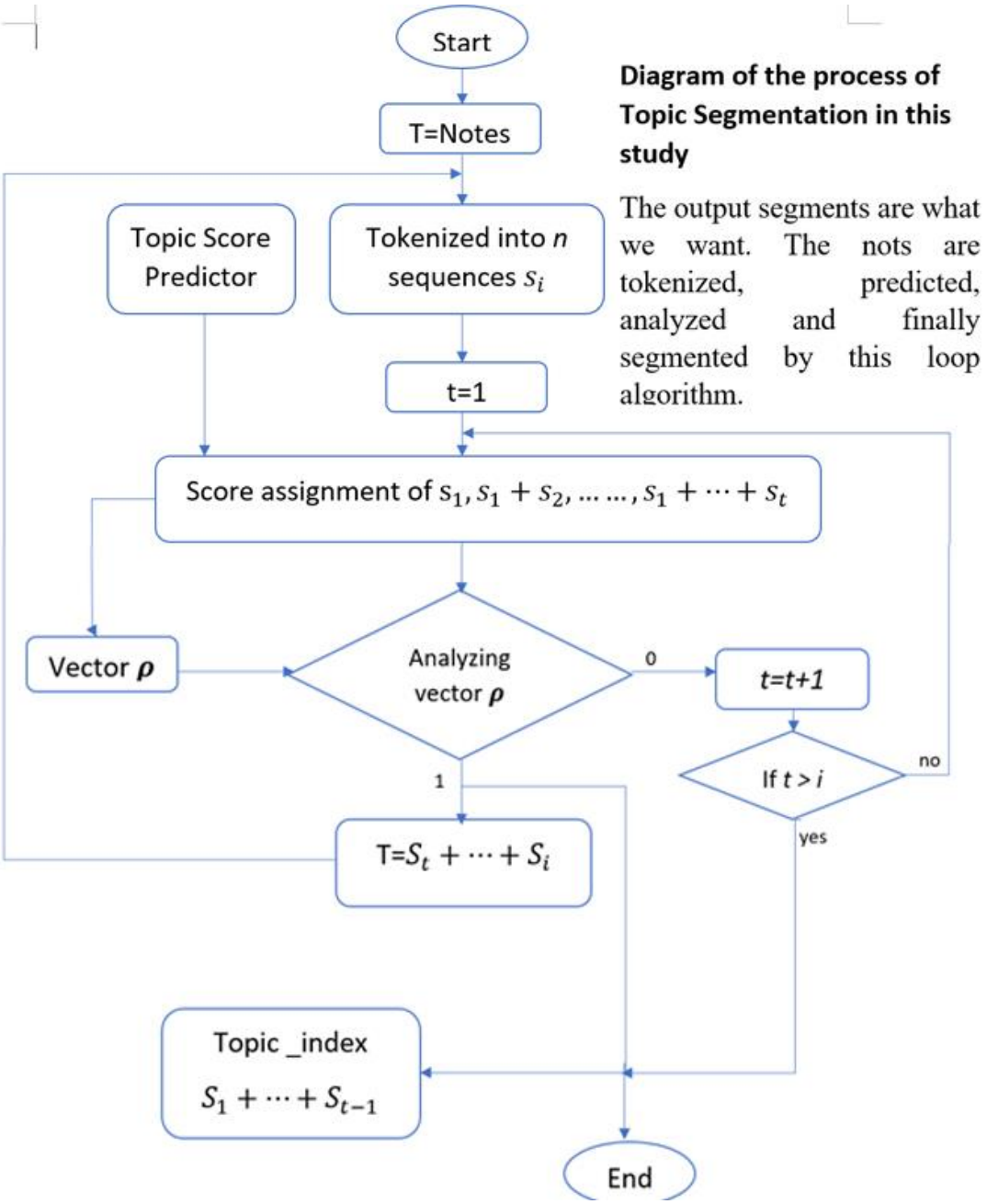

Figure 27: Diagram of the process of Topic Segmentation in this study

Table 8: An example of a part of a clinical note.

History of present illness:

This is a 54-year-old female with a history of cardiomyopathy, hypertension, diabetes type 2, end-stage renal disease on hemodialysis who had a reported cardiac arrest while receiving dialysis on 8/1/06 in her outpatient clinic at Richgo Ster Ha She began to feel nauseated and then vomited. The patient then reportedly went into VFib and was shocked once by EMS, resulting in a narrow QRS complex rhythm. She was intubated, received amiodarone and dopamine, as her BP was approximately 70s systolic over palpable diastolic. In the ED, a portable chest x-ray revealed diffuse bilateral opacities, risk of pulmonary edema and ABG showed respiratory acidosis. Pt was transferred to the ICU for further management. Of note, she was recently hospitalized at Somver Vasky University of Medical Center on 1/5/06 through 11/11/07 for initiation of dialysis after her BUN and creatinine had risen remarkably from baseline. She was then asymptomatic at that time. A fistulogram and angioplasty of her right AV fistula was performed on 9/14/06 with prednisone premedication but it was unsuccessful and therefore a left IJ tunneled dialysis catheter was inserted on 10/18/06 with the tip ending in the right atrium. She has since received dialysis treatments with no complication.

Home Medications:

At the time of admission include amitriptyline 25 mg p.o. bedtime, enteric-coated aspirin 325 mg p.o. daily, enalapril 20 mg p.o. b.i.d., Lasix 200 mg p.o. b.i.d., Losartan 50 mg p.o. daily, Toprol-XL 200 mg p.o. b.i.d., Advair Diskus 250/50 one puff inhaler b.i.d., insulin NPH 50 units q.a.m. subcu and 25 units q.p.m. subcu, insulin lispro 18 units subcu at dinner time, Protonix 40 mg p.o. daily, sevelamer 1200 mg p.o. t.i.d., tramadol 25 mg p.o. q.6 h. p.r.n. pain.

From the Table 8, we could see that this short note has two subtitles, each with different subjects in a structured format. The first part mainly discusses the medical history of the patient's present disease. However, the other main part teaches how to take medicines. Medication terminologies have been used in this part. This note must be tokenized into sequences for the further segmentation process as described earlier in step 3.

The following shows the 23 tokenized sequences from above notes. Each line represents one sequence. Tokenized sequences do not have to be a whole sentence. It could be a short phrase or a long sentence.

1). History of present illness

2). This is a 54-year-old female with a history of cardiomyopathy, hypertension, diabetes……

3). The patient then reportedly went into VFib and was shocked once by EMS …….

4). She was intubated, received amiodarone and dopamine, as her BP ……

5). In the ED, a portable chest x-ray revealed diffuse bilateral ……

6). Pt was transferred to the ICU for further management.

7). Of note, she was recently hospitalized at Somver Vasky University Of ……

8). She was then asymptomatic at that time.

9). A fistulogram and angioplasty of her right AV fistula was performed on……

10). She has since received dialysis treatments with no complication.

---

11). Home medications

12). At the time of admission include amitriptyline 25 mg p.o. bedtime

13). enteric-coated aspirin 325 mg p.o. daily

14). enalapril 20 mg p.o. b.i.d.,

15). Lasix 200 mg p.o. b.i.d.,

16). Losartan 50 mg p.o. daily,

17). Toprol-XL 200 mg p.o. b.i.d.,

18). Advair Diskus 250/50 one puff inhaler b.i.d.,

19). insulin NPH 50 units q.a.m. subcu and 25 units q.p.m. subcu,

20). insulin lispro 18 units subcu at dinner time,

21). Protonix 40 mg p.o. daily,

22). sevelamer 1200 mg p.o. t.i.d.,

23). tramadol 25 mg p.o. q.6 h. p.r.n. pain.

Vector $\boldsymbol{\rho} = [\boldsymbol{as_1}, \boldsymbol{as_2}, \ldots\ldots, \boldsymbol{as_t}]$ contains the information of sequences' accumulated probabilities of belonging to corresponding topics. It is a $5 \times t$ dimensional matrix. The row elements respectively represent corresponding topic score, while the column elements refer to relative sequences. In order to be able to understand the distribution vector $\boldsymbol{\rho}$, the data in vector $\boldsymbol{\rho}$ are plotted in Figure 28 (for illustrating the distribution of probabilities in this example, we plot $t = 23$) from which it can be seen that the probabilities of belonging to each topic all increase with more sequences predicted using NB-based Topic Score Predictor Figure 28 (top), while it shows the probabilities reduce except "medication" which suddenly goes up after 10^{th} sequences using

SVM-based Topic Score Predictor in Figure 28 (bottom). In other words, the probability of medication changes greatly to the point that 10^{th} sequence is present. The reason why nearly all the lines stabilize except for the drug line which increases sharply is that the first to ten consecutive sequences primarily describe disease in the history of patients; however, suddenly the subject changes in "medication" starting at the end of the sequence 11^{st}.

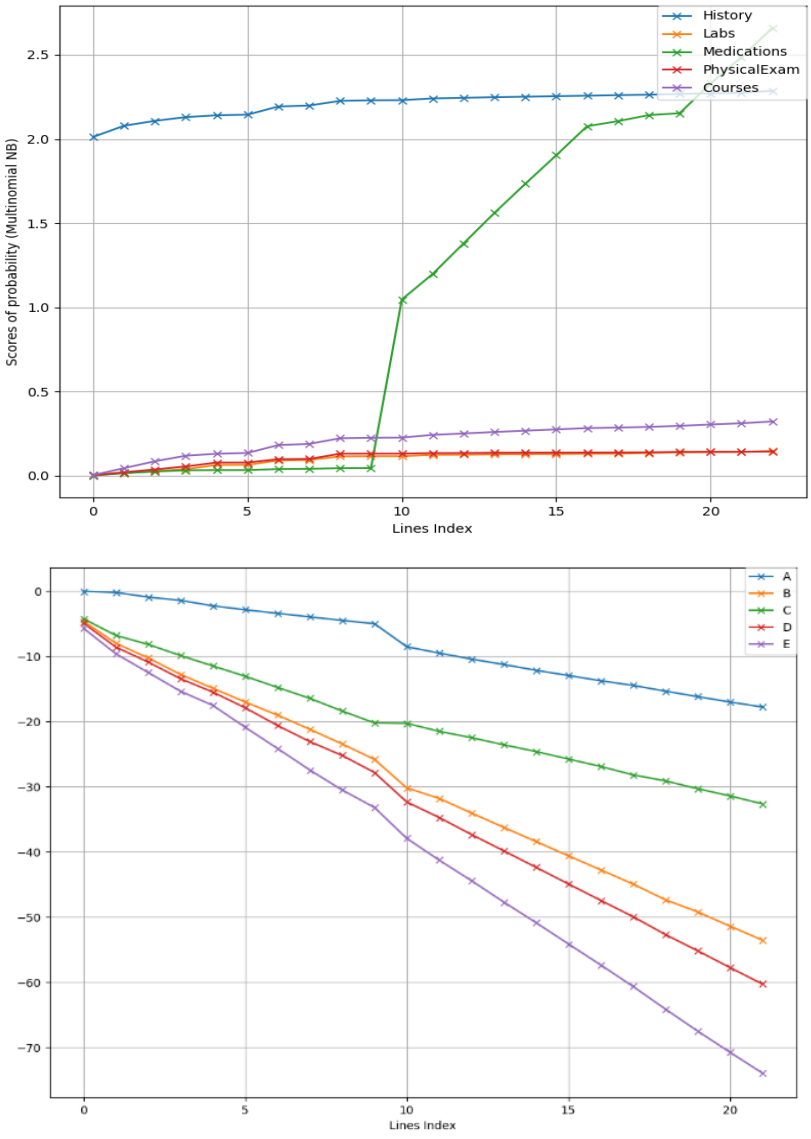

Figure 28: The accumulated score of probability in vector $\boldsymbol{\rho} = [\boldsymbol{as_1}, \boldsymbol{as_2}, \ldots\ldots, \boldsymbol{as_t}]$ *obtained using NB (top) and SVM (bottom) based topic score predictor. Each vector* $\boldsymbol{v}$ *is plotted in vertical axis.*

In order to have a close look at the variety of these five scores, we firstly initialize each accumulated score vector $\mathbf{as_i}$ by being subtracted by the maximum value in the vector:

$$\boldsymbol{as_i'} = max(\boldsymbol{as_i}) - \boldsymbol{as_i}$$

A new vector $\boldsymbol{\rho_{initial}} = [\boldsymbol{as_1'}, \boldsymbol{as_2'}, \ldots\ldots, \boldsymbol{as_t}']$ would be obtained. We then take a backward difference of the former formula. As for the vector $\boldsymbol{\rho_{initial}}$, it could be the following:

$$\boldsymbol{\rho_{initial}}' = [\boldsymbol{as_2}' - \boldsymbol{as_1}', \boldsymbol{as_3}' - \boldsymbol{as_2}', \ldots\ldots, \boldsymbol{as_t}' - \boldsymbol{as_{t-1}}']$$

Vector $\boldsymbol{\rho_{initial}}'$ is plotted on Figure 29 showing that there is a sharp peak at 10^{th} sequence where is the boundary between “history” and “medication”.

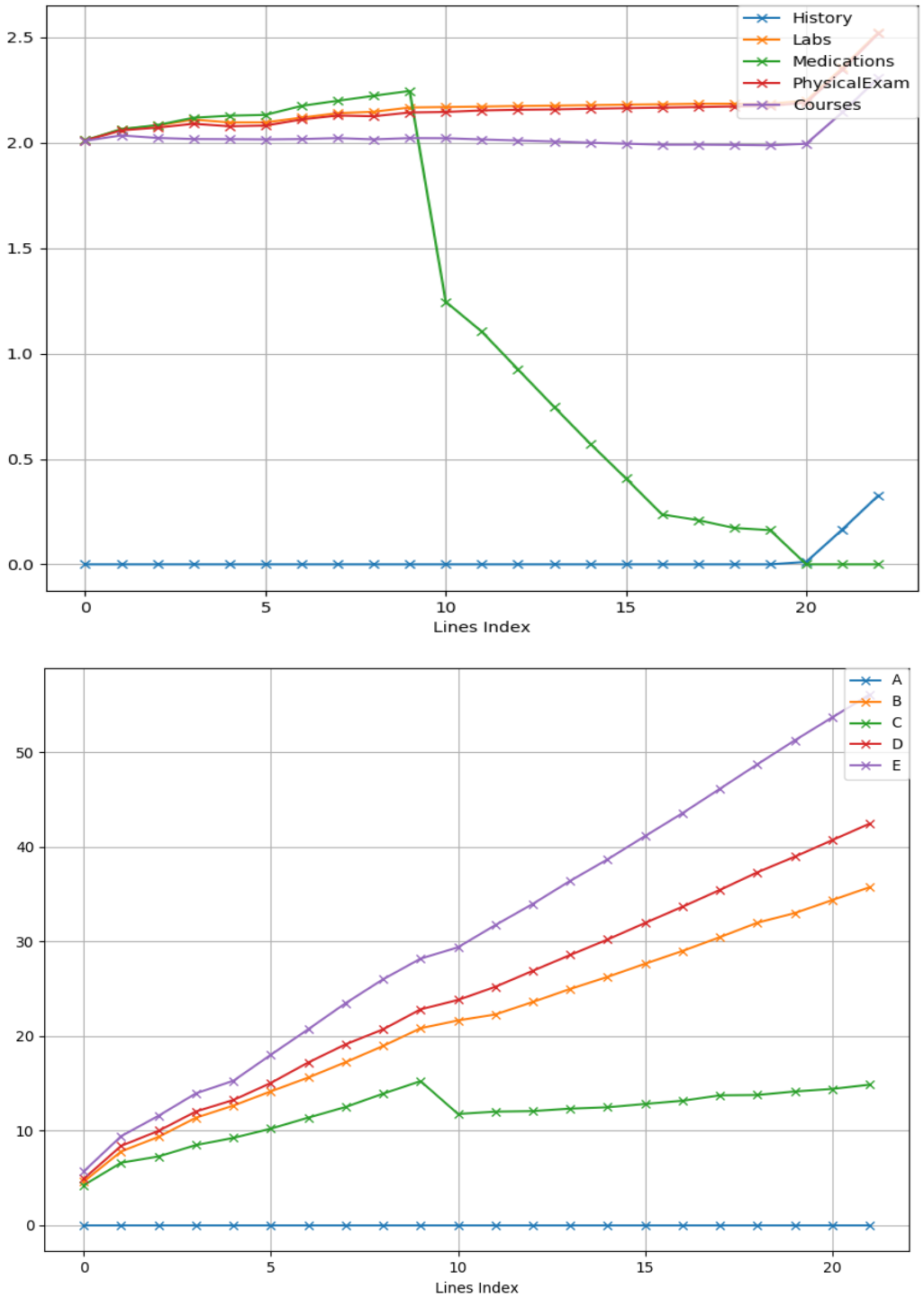

Figure 29: Vector $\boldsymbol{\rho_{initial}} = [\boldsymbol{as_1'}, \boldsymbol{as_2'}, \ldots\ldots, \boldsymbol{as_t'}]$ *obtained by being subtracted by the maximum value in the vector* $\boldsymbol{\rho} = [\boldsymbol{as_1}, \boldsymbol{as_2}, \ldots\ldots, \boldsymbol{as_t}]$. *(top: NB-based, bottom: SVM-based)*

The following is the final step for analyzing the vector $\boldsymbol{as_t}' - \boldsymbol{as_{t-1}}'$:

If $min(\boldsymbol{as_t}' - \boldsymbol{as_{t-1}}') - \text{mean}(sum\ (\boldsymbol{as_t'} - \boldsymbol{as_{t-1}'}) - min(\boldsymbol{as_t}' - \boldsymbol{as_{t-1}}')) > threshold$:

$return \quad 1 \quad and \quad topic_{index} = index(\boldsymbol{as_t}'(0))$

else

$return\ 0$

The best threshold currently tested is 0.3 for NB-based topic score predictor and for SVM-based predictor. It should be noted that the probability of belonging to each class of the samples predicted by topic score predictor is expressed in log-space $P = \log\big(P(S_i)\big)$.

For SVM classifier, after obtaining the signed distance of each data from the boundary in SVM model, we applied Platt scaling for computing the confidence score which represents the probabilities that a given data belongs to a particular class. The Platt scaling is shown below.

$$P\left(\frac{class}{data}\right) = 1/(1 + \exp(A) * f(data) + B)$$

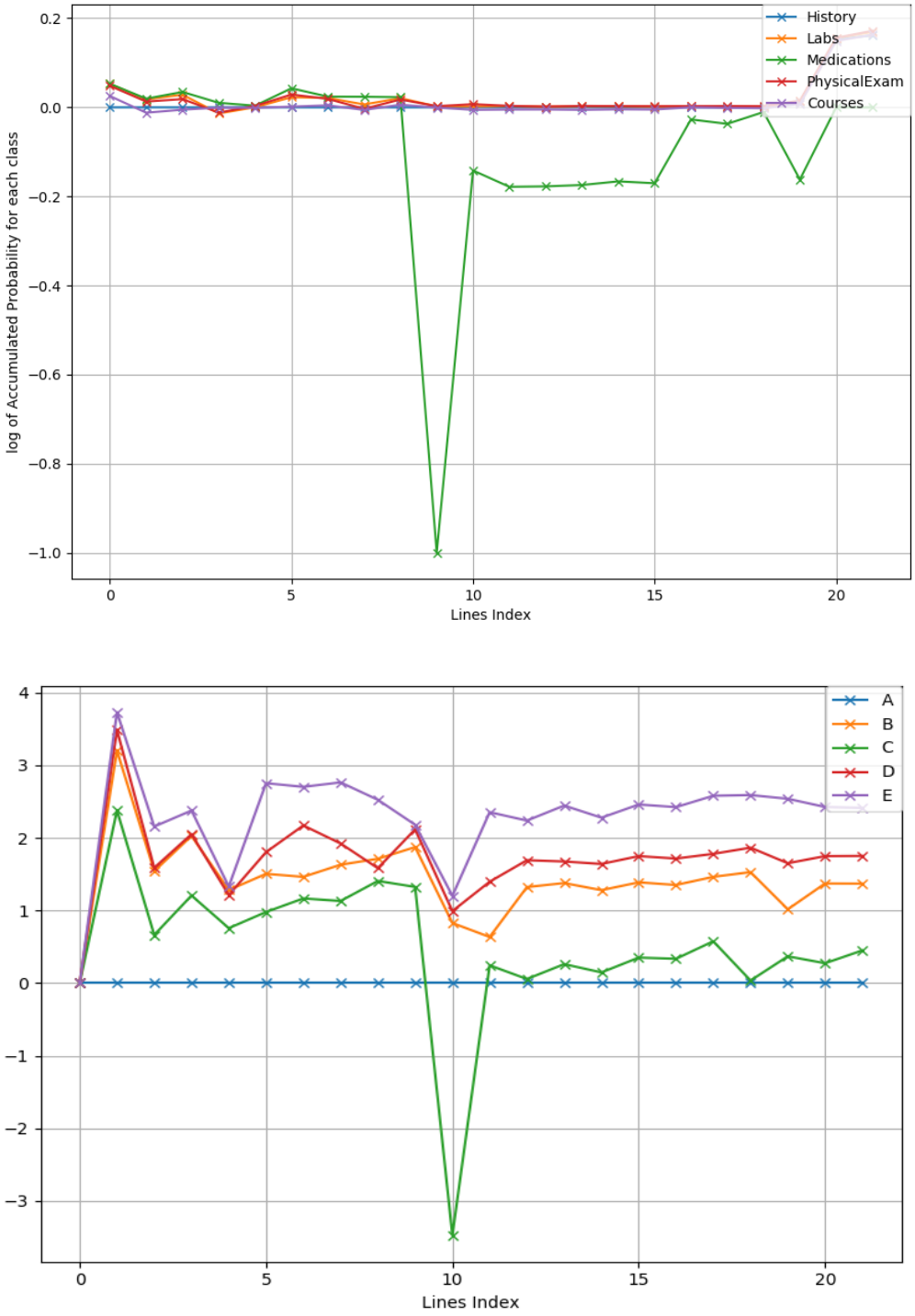

Figure 30: Vector $\boldsymbol{\rho_{initial}}' = [\boldsymbol{as_2'} - \boldsymbol{as_1'}, \boldsymbol{as_3'} - \boldsymbol{as_2'}, \ldots\ldots, \boldsymbol{as_t'} - \boldsymbol{as_{t-1}'}]$ *obtained by taking the backward difference* $\boldsymbol{\rho_{initial}} = [\boldsymbol{as_1'}, \boldsymbol{as_2'}, \ldots\ldots, \boldsymbol{as_t}']$. *(Left: NB-based, right: SVM-based)*

Same as NB scoring model, we convert the confidence score into log space. The accumulated score could be:

$$\mathrm{P}(\mathrm{sequence}_1) + \mathrm{P}(\mathrm{sequence}_2) = \log\big(\mathrm{P}(\mathrm{sequence}_1)\big) + \log(\mathrm{P}(\mathrm{sequence}_2))$$

Since NB considers each word in the document is independent, when NB classifier calculates each sequence probabilities, it should be as following:

$$\begin{aligned} P(sequence_1) &= P(word_1 + word_2 + \cdots + word_t) \\ &= P(word_1) + P(word_2) + \cdots + P(word_t) \end{aligned}$$

For predicting accumulated probabilities, it should be:

$$\begin{aligned} P(sequence_1 + sequence_2) &= P(sequence_1 + sequence_2) \\ &= P(word_1) + P(word_2) + \cdots + P(word_t) + P(word_1') + P(word_2') + \cdots \\ &+ P(word_t') \end{aligned}$$

The two scoring models both could assign score to sequences for boundary detection. However, they have different performance on boundary detection.

Topic Segmentation is an important process of Information Extraction. In this study, we applied Topic Segmentation before Named Entities Extraction for the improvement of accuracy. Chapter 6 will provide details of the performance of both segmenters and the comparison of segmented-dataset and non-segmented-dataset to see if Topic Segmentation could help improve the accuracy of Information Extraction.

Chapter 4. Medical Named Entities Extraction

In last chapter, we present the methodology of Topic Segmentation applied for clinical notes. Topic Segmentation is used to divide the clinical notes into single segments. However, this is not the final goal in this study. This chapter will introduce a methodology of Medical Named Entities Extraction based on Conditional Random Fields model. We first give an overview of the medical entities, such as medical imaging procedure entities and medical entity, to be extracted from clinical notes. After that, the approach is detailed out. The dataset used in this section is the same as that used for training topic segmenter – I2B2 dataset. We finally detail out all the features extracted for training CRF model – word-based feature, POS feature, semantic feature and other feature.

4.1. Overview

Last chapter discussed about Topic Segmentation and introduced the algorithm of Topic Segmentation applied in this study for information extraction. Our final goal of this study is to visualize the patients' medical information from clinical notes on a 2D graph for helping better and instantly understanding patients' past and current situation. Named entities recognition could help extract that crucial information. This chapter would primarily detail out the methodology of named entities extraction from segmented clinical notes.

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 (Ruan & Lee, 2018b). The medical imaging procedure is one of the

non-invasive diagnostic tests (some involve exposure to ionizing radiation), enabling physicians to look inside the body to determine best treatment options for patients. Study (Fazel et al., 2009) and (Dorfman et al., 2011) showed that, despite raising concern about low dose ionizing radiation exposure in the general population, the medical imaging procedure played an important role in the process of diagnosis, rapidly growing use of medical diagnostics imaging methods.

Figure 31 shows how a physician describes an abdominal ultrasound of a patient. The phrase is short. It does contain information, however, on parts of the human body, medical imaging procedures, date taken and results of medical imagery. This information could to a large extent contribute to future diagnosis. For instance:

- 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 in comparison with new images;
- Medical imaging results could help the doctor determine further treatment for patients obviously;

It is therefore essential to extract information from text - free clinical notes from medical imaging procedures.

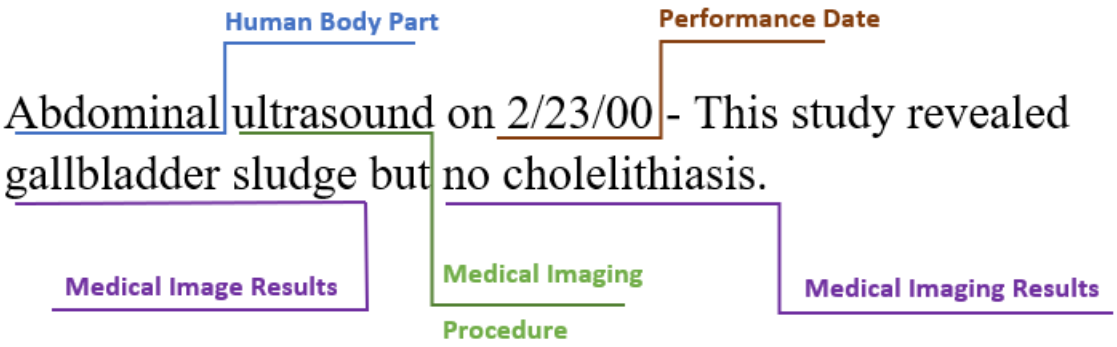

Figure 31: A sample of how a doctor describes a patient's ultrasound information

Other than medical imaging procedures, we also extracted medication entities for further information visualization. Figure 32 shows that doctors usually mention medications name, their quantities, how to take them and frequency for giving instruction. In this study, we consider medications name and their quantities as an entire medication named entity.

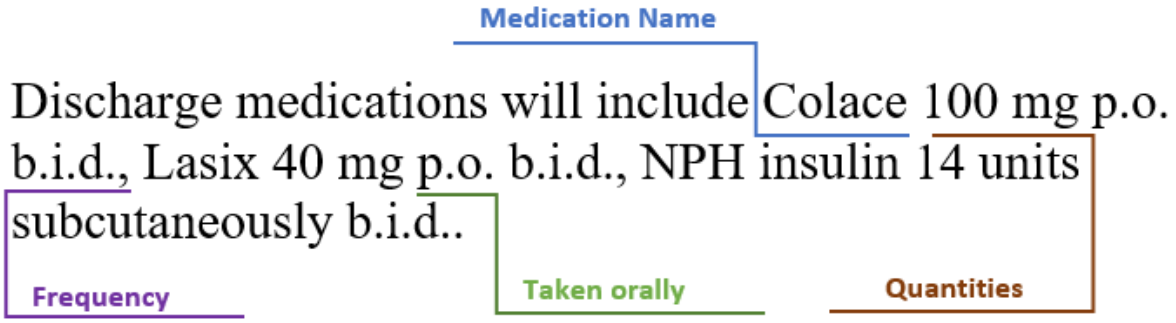

Figure 32:A sample of how a doctor instructs patient taking medications

4.2. Methodology

In order to visualize the information about patients' illness and its location, medical imaging, medication and performed date for our final goal, we proposed a system of extracting and recognizing named entity of medical imaging processing. Our system for recognizing named medical imaging entities is based on CRF. Lafferty, McCallum and Pereira (Lafferty et al., 2001), defined CRF (Conditional Random Fields) as an efficient means of performing sequence segmentation and labeling on discriminatory models. One of the advantages of CRF is that the model is flexible enough to select the feature and does not need to be conditionally independent. We collected and tagged I2B2 data and used 70% of it for training and the rest for testing. The characteristics we extracted for our trained CRF model consist of word-based features, POS features, semantic features and etc. (see B Feature Extraction).

4.2.1. Dataset Used

We sampled 1,100 sentences of I2B2 2009 Medication Challenge Corpus relating to diagnostic procedures in medical imaging. Each sentence token is classified and labeled manually, with tags that match its category (see Figure 33). Table 9 shows all tags and descriptions.

We applied IOB tags format (short for inside-outside-beginning) for encoding medical named entities in this study. This format presented by Ramshaw and Marcus (Ramshaw & Marcus, 1999) is a common tagging format for named entity recognition tagging tokens in a chunking task. There are three types (B-, I-, O) of prefixes using this format. The B-prefix indicates that the tag is the beginning of a chunk, and the I-prefix indicates that the tag is within a chunk. The B-tag is only used when a tag with no O tokens between them is followed by a tag of the same type. An O tag shows that no chunk belongs to a token.

We also extracted 500 sentences associated with I2B2 corpus drug entities for training and testing of medicines called entity recognition to compare the performance of our system with (Kumar et al., 2014; Patrick & Li, 2010).

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. .*

Figure 33: Samples of manually labelled sentences. MRI_MDI means MRI is labelled as MDI. Words without any tag are considered as "No chunk" with tag "O".

4.2.2. Feature Extraction

Metamap is a powerful tool for recognizing in the text the concepts of UMLS (Unified Medical Language System). For the extraction of features, we mainly use the semantic types that are classified by Metamap for each token. The details of the features we extracted are as follows:

- **Word-based features:**

One of the characteristics is the low case of this word. There are other word-based characteristics such as "title or non-title" (whether in the string following non-cash-based letters all case-based characters are upper case, or all other case-based characters are less case), prefix, and the suffix of these words.

Table 9: Categories and tags of our NER system using IOB format. Categories and tags of our NER system using IOB format. (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.)

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

- **POS features**

Words can be considered in the English language as the smallest elements with distinctive meanings. Words are categorized into several types or parts of speech based on their usage and functions. English grammar contains eight major parts of the speech[14]: noun, pronoun, verb, adverb, adjective, conjunction, preposition and interjection (see Figure 34).

- Nouns are the simplest among the 8 parts of speech, such as dogs, cats, birthday… …;
- Pronoun functions as a replacement for a noun. Some examples are: I, it, he, she, mine….;
- Adjectives are commonly used to describe nouns or pronouns and specify the quality, size and number of nouns or pronouns ;
- Verb, the most important part of a speech, shoes an action or state of being of the subject in a sentence. Examples are: is, am, are……;
- Adverbs are used to describe adjectives, verbs or another adverb;
- Prepositions basically refer to words that specify location or a location in time. Such as: above, below, throughout, outside, before and so on;
- Conjunctions join words, phrases or clauses together. Examples are; and, yet, but, for and so on;
- Interjections are used to express and convey emotions;

[14] http://partofspeech.org/

Figure 34: 8 types of part of speech in English grammar. (Source: http://partofspeech.org/)

Stanford POS (Part-Of-Speech) tagger (Toutanova & Manning, 2000; Toutanova, Klein, Manning, & Singer, 2003) is a famous POS tagger for assigning parts of speech to each word. In this study, we applied Stanford POS tagger for the Part-of-Speech extractions of each word.

- **Semantic features**

Metamap is used to identify the current word semantic type and semantic group. For example, "MRI," in the semantic type and in the semantic group respectively, is classified as "Diagnostic Procedures" and tagged as "diap" and "PROC." If there are no correct Metamap tags for this word, the term is tagged as "none" and "NONE." Appendix shows a list of Semantic Types and the abbreviations to their full names in Metamap.

- **Other features**

We also consider whether this word is the first or last string of the sentence. In order to comprehend the current word $X[i]$, and its context, we have removed the $[i-2], X[i-1], x[i+1], x[i+2]$ and its word, semantic and POS features.

Chapter 5. Information Pictorial Visualization

This chapter offers a good demonstration of the mobile-based Pictorial Information Visualization System which is different from the web-based PIVs introduced in Chapter 2. The difference between these two systems would be discussed from two aspects: concepts and interface design, in the section of 5.1. In addition, we eventually come up with an idea of how to integrate our Information Extraction Methodology with the Pictorial Information Visualization System.

5.1. Overview

Medical records are main resources for patients' health history and serve a major part in patient care. The topic of medical records has been discussed in Chapter 2 Literature Review, for more information about EMR please go back Chapter 2.

Information is conveyed effectively when it is represented as an image, rather than a lengthy text. In the field of medicine, for a physician to analyze the medical history of a patient is the most strenuous task. To find and read all the relevant records from past consumes a lot of time and delay in treatment. This is also a reason of long waiting times for other patients. By visualizing this textual data and presenting in an effective way will reduce wait time and eases the process of diagnosing a patient. Pictorial Medical Information Visualization System mentioned in Chapter 2 could visualize the medical information on the categorized interface for helping patients quickly learn about their health condition and the process of diagnosis according to historical records. The previous Pictorial Medical Records developed by (Suo, 2017) is web-based application. This project currently is extended to IOS system (developed by Naveenkumar Appasani) as well. The

next section would present the current progress of design of interface. It can work on the mobile device.

5.2. Current Pictorial Visualization System

5.2.1. Concepts and Plan

Currently, the application of Pictorial Information Visualization System has been developed on from web to mobile system, mainly on IOS system. The idea of developing this mobile application is initiated from the existing web-based EMR system, developed by (Suo, 2017). However, there is a slight difference between the web-version system and IOS-based system. In the existing web-based system, there are 11 physiological systems. A team involving three medical students worked on categorizing the 11 systems into 6. Table 10 and Figure 35 show how the physiological systems are merged.

Table 10: Classification of Physiological system of Web-based and Mobile-based Version

Web-based Version	Mobile-based Version
Skeletal	MusculoSkeletal/Deramatological
Muscular	
Skin	
Cardiovascular	Cardiovascular/Respiratory
Respiratory	
Digestive	Gastrointestinal
Nervous	Nervous
Immune	Immune/Endocrine
Endocrine	
Reproductive	Genitourinary
Urinary	

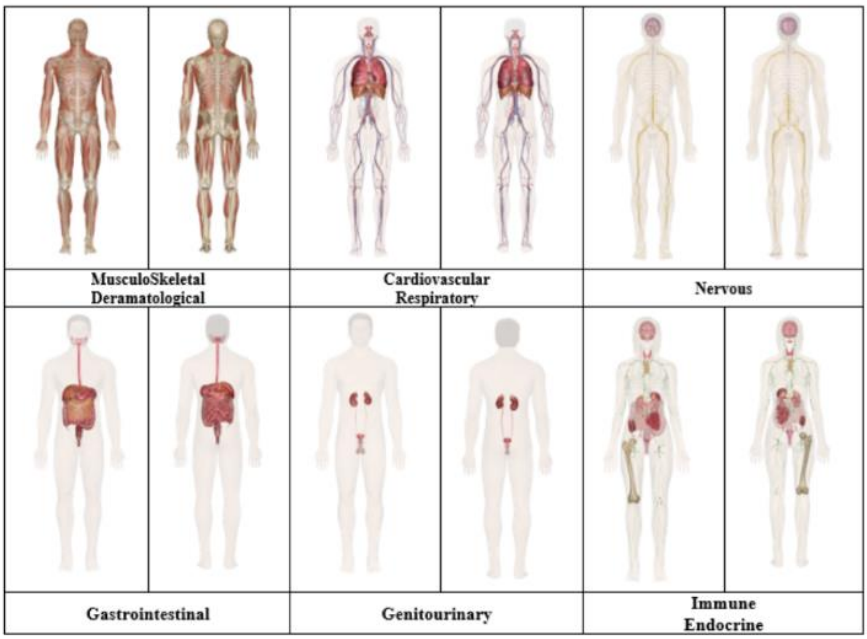

Figure 35: Display of integrated system and physiological systems in front and back views. Musculoskeletal/Dermatological, Cardiovascular/Respiratory, Gastrointestinal, Nervous, Immune/Endocrine and Genitourinary. (Ruan et al., 2018)

Figure 36 demonstrates an example of cardiovascular image and respiratory images from web-based system, which are merged to generate a new updated system representation for mobile-version system.

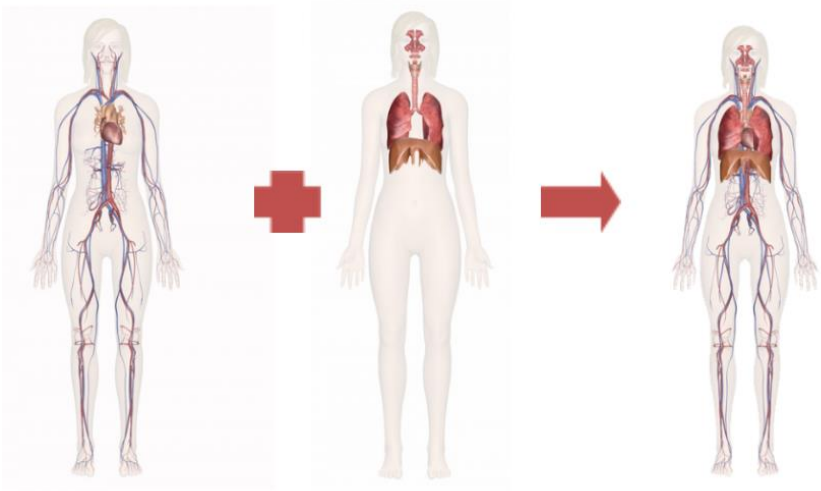

Figure 36: How two different physiological system merged (Source:Naveenkumar)

Meanwhile, the mobile-version application has five clinical events for temporal visualization which are: Imaging: Scanning, X-Ray, MRI etc., Medication: Drug Information, Vaccines etc., Visit: Appointment, initial notes, Lab test: Blood test, Urine test etc., Treatment: Physiotherapy, Operation etc.

5.2.2. Interface Design

The interface of the mobile-version system is built by using the following tools shown on Figure 37 and Figure 38:

- Sketch - a proprietary vector graphics editor for Apple's macOS, developed by the Dutch company Bohemian Coding.
- XCode 9 - Integrated macOS development environment (IDE) with Apple software development tools for the development of macOS, iOS, watchOS, and tvOS software.
- Swift 4 - a compiled programming language for iOS, macOS, watchOS, TVOS, and Linux developed by Apple Inc.
- Prepo – Different size Icon generator application for iOS
- MySQL – Database engine used to create backend for the application

Figure 37: The tools used to build mobile-version pictorial medical information visualization systems

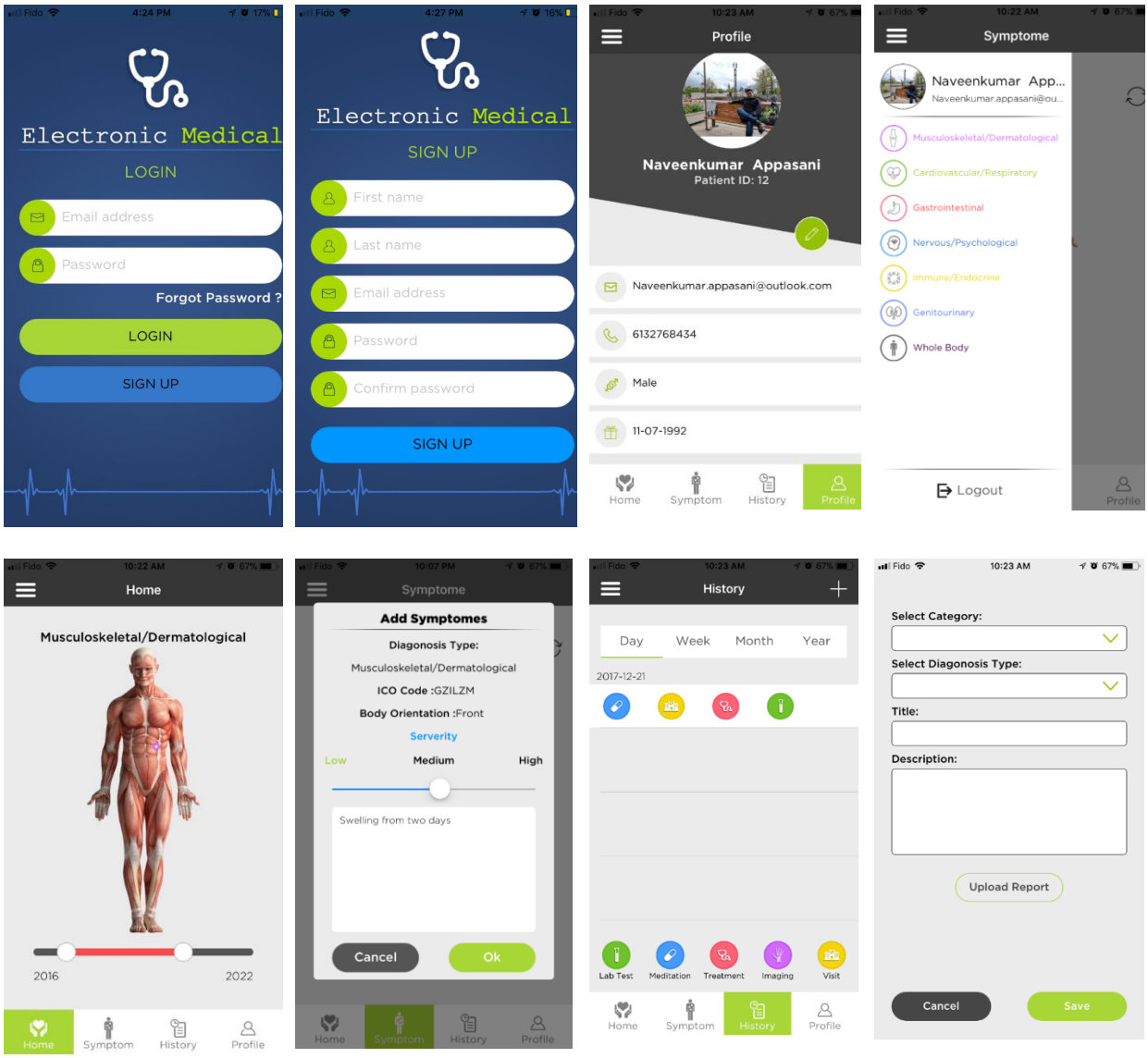

Figure 38: Interface of mobile-version system on IOS system- Interfaces of login, signup, profile, physiological system, temporal system and symptoms report interface.(Source: Naveenkumar)

5.3. Experiment of Integrating with Pictorial Visualization System

Our idea of integrating the information extraction system in this study with the Pictorial Visualization system is based on building database of the coordinates of physiological images and their corresponding body part common names and medical terminologies. For displaying the

extracted information on the physiological image, the first thing is to find the location to show the data. Human body parts entities could provide the medical terminologies or the common names which instruct us the location of the symptoms. Usually, one location might have multiple human body part names. The following Figure 39 illustrates a simple example. According to the coordinate (5,11) (see Table 11), there are at least 4 different keywords such as: Left Lung, Left Chest, 5^{th} Rib and Left Upper Back respectively corresponding to Cardiovascular/ Respiratory, Integration, Skeleton and Muscular systems. Based on this correspondence, a small database could be built:

Table 11: An example of the coordinate (5,11) and its corresponding human body-parts common name or medical terms.

Coordinates	*Keywords*	*Physiological System*
(5,11)	*Left Lung*	*Cardiovascular/ Respiratory*
	Left Chest	*Integration*
	5^{th} Rib	*MusculoSkeletal/Deramatological*
	Left Upper Back	*MusculoSkeletal/Dermatological*

The use of keywords to describe the symptoms might indicate the symptoms. For example: when “left lung” is used in the clinical notes, the patients might have breathing problem; When “rib” is used, that probably means the patients has fracture. This correspondence, however, is not absolute. As long as the entities have the keyword, the coordinate could be mapped.

For temporal visualization, the date entities could definitely help to visualize the when the imaging procedure has been performed and what the medication the patients have taken.

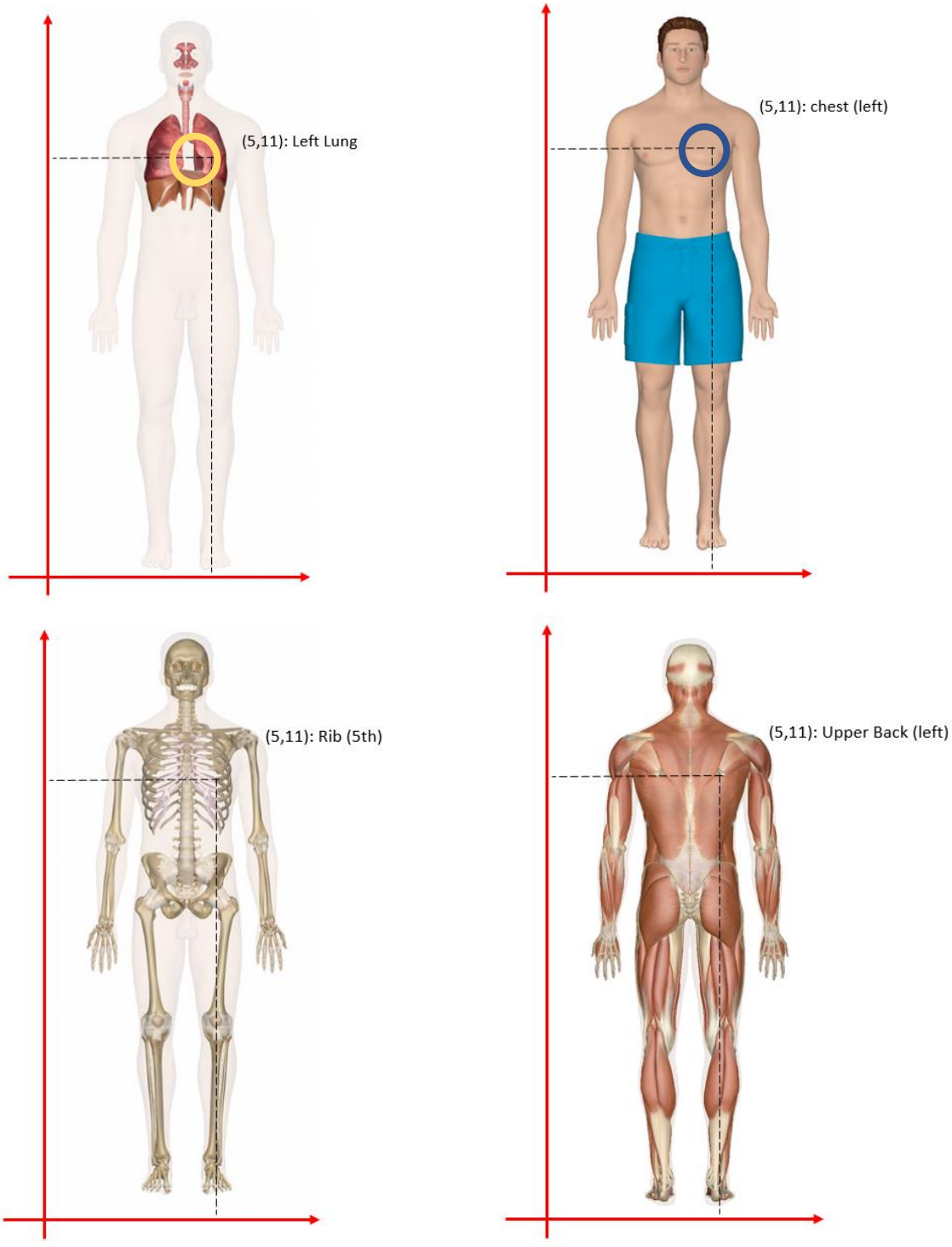

Figure 39: An example of a coordinate on different physiological image with its corresponding keywords

Chapter 6. Results and Evaluation

Evaluation is the most important part of a research study. This chapter would introduce how we assess our information extraction systems (Topic Segmentation and Named Entities Extraction) and how the systems perform. Almost a half of the pages in this chapter talk about the evaluation metrics, especially for Topic Segmentation metrics: traditional metrics (Windiff and P_k) and Boundary-Edit-Distance-Based metrics (Boundary Similarity and Segmentation Similarity). Precision recall and F-score, famous metrics in the domain of Information Extraction, are also mentioned in the corresponding section, At the end of this chapter, we applied the our NER system to extract medication entities from two different datasets: "segmented dataset" and "original dataset". The results show that Topic Segmentation could enhance the accuracy for information extraction.

6.1. Evaluation Metrics

This section will discuss the well-known evaluation metric for Topic Segmentation and Information Extraction. For assessing Topic Segmentation algorithms, the famous traditional metrics are P_k and Windiff, proposed by Beeferman et al. (Beeferman et al., 1999) and Pevzner et al. (Pevzner & Hearst, 2002) respectively. P_k, a window-based metric, attempts to solve the harsh near-miss penalization of precision, recall and F-score. The subscript k represents the window size, a half of the mean reference segment size. Another traditional metric is Windiff which is able to overcome the shortcomings in P_k. In addition, two boundary-edit-distance-based metrics: segmentation similarity and boundary similarity, are mentioned. Finally, we also detail out the famous metric-precision, recall and F-score for evaluating information retrieve and classification.

6.1.1. P_k

Segmentation of the topic has two criteria to consider: precision of segmentation and accuracy of identification of the topic. The popular metrics of assessing the quality of classification algorithms are precision and recall statistics. Precision measures the fraction of boundaries identified by an automatic segmenter which are actual boundaries for the segmentation task, while recall measures the fraction of actual boundaries correctly identified by an automatic segmenter. However, there are some shortcomings of the precision and recall metrics. Hypothesizing more boundaries increases recall at the expense of precision and recall. Consequently, the algorithm designers might tweak parameters to trade between the two in a way that matches that demands of the applications. There is one compromise approach between precision. That would be F-measure, a weighted combination of precision and recall. However, F-measure is difficult to interpret as a meaningful performance measure. Beeferman et al. (Beeferman et al., 1999) proposed an approach to evaluate a segmentation with redefining correct to mean "hypothesized within some constant-sized window of units away from a reference boundary" (see Figure 40).

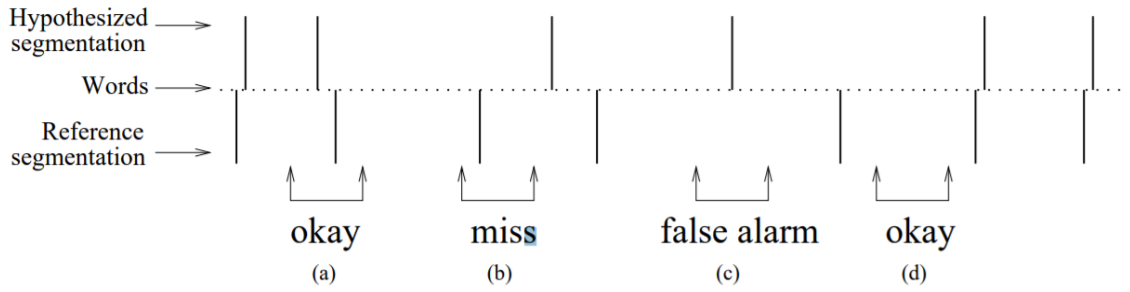

Figure 40: Failure modes of a decision procedure for segmentation. The lower vertical lines represent "true" segment breaks and hypothesized breaks are represented by the upper vertical lines. A fixed - width window slid across the corpus results in an acceptable (a and d) result in both the present and the absent hypothesized break ; a false negative (b), where there is a true break but not a hypothesized break ; and the misleading alert (c), where there is a hypothesized break, but not an true break (Beeferman et al., 1999).

The measure P_k the proposed is a probability and thus a real number between zero and one. The use the following formula:

$$p(error|\boldsymbol{ref}, \boldsymbol{hyp}, k)$$
$$= p(miss|\boldsymbol{ref}, \boldsymbol{hyp}, different\ \boldsymbol{ref}\ egments, k)p(different\ \boldsymbol{ref}\ segments\ |\ \boldsymbol{ref}, k)$$
$$+ p(false\ alarm|\boldsymbol{ref}, \boldsymbol{hyp}, same\ \boldsymbol{ref}\ segment, k)p(same\ \boldsymbol{ref}\ segment\ |\boldsymbol{ref}, k)$$

where ***k*** is a distance of ***k*** words;

This probability of measurement can be broken down into conditional probabilities, called miss and false alarm probabilities, which give a more detailed look at the error, allowing for accuracy and recall assessment.

6.1.2. WindowDiff

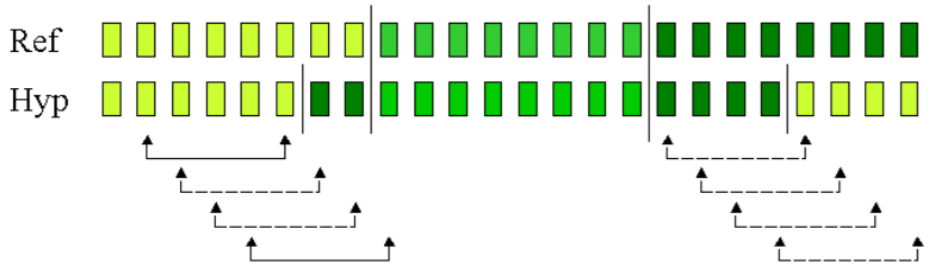

Figure 41: An example of how the metric P_k handles false positives. Boxes indicate phrases or other units of a subsection; and four are the width of the window (k), which means that four possible boundaries fall between the two ends of the probe. The lines indicate both poles of the probe as they move from right to left. No penalty is imposed on solid lines, dashed lines indicate a penalty is imposed. For false negatives, the total penalty is always k. The total penalty depends on the distance between the false positive and the correct borders. On average it is k/2 provided the boundaries are divided across the document in a uniform manner. (Pevzner & Hearst, 2002)

The P_k metric evaluation was the standard measure for evaluating algorithms for text segmentation for a long time. There are some issues, though. First, there is less penalization of false positives than of false negatives. Suppose a text has segments of average size $2k$ (k refers to the distance between the P_k probe's two ends). It the segmenter produces a false negative, it receives *k* penalties using P_k metric. (See Figure 41)

Secondly, P_k metric allows some errors to go unpenalized. Specifically, the number if segment boundaries between the two ends of the probe are not taken account by metric P_k. (see Figure 42).

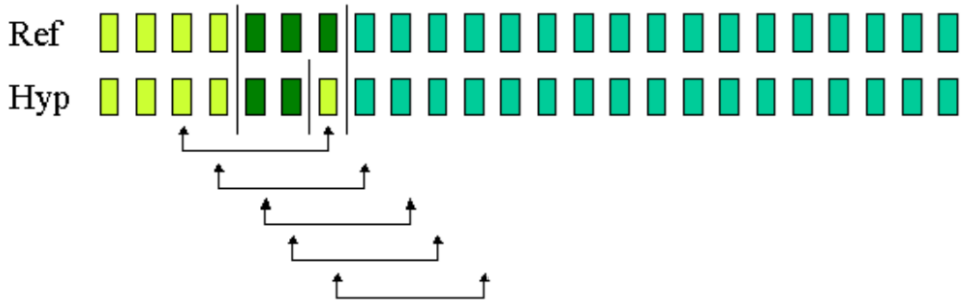

Figure 42: An illustration of the fact that the P_k metric fails to penalize false positives that fall within k (Pevzner & Hearst, 2002)

In addition, the metric P_k is sensitive to variations in segment size. The false negative (missing boundaries) penalty changes as the segment size becomes smaller. Errors in smaller-than-average segments significantly decrease the penalty for false positives and false negatives, while errors in larger-than-average segments slightly or completely increase the penalty.

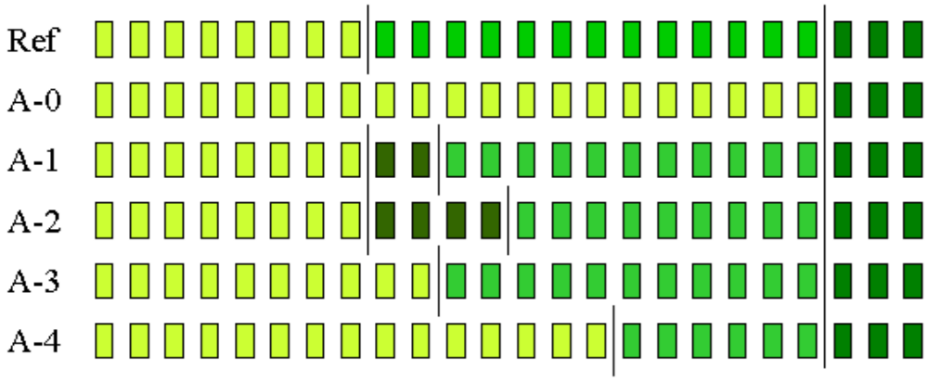

Figure 43: A reference segmentation and five different hypothesized segmentations with different properties (Pevzner & Hearst, 2002)

Another problem with metric P_k is that near-miss error is penalized too much. Figure 43 shows five different segmentation algorithms with a mistake on the boundary with the first and

second segment. Algorithms A-0 and A-2 respectively contain a pure false negative and false positive. However, algorithm A-4 should be considered as the worst segmentation because it has simultaneously a false positive and false negative. A-3 should be better in comparison with the A-1 algorithm, as it detects only one boundary rather than two, and A-1 has a supplementary boundary. Therefore, the algorithm A-3 is better than the algorithm A-1. It'd be different to use metric P_k. Assume both algorithms A-1 and A-3 both contain a small distance from the actual one, which is incorrect. Then the penalty for A-1 is e, while for A-3 the penalty for A-3 is 2e. That must not be the case; algorithm A-1 is better than algorithm A-3, as it is nicer to get a close mistake than a pure false positive, even if it's close to the limit.

For these reasons, Pevzner and Hearst (Pevzner & Hearst, 2002) proposed a new segmentation evaluation metric called WindowDiff that moves a fixed - size window across the text and punishes the algorithm if the number of boundaries within the window does not match the true number of boundaries for that text window. Formally,

$$WindowDiff(ref, hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0)$$

Where $b(i,j)$ refers to the number of boundaries between positions i and j in the text and N means the number of sentences in the text.

The asymmetry between the false negative and false positive penalties in the metric P_k are clearly eliminated by this methodology of assessing segmentation algorithm. Moreover, within segments of length less than *k*, Windiff also catches false positives and false negatives.

6.1.3. Boundary-Edit-Distance-based Metrics

6.1.3.1. Segmentation Similarity

A segmentation evaluation metric, called segmentation similarity proposed by Fournier and Inkpen (Fournier & Inkpen, 2012), is defined as “quantifies the similarity between two segmentations as the proportion of boundaries that are not transformed when comparing them

using edit distance, essentially using edit distance as a penalty function and scaling penalties by segmentation size". In this metric, the entire segmentation is conceptualized as having mass (i.e., unit magnitude/length). The Figure 44 shows the concept of annotation of segmentation mass.

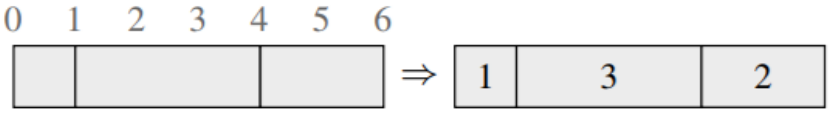

Figure 44: Annotation of segmentation mass (Fournier & Inkpen, 2012)

After the segmentation mass has been annotated, the parallel sequence of the boundary sets is then converted, each one of which has its own boundary type. If there is no boundary, the set is empty. (see Figure 45)

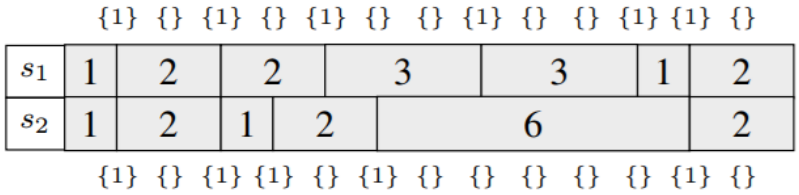

Figure 45: Segmentations annotated with mass and their corresponding boundary set sequences (Fournier & Inkpen, 2012)

Segmentation Similarity uses the following equation:

$$\mathrm{S}(s_{i1}, s_{i2}) = \frac{\mathbf{t} \cdot \mathrm{mass}(i) - \mathbf{t} - \mathrm{d}(s_{i1}, s_{i2}, T)}{\mathbf{t} \cdot \mathrm{mass}(i) - \mathbf{t}}$$

As shown on Figure 46, the numerator is normalized by the total number of potential boundaries per boundary type. The results in a function with a range of [0,1]. It returns 0 when one segmentation contains no boundaries, and the other contains the maximum number of possible boundaries. Oppositely, it returns 1 when both segmentations are identical (Fournier & Inkpen, 2012).

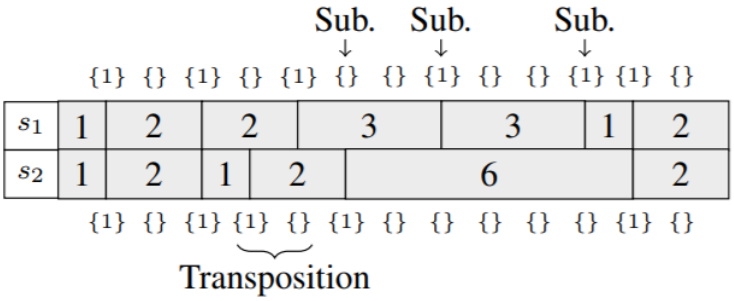

Figure 46: Edit operations performed on boundary sets (Fournier & Inkpen, 2012)

6.1.3.2. Boundary Similarity

Fournier (Fournier, 2013) also proposed another segmentation evaluation metric, named boundary similarity, based on boundary edit distance. In this work, Boundary Edit Distance has three main edit operations: 1) additions/deletions (AD) for full misses; 2) substitutions (S) for confusing one boundary type with another; 3) n-wise transpositions (T) for near misses; See Figure 47 below:

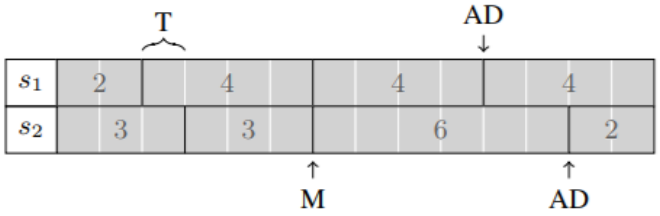

Figure 47: Boundary edit operations (Fournier, 2013)

The boundary similarity is defined as shown in the following equation:

$$B(s_1, s_2, n_t) = 1 - \frac{|A_e| + w_t\_span(T_e, n_t) + w_s\_ord(S_e, T_b)}{|A_e| + |T_e| + |S_e| + |B_M|}$$

where:

A_e is the set of additions/deletions;

T_e represents the set of n-wise transpositions;

S_e is the set of substitutions;

B_M refers to the set of matching boundary pairs;

6.1.4. Precision Recall and F-Score

Precision and recall are the most important metrics for evaluating the algorithms of information retrieval, pattern recognition and binary classification. Precision is a fraction of the relevant instances between the instances selected, and recall is the fraction that was selected for the total number of relevant instances. According to the following Table 12, precision and recall are defined as:

$$\text{Precision} = \frac{true\ positive}{true\ positive + fasle\ positive}$$

$$\text{Recall} = \frac{true\ positive}{true\ positive + false\ negative}$$

The best situation is that precision and recall are both close to 1, which means that the algorithm is the best. However, Precision and recall are usually contradictory. More specifically, if precision is 100%, recall would be low. In contrast, precision would be low if recall is very high. The best solution to avoid this problem is using F-measure, also named as F-score:

$$\text{F} = \frac{(\alpha^2 + 1)precision * recall}{\alpha^2(precision + recall)}$$

When $\alpha = 1$, F-score would be the well-known F_1 score.

Table 12: Confusion Matrix15 with true positive, false positive, false negative and true negative (source: Wikipedia)

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

6.2. Topic Segmentation Results

Topic Segmentation has two tasks: Text Segmentation and Topic Identification. We applied the evaluation metrics P_k, Windiff, F score, Boundary and Segmentation Similarity for comparing the NB-based segmentation algorithm and SVM-based algorithm. In addition, comparing topic identification accuracy is also crucial for evaluation. We refer to the approach of evaluating text classification to evaluate whether the algorithms have a good performance on topic recognition. Accuracy is calculated by counting the number of tokenized sentences whose reference label does not agree to hypothesized label, as shown in Figure 48.

[15] https://en.wikipedia.org/wiki/Precision_and_recall

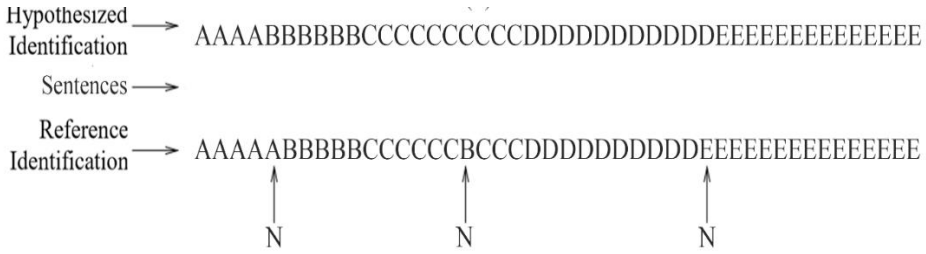

Figure 48: $\boldsymbol{F_1}$ score method of topic identification evaluation

Table 13 shows the results tested on 100 clinical notes collected from I2B2. "NB-based segmenter" refers that the score vectors $\boldsymbol{v}$ assigned to each tokenized sentence is predicted by the Topic Score Predictor which is trained using Multinomial Naïve Bayes algorithm. Similarly, "SVM-based segmenter" means the Topic Score predictor is built with the model of Support Vector Machine. The results of Windiff and P_k with range of [0,1] shows the performance of boundary detection. The result closer to 0 means the reference boundary and predicted boundary is more similar. The F-1 score, however, shows the accuracy of classification with the range of [0,1]. The higher the score is, the better the performance is.

Table 13: The results of NB-based and SVM-based segmenter assessed using Windiff, $\boldsymbol{P_k}$, Boundary Similarity, Segmentation Similarity and $\boldsymbol{F_1}$ score. The value of Windiff and $\boldsymbol{P_k}$ is lower, which means the segmenter performs better. Results show that NB-based segmenter has a better performance.

Metrics / Segmenter	Windiff	P_k	BSimilarity	SSimilarity	F_1
NB-based	**0.1121**	**0.0910**	**0.6531**	**0.9716**	**0.9181**
SVM-based	0.3084	0.2227	0.4010	0.9369	0.6669

From the above table, we can see that NB-based segmenter wins at no matter which evaluation metric is used. NB-based segmenter obtains 0.1121 and 0.0910 on Windiff and P_k, while the SVM-based segmenter performs worse with 0.3084 Windiff and 0.227 P_k. For F_1 score,

these two segmenters has approximately 24% difference. In other words, F_1 score of SVM-based segmenter is 0.24 lower than that of NB-based segmenter.

Figure 49 illustrates the F_1 score measured on 50 clinical notes using NB-based and SVM-based segmenters each of which is represented with the line of colour of blue and orange respectively. These two lines have a similar trend. However, the blue one is almost always above the orange line, although there are overlaps at some points (for example at the point 39 and 40).

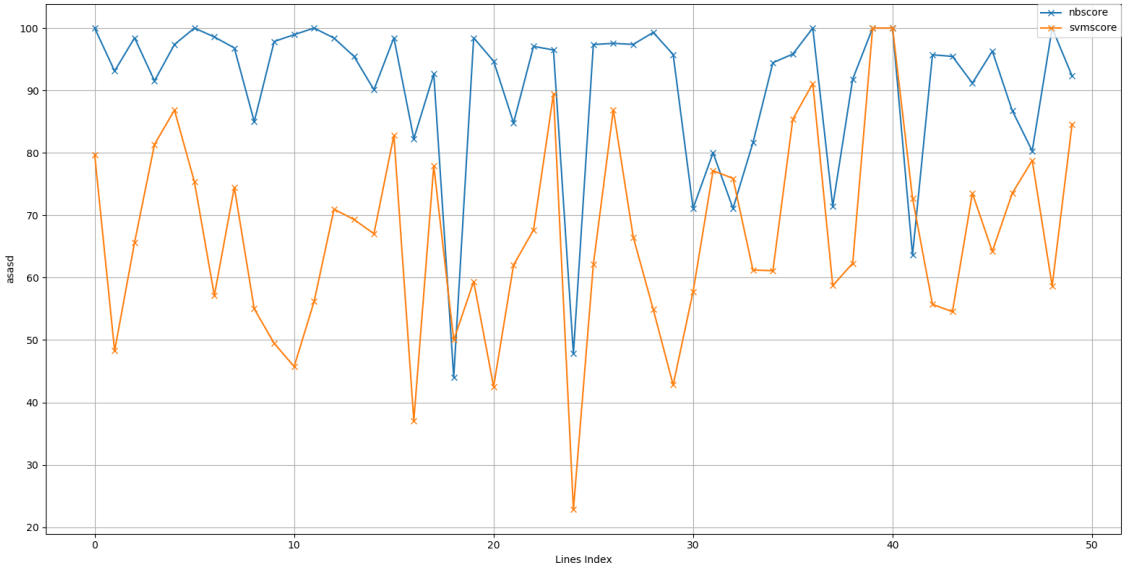

Figure 49: F-1 score measured on 50 clinical notes; the blue coloured NB-based method shows higher . F_1 score than the orange coloured SVM-based method for most clinical notes.

Let us focus on a single clinical note and see the difference between these two segmenters' performance. Figure 50 shows the three specific segmentations of a clinical notes: reference segmentation, NB-based segmentation and SVM-based segmentation. Different colours are used to highlight different segments. Blue, red, green, yellow and purple coloured the part of history, physical exams, medications, labs and hospital course respectively. Obviously, the middle one is much closer to the reference segmentation than the third one. From the third figure we can see that some parts of the physical exams part are mis-labelled as medications and labs. Similarly, about a

quarter of hospital course is wrongly classified as medications. This is another evidence that NB-based segmenter performs better than SVM-based segmenter.

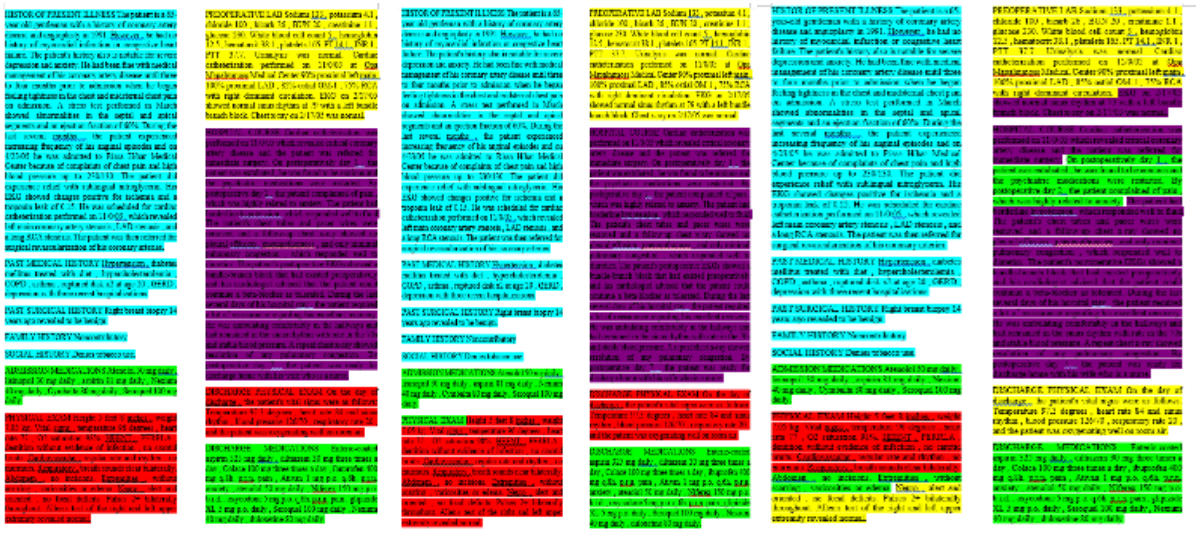

Figure 50: Reference Segmentation (left) and Hypothesized Segmentation (NB-based: middle, SVM-based: right); Blue, red, green, yellow and purple coloured the part of history, physical exams, medications, labs and hospital course respectively. Intuitively, the middle one is closer to the left one, which means NB-based segmenter performs better.

6.3. Medical Named Entities Extraction Results

Our medical named entities extraction 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 $\boldsymbol{F_1}$ score metrics are used to evaluate our model.

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

The results (Table 14) show that our named entities extraction system has different performance on different entities extraction. For recognizing "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 14: The results of Precision, Recall and F-1 score of each Named Entity based on per token

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

However, evaluating NER models should consider entity level beside tokens level. Figure 51 shows our modified method of counting the correct number of matches.

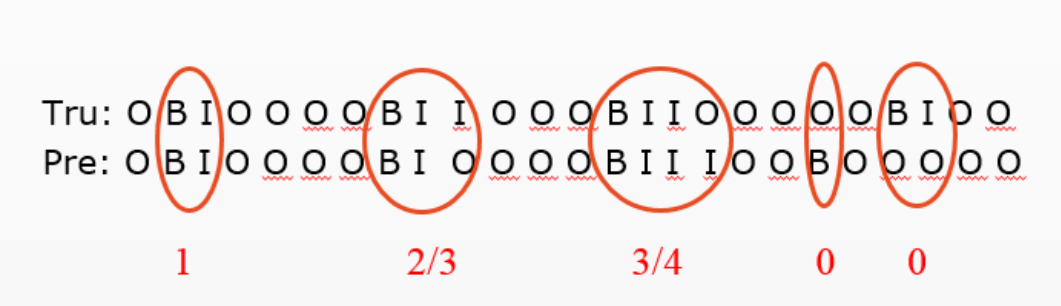

Figure 51. Our method of comparing reference sequence and predicted sequence for counting the correct number of labeling

Table 15 illustrates the results of each named entities extraction and shows that entities MDI and HMB have acceptable accuracy with 92.3% and 77.6% respectively.

Table 15: The results of Precision, Recall and F-1 score of each Named Entity based entities

Entities	Precision	Recall	F-1 Score
O	0.867	0.953	0.913
MDI	0.907	0.94	0.923
HMB	0.846	0.716	0.776
MID	0.914	0.632	0.747
MIR	0.784	0.575	0.663
ST	0.545	0.298	0.385

The International Workshop on Semantic Evaluation, also known as SemEval for short, introduced four different ways to measure precision, recall and F1-score results based on the metrics defined by MUC. They are four ways as follows:

Strict: exact boundary surface string match and entity type;

Exact: exact boundary match over the surface string, regardless of the type;

Partial: partial boundary match over the surface string, regardless of the type;

Type: some overlap between the system tagged entity and the gold annotation is required;

As shown in Table 15, the following table (

Table 16) also shows that entities HMB and MID obtain satisfied results (93.7% and 95.3% on Partial respectively) measured by SemEval. The entity ST, however, has a lower accuracy with 50.4%, 61.7%, 59.6% and 46.1% on Type, Partial, Exact and Strict respectively. From Table 16 it is clear to see that our Named Entity Recognition System has a good performance on partial entities extraction while on exact entities extraction it needs improvement. In other words, the system could accurately recognize the entities but not all tokens could be extracted.

Table 16: The results of Precision, Recall and F-1 score of each Named Entity evaluated based on type, partial, exact and strict.

Entities	Measure	Type	Partial	Exact	Strict
HMB	Precision	0.845	0.941	0.947	0.841
	Recall	0.833	0.933	0.935	0.834
	F1	0.839	0.937	0.941	0.837
MID	Precision	0.81	0.961	0.952	0.802
	Recall	0.798	0.946	0.937	0.791
	F1	0.804	0.953	0.944	0.796
MIR	Precision	0.731	0.854	0.823	0.698
	Recall	0.705	0.842	0.795	0.678
	F1	0.718	0.848	0.809	0.688
ST	Precision	0.513	0.623	0.601	0.490
	Recall	0.495	0.612	0.591	0.435
	F1	0.504	0.617	0.596	0.461

Table 17 shows the performance of our system on Medication Entities Extraction tested using SegEval's four different ways. The results show that our NER system also performs well on Medication Entities Extraction with 92.9%, 88.9%, 85.1% and 85.1% on Type, Partial, Exact and Strict respectively.

Table 17: The results of Medication Entities Extraction evaluated on type, partial, exact and strict.

Measure	Type	Partial	Exact	Strict
Precision	0.952	0.911	0.871	0.871
Recall	0.908	0.869	0.831	0.831
F1-Score	0.929	0.889	0.851	0.851

Table 18 compares the performances of our system on medication entities extraction with that of other researches. Obviously, our system wins on both measurement with 88.9% on Partial and 85.1% on Exact.

Table 18: The results of Medication Entities Extraction comparing with KUMAR'S, PATRICK'S AND WANG'S.

	Ours (Partial)	**Ours (Exact)**	**Kumar's**	**Patrick's**	**Wang.Y's**
Precision	0.911	0.871	0.913	0.896	0.833
Recall	0.869	0.831	0.708	0.814	0.768
F_1 Score	**0.889**	**0.851**	0.798	0.849	0.799

6.4. Is segmentation useful?

For evaluating our system and Topic Segmentation idea, we tested our Named Entity Recognition model with Topic Segmentation on segmented dataset and dataset without segmentation to see if the Topic Segmentation is necessary for information extraction. We first applied our Topic Segmentation algorithm on 50 clinical notes and obtained 50 segmented clinical notes. Our named entity recognition model is then used to extracted medication entity on both datasets- original dataset and segmented dataset. When extracting from the segmented dataset, the NER model goes straight to "Medication" parts for medication entity recognition. Table 19 shows the results:

Table 19: The results of Medication Entities Extraction using our NER system tested on Original Dataset and Segmented Dataset

Metric	**Precision**	**Recall**	**F-1 Score**
Original Dataset	0.8412	0.8737	0.8571
Segmented Dataset	0.9851	0.8721	0.9251

From the above table, we can see that the segmented-dataset has a higher accuracy than the original dataset at F-1 score 0.9251, which proves that Topic Segmentation could help enhance the accuracy of information extraction to some extent.

In this small test, we consider the discharge or admission medication as the desired information to be extracted. The reason why the result on segmented dataset and original dataset has a that slight difference is that segmentation algorithm could ignore those paragraphs which are not related to medications. In original dataset, however, medication entity might appear in anywhere not only the medication part, but also hospital course, medical history and so on. This situation makes the original dataset have the same true positives and false negatives as segmented dataset. Unfortunately, it could also increase the false positives, which brings that original dataset obtains a lower precision that segmented dataset.

Chapter 7. Conclusion and Future Work

7.1. Conclusion

In this thesis study, we developed a medical information extraction system for a pictorial information visualization system for visualizing summaries of medical data on graphical user interfaces. As input data, we have collected clinical notes which are somewhat structured with subsections and their titles, but they are not consistent manner. So, we applied natural language processing techniques to extract the information used for medical record summarization, such as medication names and medical imaging information as output.

Our system is composed of two parts: Topic Segmentation, to partition a clinical note into predefined topics and named entities recognition, to extract desired medical information. Our goal is to realize specific named entity extraction from a clinical note. Segmentation of the note according to topics is added to raise the accuracy of named entity extraction.

For the part of Topic Segmentation, we proposed a novel algorithm to automatically divide the clinical notes into five parts: history, medications, hospital courses, laboratories and physical examinations. We applied Naïve Bayes and Support Vector Machine models for training the predictors which could estimate the topic probabilities of tokenized sequences and assign five topic confidence scores to each of them. By taking the difference of confidence score of tokenized sequences, the position of each boundary, where the confidence score has a big change, would be detected. The process of score assignment and segmentation are done simultaneously while these two steps are generally separated in other algorithms. Consequently, our algorithm could potentially be more efficient when dealing with a large size of the dataset.

For the part of named entities extraction, we applied Conditional Random Fields for training the system of recognizing the medical entities: Medical Imaging Procedure Entities, Procedure Locations, Medications and so on, from clinical notes. We experimented using Metamap to map each medical terminology to its semantic type and group as one of the features for Conditional Random Fields model and found our system has a better performance on medication entities recognition than that of (Kumar et al., 2014; Patrick & Li, 2010; Yefeng Wang, 2009).

Also, we prove the usefulness of Topic Segmentation for Information Extraction in this study. Topic Segmentation, to some extent, could enhance the accuracy of information extraction because Topic Segmentation has the ability to reduce false positives.

Even though out methodology has been tested with clinical note information extraction, it is a general methodology which could be applied to realize any other Information Extraction tasks, not limited into clinical notes.

7.2. Future Work

In the future work, we will firstly aim to improve the performance of information extraction, no matter on Topic Segmentation side or Named Entities Recognition part. Increasing the dataset for training the models is a good way to improve the accuracy. This current study of Topic Segmentation is based on a small set of datasets, which brings that the performance of SVM-based segmenter is not satisfying. Expanding the dataset probably could improve it.

Since clinical notes contain not only medications, laboratories, hospital courses, physical examinations and history but also allergies which are crucial for the process of diagnosis. The next step is to expend our topics from five to more. Another idea is to refine current topics, such as form history to “history of illness”, “medical history” and “family history”. Finally, we will integrate our system with the medical information pictorial visualization system.

References

Assal, H., Seng, J., Kurfess, F., Schwarz, E., & Pohl, K. (2011). Semantically-enhanced information extraction. In *2011 Aerospace Conference* (pp. 1–14). IEEE. https://doi.org/10.1109/AERO.2011.5747547

Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, *34*(1/3), 177–210. https://doi.org/10.1023/A:1007506220214

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022. Retrieved from http://www.jmlr.org/papers/v3/blei03a.html

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(90001), 267D–270. https://doi.org/10.1093/nar/gkh061

Byrd, R. J., Steinhubl, S. R., Sun, J., Ebadollahi, S., & Stewart, W. F. (2014). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International Journal of Medical Informatics*, *83*(12), 983–992. https://doi.org/10.1016/J.IJMEDINF.2012.12.005

Chieu, H. L., & Ng, H. T. (2002). Named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics* - (Vol. 1, pp. 1–7). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1072228.1072253

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. Retrieved from http://arxiv.org/abs/cs/0003083

Dietterich, T. G. (2002). Machine Learning for Sequential Data: A Review (pp. 15–30). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-70659-3_2

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*, *121*, 279–290. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17095826

Dorfman, A. L., Fazel, R., Einstein, A. J., Applegate, K. E., Krumholz, H. M., Wang, Y., … Nallamothu, B. K. (2011). Use of Medical Imaging Procedures With Ionizing Radiation in Children. *Archives of Pediatrics & Adolescent Medicine*, *165*(5), 458–464. https://doi.org/10.1001/archpediatrics.2010.270

Edinger, T., Demner-Fushman, D., Cohen, A. M., Bedrick, S., & Hersh, W. (2017). Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2017*, 660–669. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/29854131

Fazel, R., Krumholz, H. M., Wang, Y., Ross, J. S., Chen, J., Ting, H. H., … Nallamothu, B. K. (2009). Exposure to Low-Dose Ionizing Radiation from Medical Imaging Procedures. *New England Journal of Medicine*, *361*(9), 849–857. https://doi.org/10.1056/NEJMoa0901249

Fournier, C. (2013). Evaluating Text Segmentation using Boundary Edit Distance. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *1*, 1702–1712. Retrieved from https://aclanthology.info/papers/P13-1167/p13-1167

Fournier, C., & Inkpen, D. (2012). Segmentation Similarity and Agreement. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 152–161. Retrieved from http://arxiv.org/abs/1204.2847

Galley, M., McKeown, K., Fosler-Lussier, E., & Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03* (Vol. 1, pp. 562–569). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1075096.1075167

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. In *International journal of pattern recognition and artificial intelligence* (pp. 9–41). https://doi.org/10.1142/9789812797605_0002

Ginter, F., Suominen, H., & Pyysalo, S. (2009). Combining hidden Markov models and

latent semantic analysis for topic segmentation and labeling: Method and clinical application. *International Journal of Medical Informatics*, *78*(12), e1–e6. https://doi.org/10.1016/J.IJMEDINF.2009.02.003

Hearst, M. A. (1993). *TextTiling: A Quantitative Approach to Discourse Segmentation*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9813&rep=rep1&type=pdf

Hearst, M. A. (1994). *Context and Structure in Automated Full-Text Information Access*. Retrieved from https://apps.dtic.mil/docs/citations/ADA632259

Hearst, M. A. (1997). *TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*. Retrieved from http://acl-arc.comp.nus.edu.sg/archives/acl-arc-090501d3/data/pdf/anthology-PDF/J/J97/J97-1003.pdf

Jensen, K., Soguero-Ruiz, C., Oyvind Mikalsen, K., Lindsetmo, R.-O., Kouskoumvekaki, I., Girolami, M., … Augestad, K. M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, *7*(1), 46226. https://doi.org/10.1038/srep46226

Jin, Y. (2016). *Interactive Medical Record Visualization based on Symptom Location in a 2D Human Body*. University of Ottawa. Retrieved from https://ruor.uottawa.ca/handle/10393/34255

Kaelber, D., Greco, P., & Cebul, R. D. (2005). Evaluation of a commercial electronic medical record (EMR) by primary care physicians 5 years after implementation. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2005*, 1002. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16779289

Kapler, T., & Wright, W. (2005). GeoTime Information Visualization. *Information Visualization*, *4*(2), 136–146. https://doi.org/10.1057/palgrave.ivs.9500097

Kumar, A., Alam, H., Kumar, R., & Sheel, S. (2014). Understanding Medical Named Entity Extraction in Clinical Notes. In *Health Informatics and Medical Systems* (pp. 201–204). Retrieved from https://www.i2b2.org/NLP/DataSets/

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic

Models for Segmenting and Labeling Sequence Data. *Departmental Papers (CIS)*. Retrieved from https://repository.upenn.edu/cis_papers/159

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

LePendu, P., Iyer, S. V, Fairon, C., & Shah, N. H. (2012). Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. *Journal of Biomedical Semantics 2012 3:1*, *3*(1), S5. https://doi.org/10.1186/2041-1480-3-s1-s5

Li, H., & Yamanishi, K. (2000). Topic analysis using a finite mixture model. In *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* - (Vol. 13, pp. 35–44). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1117794.1117799

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to Information Retrieval. *Natural Language Engineering*, 100–103. Retrieved from http://eprints.bimcoordinator.co.uk/35/

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482–489. https://doi.org/10.1016/J.CSI.2012.09.004

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* - (Vol. 4, pp. 188–191). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1119176.1119206

Mccallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification, 41--48. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). *Spam Filtering with Naive Bayes-Which Naive Bayes?* Retrieved from http://www.iit.demokritos.gr/skel/i-config/

Mohanty, M., Atrey, P., & Ooi, W. T. (2012). Secure cloud-based medical data visualization. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12* (p. 1105). New York, New York, USA: ACM Press. https://doi.org/10.1145/2393347.2396394

Mohit, B. (2014). Named Entity Recognition (pp. 221–245). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45358-8_7

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, *17*(1), 21–48. Retrieved from https://dl.acm.org/citation.cfm?id=971740

Morwal, S., Jahan, N., & Chopra, D. (2012). Named Entity Recognition using Hidden Markov Model (HMM). *International Journal on Natural Language Computing (IJNLC)*, *1*(4). https://doi.org/10.5121/ijnlc.2012.1402

Pakhomov, S., Buntrock, J., & Duffy, P. (2005). *High Throughput Modularized NLP System for Clinical Text*. Retrieved from http://umlslex.nlm.nih.gov

Patrick, J., & Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, *17*(5), 524–527. https://doi.org/10.1136/jamia.2010.003939

Patrick, J., & Min Li. (2009). Intelligent Clinical Notes System: An information retrieval and information extraction system for Clinical Notes. In *2009 11th International Conference on e-Health Networking, Applications and Services (Healthcom)* (pp. 108–115). IEEE. https://doi.org/10.1109/HEALTH.2009.5406206

Pevzner, L., & Hearst, M. A. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, *28*(1), 19–36. https://doi.org/10.1162/089120102317341756

Purver, M. (2011). *Topic Segmentation*. Spoken language understanding: systems for extracting semantic information from speech. Retrieved from

http://www.eecs.qmul.ac.uk/~mpurver/papers/purver11slu.pdf

Rabiner, L. R., & Juang, B. H. (1986). *An Introduction to Hidden Markov Models*. Retrieved from http://sistemas-humano-computacionais.wdfiles.com/local--files/capitulo%3Amodelagem-e-simulacao-de-humanos/rabiner86.pdf

Ramshaw, L. A., & Marcus, M. P. (1999). Text Chunking Using Transformation-Based Learning. In *TNatural Language Processing Using Very Large Corpora* (pp. 157–176). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-2390-9_10

Riedl, M., & Chris, B. (2012). Text Segmentation with Topic Models. *Language Technology and Computational Linguistics*, *27*, 47–69. Retrieved from https://jlcl.org/content/2-allissues/11-Heft1-2012/H2012-1.pdf#page=53

Ruan, W., Appasani, N., Kim, K., Vincelli, J., Kim, H., & Lee, W.-S. (2018). Pictorial Visualization of EMR Summary Interface and Medical Information Extraction of Clinical Notes. In *2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)* (pp. 1–6). IEEE. https://doi.org/10.1109/CIVEMSA.2018.8439958

Ruan, W., & Lee, W. (2018a). Boundary Detection by Determining the Difference of Classification Probabilities of Sequences: Topic Segmentation of Clinical Notes. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 747–750). IEEE. https://doi.org/10.1109/BIBM.2018.8621195

Ruan, W., & Lee, W. (2018b). Recognising Named Entity of Medical Imaging Procedures in Clinical Notes. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 743–746). IEEE. https://doi.org/10.1109/BIBM.2018.8621452

Sarawagi, S., & Cohen, W. W. (2005). Semi-Markov Conditional Random Fields for Information Extraction. *Advances in Neural Information Processing Systems*, 1185–1192. Retrieved from http://papers.nips.cc/paper/2648-semi-markov-conditional-random-fields-for-information-extraction.pdf

Savova, G. K., Coden, A. R., Sominsky, I. L., Johnson, R., Ogren, P. V., Groen, P. C. de, & Chute, C. G. (2008). Word sense disambiguation across two domains: Biomedical

literature and clinical notes. *Journal of Biomedical Informatics*, *41*(6), 1088–1100. https://doi.org/10.1016/J.JBI.2008.02.003

Singh, S. (2018). Natural Language Processing for Information Extraction. Retrieved from http://arxiv.org/abs/1807.02383

Suo, J. (2017). *Pictorial Visualization System with Patient Portal for Problem-based Electronic Medical Record*. University of Ottawa. Retrieved from https://ruor.uottawa.ca/handle/10393/35975

Tepper, M., Capurro, D., Xia, F., Vanderwende, L., & Yetisgen-Yildiz, M. (2012). Statistical Section Segmentation in Free-Text Clinical Records. In *LREC*. Retrieved from http://staff.washington.edu/melihay/publications/LREC_2012.pdf

Wang, Y. (2009). Annotating and recognising named entities in clinical notes. In *ACLstudent '09 Proceedings of the ACL-IJCNLP 2009 Student Research Workshop* (pp. 18–26). Suntec, Singapore. Retrieved from https://dl.acm.org/citation.cfm?id=1667888

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., … Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/10.1016/J.JBI.2017.11.011

Wu, S. T., Liu, H., Li, D., Tao, C., Musen, M. A., Chute, C. G., & Shah, N. H. (2012). Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. *Journal of the American Medical Informatics Association*, *19*(e1), e149–e156. https://doi.org/10.1136/amiajnl-2011-000744

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0

Appendix

A list of Semantic Types[16] and their abbreviations to their full names in Metamap. The format of this list is "Semantic Group Abbreviation |Semantic Group Name | Full Semantic Type Name| Semantic Type Name Abbreviation".

ACTI	Activities & Behaviors	Activity	acty
ACTI	Activities & Behaviors	Behavior	bhvr
ACTI	Activities & Behaviors	Daily or Recreational Activity	dora
ACTI	Activities & Behaviors	Event	evnt
ACTI	Activities & Behaviors	Governmental or Regulatory Activity	gora
ACTI	Activities & Behaviors	Individual Behavior	inbe
ACTI	Activities & Behaviors	Machine Activity	mcha
ACTI	Activities & Behaviors	Occupational Activity	ocac
ACTI	Activities & Behaviors	Social Behavior	socb
ANAT	Anatomy	Anatomical Structure	anst
ANAT	Anatomy	Body Substance	bdsu
ANAT	Anatomy	Body System	bdsy
ANAT	Anatomy	Body Location or Region	blor
ANAT	Anatomy	Body Part, Organ, or Organ Component	bpoc
ANAT	Anatomy	Body Space or Junction	bsoj
ANAT	Anatomy	Cell Component	celc
ANAT	Anatomy	Cell	cell
ANAT	Anatomy	Embryonic Structure	emst
ANAT	Anatomy	Fully Formed Anatomical Structure	ffas
ANAT	Anatomy	Tissue	tisu
CHEM	Chemicals & Drugs	Amino Acid, Peptide, or Protein	aapp

[16] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

CHEM	Chemicals & Drugs	Antibiotic	antb
CHEM	Chemicals & Drugs	Biologically Active Substance	bacs
CHEM	Chemicals & Drugs	Biomedical or Dental Material	bodm
CHEM	Chemicals & Drugs	Carbohydrate	carb
CHEM	Chemicals & Drugs	Chemical	chem
CHEM	Chemicals & Drugs	Chemical Viewed Functionally	chvf
CHEM	Chemicals & Drugs	Chemical Viewed Structurally	chvs
CHEM	Chemicals & Drugs	Clinical Drug	clnd
CHEM	Chemicals & Drugs	Eicosanoid	eico
CHEM	Chemicals & Drugs	Element, Ion, or Isotope	elii
CHEM	Chemicals & Drugs	Enzyme	enzy
CHEM	Chemicals & Drugs	Hazardous or Poisonous Substance	hops
CHEM	Chemicals & Drugs	Hormone	horm
CHEM	Chemicals & Drugs	Immunologic Factor	imft
CHEM	Chemicals & Drugs	Inorganic Chemical	inch
CHEM	Chemicals & Drugs	Indicator, Reagent, or Diagnostic Aid	irda
CHEM	Chemicals & Drugs	Lipid	lipd
CHEM	Chemicals & Drugs	Nucleic Acid, Nucleoside, or Nucleotide	nnon
CHEM	Chemicals & Drugs	Neuroreactive Substance or Biogenic Amine	nsba
CHEM	Chemicals & Drugs	Organophosphorus Compound	opco
CHEM	Chemicals & Drugs	Organic Chemical	orch
CHEM	Chemicals & Drugs	Pharmacologic Substance	phsu
CHEM	Chemicals & Drugs	Receptor	rcpt
CHEM	Chemicals & Drugs	Steroid	strd
CHEM	Chemicals & Drugs	Vitamin	vita
CONC	Concepts & Ideas	Classification	clas
CONC	Concepts & Ideas	Conceptual Entity	cnce

CONC	Concepts & Ideas	Functional Concept	ftcn
CONC	Concepts & Ideas	Group Attribute	grpa
CONC	Concepts & Ideas	Idea or Concept	idcn
CONC	Concepts & Ideas	Intellectual Product	inpr
CONC	Concepts & Ideas	Language	lang
CONC	Concepts & Ideas	Qualitative Concept	qlco
CONC	Concepts & Ideas	Quantitative Concept	qnco
CONC	Concepts & Ideas	Regulation or Law	rnlw
CONC	Concepts & Ideas	Spatial Concept	spco
CONC	Concepts & Ideas	Temporal Concept	tmco
DEVI	Devices	Drug Delivery Device	drdd
DEVI	Devices	Medical Device	medd
DEVI	Devices	Research Device	resd
DISO	Disorders	Acquired Abnormality	acab
DISO	Disorders	Anatomical Abnormality	anab
DISO	Disorders	Congenital Abnormality	cgab
DISO	Disorders	Cell or Molecular Dysfunction	comd
DISO	Disorders	Disease or Syndrome	dsyn
DISO	Disorders	Experimental Model of Disease	emod
DISO	Disorders	Finding	fndg
DISO	Disorders	Injury or Poisoning	inpo
DISO	Disorders	Mental or Behavioral Dysfunction	mobd
DISO	Disorders	Neoplastic Process	neop
DISO	Disorders	Pathologic Function	patf
DISO	Disorders	Sign or Symptom	sosy
GENE	Genes & Molecular Sequences	Amino Acid Sequence	amas
GENE	Genes & Molecular Sequences	Carbohydrate Sequence	crbs

GENE	Genes & Molecular Sequences	Gene or Genome	gngm
GENE	Genes & Molecular Sequences	Molecular Sequence	mosq
GENE	Genes & Molecular Sequences	Nucleotide Sequence	nusq
GEOG	GEOGGeographic Areas	Geographic Area	geoa
LIVB	Living Beings	Age Group	aggp
LIVB	Living Beings	Amphibian	amph
LIVB	Living Beings	Animal	anim
LIVB	Living Beings	Archaeon	arch
LIVB	Living Beings	Bacterium	bact
LIVB	Living Beings	Bird	bird
LIVB	Living Beings	Eukaryote	euka
LIVB	Living Beings	Family Group	famg
LIVB	Living Beings	Fish	fish
LIVB	Living Beings	Fungus	fngs
LIVB	Living Beings	Group	grup
LIVB	Living Beings	Human	humn
LIVB	Living Beings	Mammal	mamm
LIVB	Living Beings	Organism	orgm
LIVB	Living Beings	Plant	plnt
LIVB	Living Beings	Patient or Disabled Group	podg
LIVB	Living Beings	Population Group	popg
LIVB	Living Beings	Professional or Occupational Group	prog
LIVB	Living Beings	Reptile	rept
LIVB	Living Beings	Virus	virs
LIVB	Living Beings	Vertebrate	vtbt
OBJC	Objects	Entity	enty

OBJC	Objects	Food	food
OBJC	Objects	Manufactured Object	mnob
OBJC	Objects	Physical Object	phob
OBJC	Objects	Substance	sbst
OCCU	Occupations	Biomedical Occupation or Discipline	bmod
OCCU	Occupations	Occupation or Discipline	ocdi
ORGA	Organizations	Health Care Related Organization	hcro
ORGA	Organizations	Organization	orgt
ORGA	Organizations	Professional Society	pros
ORGA	Organizations	Self-help or Relief Organization	shro
PHEN	Phenomena	Biologic Function	biof
PHEN	Phenomena	Environmental Effect of Humans	eehu
PHEN	Phenomena	Human-caused Phenomenon or Process	hcpp
PHEN	Phenomena	Laboratory or Test Result	lbtr
PHEN	Phenomena	Natural Phenomenon or Process	npop
PHEN	Phenomena	Phenomenon or Process	phpr
PHYS	Physiology	Cell Function	celf
PHYS	Physiology	Clinical Attribute	clna
PHYS	Physiology	Genetic Function	genf
PHYS	Physiology	Mental Process	menp
PHYS	Physiology	Molecular Function	moft
PHYS	Physiology	Organism Attribute	orga
PHYS	Physiology	Organism Function	orgf
PHYS	Physiology	Organ or Tissue Function	ortf
PHYS	Physiology	Physiologic Function	phsf
PROC	Procedures	Diagnostic Procedure	diap
PROC	Procedures	Educational Activity	edac

PROC	Procedures	Health Care Activity	hlca
PROC	Procedures	Laboratory Procedure	lbpr
PROC	Procedures	Molecular Biology Research Technique	mbrt
PROC	Procedures	Research Activity	resa
PROC	Procedures	Therapeutic or Preventive Procedure	topp