EVALUATION OF MACHINE LEARNING MODELS FOR PATIENT DATA DE-IDENTIFICATION IN CLINICAL RECORDS

by

Yamani Kakarla

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2018

To my parents and sisters

Table of Contents

List of Tables

List of Figures

Abstract

In research that involves medical records, it is important that patient-identifiable details are removed before the records are made available for research, a requirement enforced by the HIPAA Privacy Rule and Public Law 104-191. De-identification is the redaction or masking of individually identifiable pieces of patient health information (PHI) from the clinical notes to protect the patient's identity from being exposed. With an increasing adoption of electronic health records (EHRs) in healthcare industries, there is an increasingly large amount of medical information available in digital format. Performing de-identification on such large collections of records is a challenging task to complete manually. Automated de-identification systems address this issue by automatically tagging the free-text medical records.

The primary objective of this research is to explore automated techniques in *natural language processing* (NLP) for de-identifying unstructured health records. To facilitate studies in automatic de-identification using statistical models, my work provides an overview of the evaluation results of a core NLP based de-identification model. My thesis describes the complexities in learning the variants of the model in the parameter space, explains performance metrics (precision, recall, and F1 measure) of the models, compare results with a rule-based de-identification system and finally provides directions for future research. The data used for evaluation consisted of three different types of medical notes: discharge summaries, longitudinal medical records, and nursing notes. Through model-specific feature engineering and introduction of hidden neural gates (model parameter) to the core model, a highest tag-level F1-measure of 0.967 on discharge summaries was achieved. For this task, in cases where more importance should be given to precision, the F1 measure can over-weight recall. The performance results from all models are encouraging and provide scope for future work. Overall this thesis intends to increase practitioners' understanding of the nature of de-identification models and how they are trained, to help preserve medical information while not compromising the privacy of individuals.

List of Abbreviations Used

AHIA	Alberta's Health Information Act
ANN	Artificial Neural Networks
CLL	Conditional Log-Likelihood
CRF	Conditional Random Field
DAG	Directed Acyclic Graph
EHR	Electronic Health Records
EMR	Electronic Medical Record
HIPAA	Health Insurance Portability and Accountability Act
HITECH	Health Information Technology for Economic and Clinical Health
HMM	Hidden Markov Models
I2b2	Informatics for Integrating Biology and the Beside
JSON	JavaScript Object Notation
JVM	Java Virtual Machine
LSTM	Long Short Term Memory
MAT	MIST Annotation Toolkit
MEMM	Maximum Entropy Markov Model
MIST	MITRE Identification Scrubber Toolkit
ML	Machine Learning
MRF	Markov Random Fields
MUC	Message Understanding Conference
NER	Named Entity Recognition
NLP	Natural Language Processing
NPS	National Physician Survey
PGM	Probabilistic Graphical Models
PHI	Patient Health Information
PHIA	Personal Health Information Act
PHIPA	Personal Health Information Protection Act
POS	Part-Of-Speech
PPM	Positive Predictive Value
PSA	Periodic Step-size Adaptation
RNN	Recurrent Neural Networks
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine

Acknowledgements

Firstly, I express my sincere and greatest gratitude to my supervisor, Dr.Stan Matwin, for supporting and guiding me throughout my Master program. Thank you for giving me the opportunity to work under your guidance. Your precious time and insightful comments significantly contributed to this thesis.

My sincere gratitude also goes to my co-supervisor, Dr.Aaron Gerow, for the constant support and constructive comments on my work. Your patience and guidance have been my strength in the completion of my thesis. Thank you for all the support you have extended to me.

I would like to thank my committee members, Dr.Vlado Keselj and Dr.Srinivas Sampalli, for serving on my thesis and giving insightful comments on my work.

I thank, the Canadian Institutes of Health Research (CHIR) and the Institute for Big Data Analytics for supporting this project. Above all I would like to thank Dalhousie University, for being the platform to my passion for learning.

Thanks to all my lab-mates in the Big Data Institute: Xiang, Fateha, Eman, Dr.Amilcar, Behrouz, Habibeh and Salil for all the discussions and fun time we had during the past two years. Special mentions to Xiang, for helping me clarify my doubts anytime.

To my Mom (Kameswari) and Dad (Dass Reddy), I feel immensely proud to be your daughter. Thank you for supporting me to pursue my dream. A warm and heartful thanks to my family members: Bhagya, Mahesh, Mamatha, and Shashi for being the pillar behind my achievement. Special mentions to my husband, Dileep, who has been the traveler in disguise throughout this journey. I extend my thanks to my friends: Dhivya Jayaraman, Pooja Srinivasan, Ruhi Madiwale and Jeybalaji, for being my crazy squad lifting me up in all my ups and downs. Thank you for the companionship.

Finally, I thank God almighty for all his blessings throughout this journey.

Chapter 1

Introduction

This chapter begins by exploring the motivation of this thesis work from the perspective of medical data and privacy, then explains the aim and objectives, scientific contribution and later discusses the organization of the thesis.

1.1 Motivation

With a massive increase in the number of clinical records generated from healthcare industries, the scope of text analysis in healthcare data has also increased dramatically. As the complexities in patient disease trajectories increase, it is a challenge for the physicians to make clinical decisions based on the complexity of patient's clinical history. One obvious strength of text analysis in healthcare is that a lot of data, that might be out of reach for any single practitioner, can be combined to find out interesting patterns from the clinical notes.

Applying text mining techniques on a huge amount of clinical records help in identifying the underlying patterns and developing new ideas for enhanced treatments. Eventually, it improves the healthcare quality and promotes clinical and research initiatives. Several text mining challenges such as identifying unknown disease correlations and genotype-phenotype relationships [40], text mining techniques in molecular biology and biomedicine to extract bioentities such as chemical compounds and proteins [45], biomedical text mining and its applications in cancer research [87] and other major applications emphasize the importance of text mining in healthcare data.

Most of the healthcare data is stored in the form of unstructured free-text notes: discharge summaries, clinical examination reports, nursing notes and so on. A significant amount of clinical data is recorded as narrative text in the clinical notes. While it is easy to process or analyze structured data as in data formatting and content requirements [55], it is a tedious task to perform analysis on unstructured medical records [46]. Currently, most of the medical notes are in a computerized format. Prior to this digital storage, clinical records were stored in a paper file system. The transition of paper-based medical records to *electronic health records* (EHR) has been adopted in most of the healthcare systems as part of the Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted as part of the American Recovery and Reinvestment Act of 2009 [36].

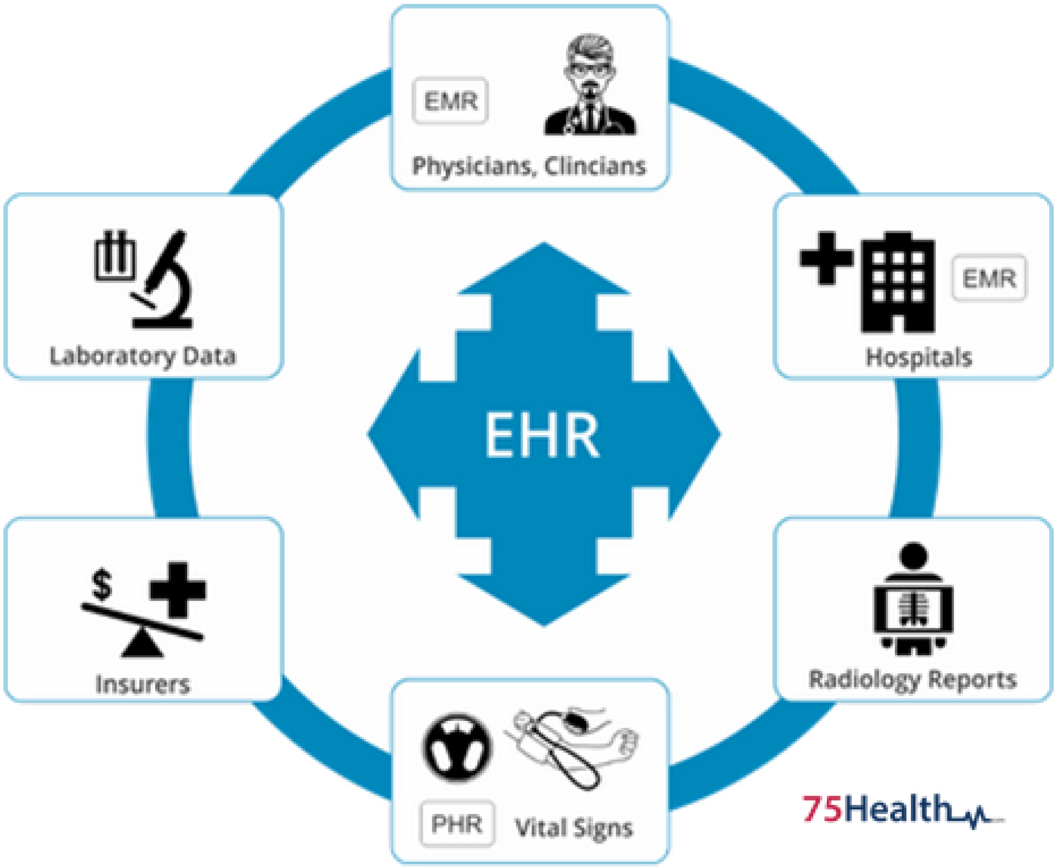


Figure 1.1: As depicted in [10], the image shows different categories of data in an electronic health record (EHR). In the image, electronic medical record (EMR) refers to the digital version of the patient information.

1.1.1 Electronic Health Records

EHRs were introduced at the beginning of 1991 by the *Institute of Medicine* (a division of the National Academics of Sciences, Engineering, and Medicine) as part of their sponsored studies [30]. As depicted in Fig 1.1, EHR contains a wide range of data including patient's medical history, prescription details, laboratory results, radiology images, discharge summary notes, physician notes, physician name, demographic details, personal statistics: age, weight and height, identification details: healthcare id, contact address, telephone number, email address and billing information. These categories are individually referred to as electronic medical r ecords (EMRs). EMRs are restricted to an individual practice and are mainly used for diagnosis and treatment. But EHRs allow sharing of medical and other information about the patient for decision-making purposes. For the outcome of patient care, EHRs have become one of the potentially comprehensive sources of information.

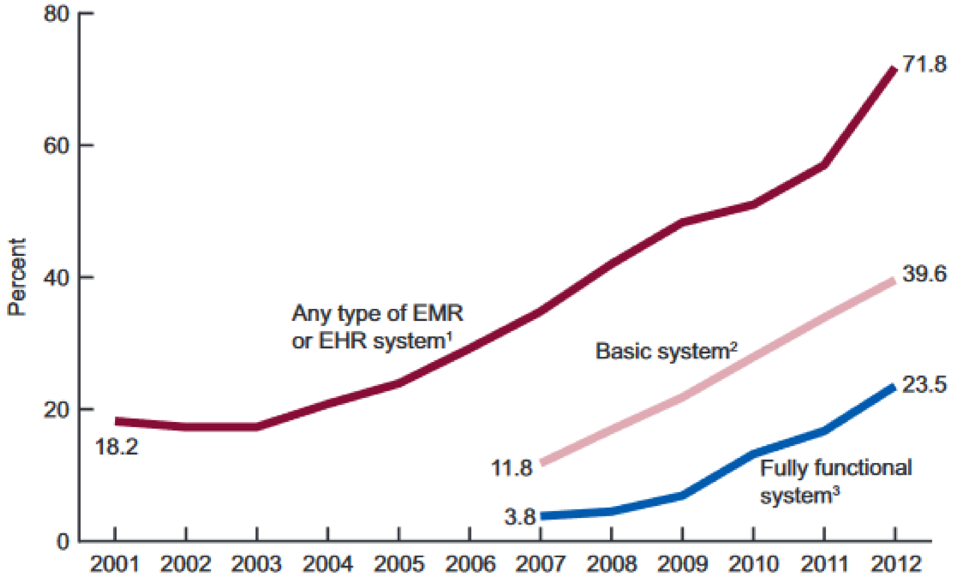


Figure 1.2: As shown in [37], the graph shows the trend in adoption of EHRs by office-based physicians for 2007 through 2012. It differentiates the trend by comparing adoption of any type of EHR to that of a basic and fully functional system.

Fig 1.2 shows the trend in the adoption of EHRs by office-based physicians for 2007 through 2012 in the United States [37]. It can be observed from the survey details, that the percentage increased to 71.8% in 2017 up from 34.8% in 2007 in the adoption of any type of EHRs by the office-based physicians. In Canada, a similar survey shows that the adoption rate of EMRs has significantly increased. As per the National Physician Survey (NPS) in 2013, the overall adoption rate of EMRs in Canada has increased from 20% of practitioners in 2006 to 62% of practitioners in 2013 [21]. As EHRs contain rich clinical information, they are being used for an array of medical researches [15, 56, 59]. Increase in the percentage of adopting EHR increases the amount of clinical data available for research. Performing analysis on EHR helps in understanding the patient population of a country.

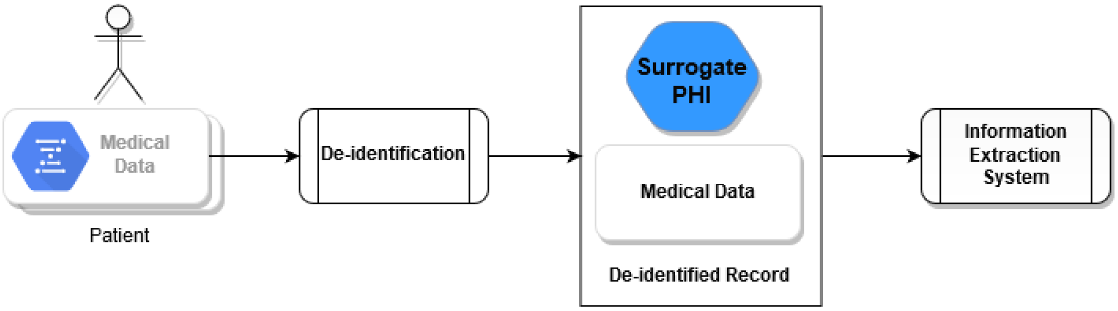


Figure 1.3: Medical data and their usage in Information Extraction Systems (Downstream Applications).

In case of data analytics, importance is given to the medical data and not the information about the owner of the data as shown in 1.3. For research that focuses on monitoring the trends in dealing with a drug, importance is given to the medical information and the time period but not on the individual's identity. Thus a key challenge in using EHRs is avoiding the exposition of confidential and patient identifiable information which needs to be redacted or masked prior to sharing data and publishing results. Such types of information are legally determined in the U.S. by the *Health Insurance Portability and Accountability Act* (HIPAA 1996) and are referred to as *protected health information* categories. This privacy legislation is followed across the provinces of Canada in the form of Canadian healthcare privacy laws. Some of them are, Nova Scotia's *Personal Health Information Act*[1], Ontario's *Personal Health Information Protection Act* (PHIPA)[2] and *Alberta's Health Information Act* (HIA)[3].

1.1.2 Privacy

Privacy is always a concern for all healthcare industries. It is of utmost importance that the patient's health data and personally identifiable data be protected so that there are no chances for any unauthorized access or usage of data. This section starts with the discussion on trends in security breaches of medical data followed

[1] PHIA: `https://nslegislature.ca/legc/bills/61st_2nd/3rd_read/b089.htm`

[2] PHIPA: `https://www.ontario.ca/laws/statute/04p03`

[3] HIA: `http://www.albertanetcare.ca/learningcentre/Health-Information-Act.htm`

by elaborating the types of information to be protected in medical data and later discusses the Privacy laws enforced for healthcare data. The discussion is directed towards regulations in HIPAA due to the research requirements.

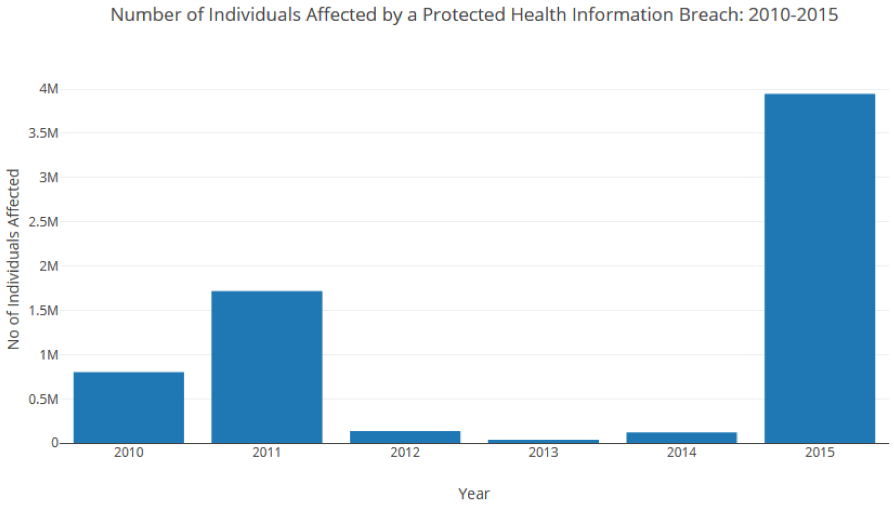


Figure 1.4: Number of individuals affected by a PHI Breach: 2010-2015 in U.S based on the data collected in [32]. Source: U.S. Department of Health and Human Services Office for Civil Rights, Feb 1, 2016.

As the medical records are stored electronically there is a greater access to this data by a wide group of stakeholders. On one hand, the exchange of data among the groups lead to security breach resulting in misuse of the patient's identity and therefore leading to a privacy breach. Security is the mechanism by which privacy is ensured and a breach in such mechanism is known as a security breach. On the other hand, privacy is the right of an individual from disclosing their identity. A compromise in the security paves the way to privacy breaches. In the last few years (2006-2012), around 767 security breaches were recorded by the healthcare providers that resulted in a compromise of confidential health information of around 23,625,933 patients [5].

According to the data in [32], collected by U.S. Department of Health and Human Services Office for Civil Rights, as of February 1, 2016, around 113 million individuals were affected by protected health information breaches. The number of breaches mentioned in [32] is categorized based on the type of security mechanism that is compromised. They are individuals affected by a hacking/IT incident breach and those affected by a non-hacking/IT incident breach. Considering the importance of sharing EMRs among researchers, only the incidents occurred due to a PHI breach in EMR are mentioned in Table 1.1.

Table 1.1: Total number of individuals affected by a PHI breach in EMR as mentioned in [32].

Source of Information	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
EMR	803600	1720064	136751	40196	121845	3948985

Fig 1.4 shows the trend in the number of individuals affected by a PHI breach in the U.S. From Table 1.2, the number of reported breaches of electronic medical records has increased to 16 in 2015 up from 3 in 2010. Similarly in the year 2017, province of Nova Scotia in Canada encountered the largest privacy breach recorded ever in the province. A total of 337 patient's health information has been accessed inappropriately during a privacy breach. This enforced healthcare industries to follow strong privacy rules which will protect the PHI. HIPAA introduced the privacy laws related to healthcare data in the U.S.A. My work is based on the HIPAA regulations.

HIPAA

The *Standards for Privacy of Individually Identifiable Health Information (Privacy Rule)* established national standards for preserving electronic health records and the personally identifiable information in the EHRs [58]. The regulations for usage and disclosure of the patient's health information are determined by the Privacy Rule. It also addresses the standards for individual's privacy rights in understanding and controlling the way their health information is used. The main objective of this rule is to provide proper protection to the patient's health information while allowing the flow of information required for promoting high quality health care.

Protected Health Information (PHI)

The privacy rule protects the *patient's individually identifiable* (PII) information and this protected information is called "protected health information". Patient's individually identifiable health information refers to the data that relates to,

1. individual's past, present and future health condition, health care provision details and payment history for the healthcare provider.

2. data that identifies an individual or that of which can be considered a reasonable basis for identifying an individual.

PHI is defined in 45 *(Code of Federal Regulations)* 160.103 and is referenced in Section 13400 of Subtitle D (Privacy) of the HITECH Act [41]. There are two rules under HIPAA: Privacy Rule protects PHI in any medium and Security Rule protects EHRs.

Table 1.2: Number of reported PHI breaches for 2010-2015 as shown in [32]. The count specified is the occurrences of breach with respect to the type and source of information. Occurrences due to EMR is denoted in bold values.

	2010	**2011**	**2012**	**2013**	**2014**	**2015**
Type of Information Breach						
Hacking/IT incident	10	16	16	23	32	57
Improper disposal	10	7	7	13	11	6
Loss	18	17	19	24	28	22
Theft	127	118	117	124	113	80
Unauthorized access/disclosure	7	26	25	63	72	100
Other breach	22	2	18	24	28	0
Source of Information Breach						
Desktop computer	28	35	23	39	29	29
Electronic medical record	**3**	**6**	**6**	**14**	**14**	**16**
E-mail	5	2	10	20	36	37
Laptop	50	38	51	67	42	38
Network server	17	16	20	30	46	41
Paper/Film	46	45	47	53	62	67
Portable Electronic Device	6	2	19	20	22	15
Other source	42	50	26	24	34	22

There are 18 categories of PHI as mentioned in Table 1.3, that needs to be redacted or masked to protect the individually identifiable health information. The standards for protecting this information is provided under the *de-identification* standard of the

HIPAA Privacy Rule [45 CFR 164.514] [41]. The process of identifying and removing/replacing the PHI terms in the medical notes is known as De-identification. This is achieved with the help of clinicians, physicians or dedicated medical personnel who is familiar with medical terminologies. There are automated ways of de-identifying medical notes and it can be broadly categorized into rule-based systems and statistical models that use *machine learning* (ML) techniques. The core implementation of automated de-identification involves *Natural Language Processing* (NLP) concepts and techniques.

Table 1.3: PHI Categories

No	PHI Category
1	Name
2	Geographic divisions smaller than a state such as address
3	Date elements except year related to individual
4	Age (if over 89)
5	Telephone numbers
6	Fax numbers
7	Electronic mail address
8	Social Security number
9	Medical record number
10	Health plan beneficiary number
11	Account number
12	Certificate/license number
13	Vehicle identifiers, serial numbers or license plate numbers
14	Device identifiers and serial numbers
15	Web Universal Resource Locators (URLs)
16	Internet Protocol (IP) address numbers
17	Biometric identifiers including finger and voice prints
18	Full face photographic images
19	Any other unique identifying number, characteristic or code

1.2 Objectives and Contribution

Based on the format and structure of the medical records, and the problem definition, many automated de-identification systems have been implemented. The main aim of this thesis is to study the nature and behavior of such models and provide an insight on the performance of statistical models to that of a rule-based system in de-identifying different types of medical records while retaining the medical information in the records. To deal with this objective, we first evaluate existing de-identification models on different types of medical notes, then discuss the challenges and factors influencing the performance of the model on each data-set and make a comparative study on performances of ML-based models and rule-based system on the different data-sets. To meet the research aim, we set the following objectives:

1. Discuss the design, implementation, and training of ML-based and rule-based de-identification models.

2. Discuss the format, nature of medical records and annotations used in exploring the de-identification systems.

3. Perform 10-fold cross-validation to evaluate and compare the performance of the statistical models.

4. Demonstrate the effectiveness of ML-based de-identification models with minimal feature engineering for the task of de-identification on different medical notes.

This thesis work aims at aiding healthcare institutions to choose the best approach in de-identifying the medical records, based on various factors such as availability of annotated data for training the models for a better performance and the number of instances required for each type of PHI category in a given data-set. This thesis contributes:

1. An exhaustive and comprehensive evaluation of the different variants of CRFs for a sequence labelling task like de-identification.

2. An approach that can be followed for estimating the performance of a de-identification system.

3. Demonstrate the importance of Artificial Intelligence for de-identification using Long Short-Term Memory (LSTM) based Recurrent Neural Networks (RNN), which has already been incredibly successful in applications like speech recognition, part-of-speech tagging, image classification, language modeling and many more.

1.3 Thesis Outline

In this thesis, I present a new approach for exploring NLP based free-text analysis techniques, that uses Conditional Random Fields, by evaluating statistical models and rule-based system for the task of de-identification on different types of medical records. The rest of the thesis is organized as follows: Chapter 2, provides a review of the background and relevant work; in Chapter 3, discuss in detail the design and implementation of the de-identification models; in Chapter 4, describe the three de-identified medical data-sets and the annotations used; in Chapter 5, discuss the evaluation experiments and the results of the models on all three data-sets and demonstrate the importance of deep learning approaches for the task of de-identification through the experimental results from a bi-directional LSTM based RNN model; finally in Chapter 6, I conclude with the future directions of this thesis.

Chapter 2

Background and Related Work

In this Chapter, we first describe the task of de-identification, from the perspective of NLP tasks. We then describe the background of CRFs, from the perspective of directed and undirected, generative and discriminative models and conclude with a discussion on CRF variants. Finally, we discuss related models in the task of de-identification.

2.1 Natural Language Processing

NLP is a branch of computer science, artificial intelligence, and computational linguistics which involves analysis and manipulation of the natural language in the form of text or speech. As described in [22], it involves computer programs and software that take natural language data as input and generate natural language data as output. Some of the major challenges of NLP include speech recognition, language translation, and text generation. The history of NLP started in the early 1950s with some interesting experiments like machine translation [39] of Russian sentences into English. Later in the 1960s, there were some notable NLP systems like ELIZA [81] one of the first chatterbots and SHRDLU[1] is a language parser which parses user instructions to move various objects in a block world. With improvements in structuring the real-world information into computer understandable data, many chatterbots were introduced in the 1970s like PARRY [19], a bot to simulate a person with paranoid schizophrenia; JABBERWACKY[2] simulates natural human chat in a humorous, interesting and entertaining way; and RACTER a bot that generates random English prose [20].

Most of the NLP systems relied on hand-written rules until the introduction of *machine learning* algorithms in the late 1980s. Initial ML algorithms like Decision

[1] SHRDLU: http://hci.stanford.edu/winograd/shrdlu/

[2] JABBERWACKY:https://web.archive.org/web/20050411014336/http://jabberwacky.com/

Trees [74] are rule-based systems, however, NLP tasks like part-of-speech tagging first used *Hidden Markov Models* (HMM) [66] which are more suited for the sequential character of the NLP data. This eventually increased the use of statistical models, for various NLP tasks, by making probabilistic decisions based on real-valued weights attached to the features of the input data [22]. Commonly used NLP tasks in real-world applications are categorized into Syntax (lemmatization, parsing, part-of-speech tagging, stemming); Semantics (lexical semantics, machine translation, named entity recognition, natural language generation, sentiment analysis); Discourse (automatic summarization, coreference resolution and discourse analysis) and Speech (speech recognition, text-to-speech).

2.1.1 Sequence labeling

Sequence labeling is a pattern recognition task in NLP, that involves recognizing the pattern of a sequence of class labels $\vec{y} = (y_1, ..., y_n) \epsilon \mathit{Y}^n$ for a given input sequence $\vec{x} = (x_1, ..., x_n) \epsilon \mathit{X}^n$. The inputs x_j in the sequence labeling task are tokens/words and the class labels y_j can be part-of-speech (POS) tags for a part-of-speech tagging task [44] or any named entity classes in a named entity recognition (NER) task [28]. In most of the sequence labeling problems, the sequence of tokens is represented in a specific format, while in certain problems a raw labeling happens. Raw labeling involves assigning a class label to a given token like in POS tagging, where each word/token gets a single tag which is the part-of-speech of the particular word.

Tasks like NER requires join segmentation and labeling. For example, in the sentence “James Cameron directed the movie, Avatar.”, the entire phrase (“James Cameron”) should be tagged as “PERSON”. It is necessary to know the beginning and end of the named entity. To achieve this, tokens are labeled in a specific notation like *BIO* [64], where “B_X” marks the beginning of the phrase of a named entity X, “I_X” marks the continuation of the phrase X and “O” specifies other tokens that are not in the named entity. Therefore the labeling of the sentence looks like: James/B_PERSON, Cameron/I_Person, directed/O, the/O, movie/O, ,/O, Avatar/B_MISC. Some of the methods that are already implemented for sequence labeling problems include HMM [66] (generative) and Maximum Entropy Markov Model (MEMM) (discriminative) [53]. This kind of approach is followed in de-identification systems. A de-identification system processes a given input sentence into tokens and assigns labels to each token resulting in a sequence labeling task.

2.1.2 Named Entity Recognition

Named Entity Recognition is a sequence labeling task under the “information extraction” domain, which involves extracting specific kinds of information, unlike general classification task, where all of the information from a given document is extracted to classify a document. Message Understanding Conference (MUC) [28] introduced the first NER task which studied various information extraction techniques for extracting a “named entity”: persons, organizations, locations, currency values, and artifacts. The type of named entities differ based on the problem and domain of the data that is involved. Evaluations done in MUC used newswire text. De-identification is a sequence labeling task and is considered as a more specific form of NER [54]. In de-identification, the inputs are the tokens extracted from the medical records (EHRs) and the class labels are the PHI annotations from Table 1.3.

Variety of approaches have been employed in implementing NER across various languages and domains with respect to the trends in implementation of NLP. Initial approaches are rule-based systems that follow simple pattern matching algorithms/techniques. They rely on manually generated features by using lexical directories, regular expressions, and hand-written rules. Stanford CoreNLP [52] toolkit uses annotator tasks to generate annotation information and produces state-of-the-art results in rule-based systems. Given the limitations like robustness and portability issues in rule-based systems, ML-based NER systems are highly in demand. ML-based systems involve training a model on a given set of annotated data and evaluating the performance of the model on new unseen data. For the models to perform well on unseen data, a huge amount of annotated data with a balanced class distribution is required for training. That is, in a rule-based system the rules are provided to the system, while in a machine learning model the rules are to be learned from the training data.

Some examples are NER using HMM-based chunk-tagger [85] and language independent NER system [63]. Some applications in medical domain include PowerBioNE [86] that identifies names in biomedical texts; TEXT2TABLE [14] medical event recognition systems for converting medical text into the table structure. While there are many NER systems used in identifying specific entity from a specific domain, de-identification focuses only on anonymization of medical records while preserving the integrity of the data as much as possible. Rule-based systems are limited to the domain of the data-sets, i.e., if the data-set contains names out of the lexical sources, the system fails to perform well. Therefore systems that rely on context-based learning are preferred to that of rule-based models. Context-based learning has become easy with the help of Conditional Random Fields (CRF) [47].

2.2 Background on CRF

Graphical models, also known as probabilistic graphical models (PGM), are probabilistic models that use graphical structures to express the conditional dependencies among random variables. It is a framework that unifies the graph theory with the probability theory for a factorized representation and inference of multivariate probability distributions. These models are distinct in its way by maintaining the control over the computational cost of formulating probabilistic models of complex phenomena in applied fields [42]. The graph structure comprises of edges and vertices, where a vertex represents a random variable and the edge represents the dependency relationship between the two random variables. Based on the direction of edges there are two categories of graphical models namely directed and undirected graphical models.

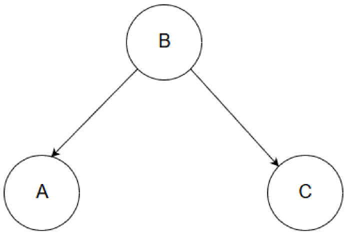


Figure 2.1: A Bayesian network, with random variables A, B, and C. In the graph, A is conditionally dependent on B, and C is conditionally dependent on B.

2.2.1 Directed and Undirected Graphical models

Bayesian networks [43] are the common directed acyclic graphs (DAG). Fig 2.1 from Cowell et al., 1999, illustrates a Bayesian network with directed edges in a graphical model. In the figure, the nodes are represented as $K = \{$A, B, C$\}$ and the edges are represented as $E = \{$(B, A), (B, C)$\}$. It is seen that A and C are conditionally independent and thus the probability is *P(A|B, C) = P(A|B)*, which means that, A is conditionally independent of C for the local probability calculation and it only depends on B. The joint probability function of all the events $\{X_1, ..., X_n\}$ is given by:

$$P[X_1, ..., X_n] = \prod_{1}^{n} P[X_i|pa_i] \tag{2.1}$$

where pa_i represents the parents of node X_i. And the joint distribution, as factorized by the Bayesian network, for the random variables A, B, C is given by:

$$P(A, B, C) = P(A|B).P(B).P(C|B) \tag{2.2}$$

In sequence labeling tasks, Bayesian Networks [69], HMM [66], and MEMMs [53] are considered as DAGs.

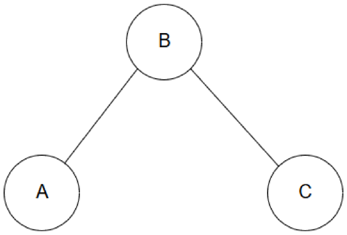


Figure 2.2: Undirected graphical model. The graph illustrates the conditional independence among the random variables A and C for given random variable B.

In a stochastic process, if the conditional probability distribution of the future state depends only on the current state and not on the set of states that preceded it, then such a memoryless state of the process is known as Markov Property. A set of random variables with this Markov property is known as Markov random fields (MRFs) or Markov Network or undirected graphical models. Fig 2.2 represents an undirected graphical model where the edges are bidirectional. Unlike directed graphs, where a causal relationship is between a vertex and its parent vertices, undirected graphs represent the correlation between the vertices i.e., it represents a non-causal relationship between the vertices. Thus it does not guarantee a consistent joint distribution when a set of conditional probabilities are multiplied, but the probability distribution can be factorized using potential or clique functions [73]. For a given set of random variables $X = (X_v)_{v \epsilon V}$, the joint density factorized over the cliques G (fully connected set of nodes), is given by the equation,

$$P(X = x) = \prod_{C \epsilon cl(G)} \phi_C(x_C) \tag{2.3}$$

where cl(G) represents the set of cliques and the function ϕ_C is known as the factor potential or clique potential. One of the notable variants of Markov random field is conditional random field. CRFs are undirected graphical models.

2.2.2 Generative and Discriminative models

Directed graphical models are most often used in generative models while CRFs are a type of discriminative undirected graphical models. This section discusses the basic understanding of CRFs from the perspective of generative and discriminative models. Let $X = \{x_1, .., x_n\}$ be a set of input observations and $Y = \{y_1, .., y_n\}$ be the output labels that need to be predicted. A generative model learns the joint probability distribution $p(x, y)$ that is factorized as $p(y, x) = p(y)p(x|y)$, where $p(x|y)$ is the class conditional probability and $p(y)$ is the class prior probability [84]. In generative models, parameter estimation involves estimating the optimal values for the parameters, that maximizes the log-likelihood. In a classification task, a generative model models the $p(x, y)$ and uses Bayesian rules to determine the posterior probabilities. Directed graphical models like Bayesian Networks, Naïve Bayes and HMMs are generative models. Discriminative models learn the conditional probability distribution $p(y|x)$ directly unlike generative models. These models learn the boundaries among classes by focusing on the differences among the categories in a data-set. Undirected graphical models, SVMs, Logistic Regression models and CRFs are discriminative models.

For sequence labeling task, HMMs and stochastic grammars [62] have been successfully used and their various applications in NLP were discussed earlier. For every production P (rewrite rules) for a nonterminal S, in a stochastic context-free grammar, there exists an associated probability such that there is a probability distribution over a set of productions. The difficulty of generative models like HMMs, in representing multiple interacting features of the given observations, motivated the usage of discriminative models like MEMMs [53]. But MEMMs and other non-generative finite-state models, which are based on next-state classifiers, have label bias problem, first described in [17]. It is, “the transitions leaving a given state compete only against each other, rather than against all other transitions in the model” [47]. CRFs address the limitations of HMMs by relaxing the independence assumptions of HMM and allows context-based learning [80] with the help of complex and highly correlated features. It solves the label bias problem of MEMMs by using a global normalizer to sum all possible states instead of per state normalizer. CRF was originally proposed by Lafferty et al. in 2001 [47].

CRFs are undirected graphical models whose nodes X represent input observations and Y represent output variables and from which the conditional probability $P(Y|X)$ is modeled. Lafferty et al. defines CRF as, *”Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \epsilon V}$, 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 Y_v obey the Markov property with respect to the graph: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G”*, where Markov Property is the state where the conditional probability distribution $P(Y|X)$ of future state depends only on the current state and not on the sequence of events that preceded it.

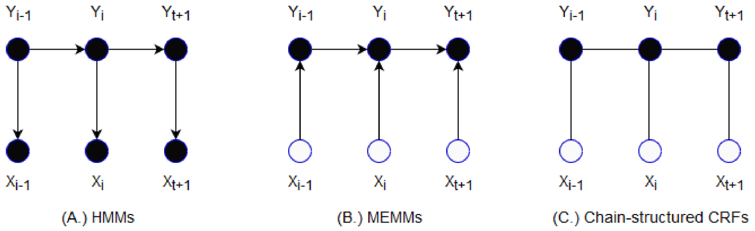


Figure 2.3: Graphical structure representations as shown in [47] for (A) HMMs (left), (B) MEMMs (center), (C) Chain-structured CRFs (right) for sequences. Open circles represent variables that are not generated by the model.

2.2.3 Variants of CRF

Based on the generalization of CRFs, there are different variants of CRF. They are linear chain CRF [73], higher-order CRFs [48], semi-Markov CRFs [65], and Neuro-CRFs [25].

Linear CRFs

Linear-chain CRFs are a combination of discriminative modeling and sequence modeling [73]. This kind of CRF uses linear chain factor graphs while a generic CRF uses more general factor graph. The conditional distribution $p(y|x)$, associated with the joint probability $p(y, x)$ that is factorized as an HMM, is known as linear-chain CRF. This type of CRF is considered HMM-like as they impose the dependencies on previous elements unlike general CRFs, where the dependencies are imposed on any arbitrary element. In linear-chain CRFs, the input and output sequences are of the same length. From the chain-structured CRF representation in Fig 2.3 (c) taken from [47], random variable Y represents the state variable to be inferred and that Y_i is structured to form a chain graph with an edge between Y_{i-1} and Y_i. Then the conditional dependency of the variable Y_i on X is defined with a set of feature functions $f(i, Y_{i-1}, Y_i, X)$. For each feature, the model assigns a weight with which the probability of Y_i is determined. The probability distribution $p(y|x)$ of a linear-chain CRF takes the form,

$$p(y|x) = \frac{1}{Z(x)} exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\}, \tag{2.4}$$

where $Z(x)$ is a normalization function and is defined as,

$$Z(x) = \sum_{y} exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\} \tag{2.5}$$

From the probability distribution equation (2.4) of linear-chain CRF, the observation argument x_t contains all the components of the global observations x needed for computing features at time t. This means that linear-chain CRFs are globally normalized. Thus for example, if a linear-chain CRF uses x_{t+1} as a feature, then the assumption is that x_t includes the identity of the word x_{t+1}. In a sequence labeling task, linear-chain graphs are of interest as they overcome the label bias problem mentioned in [47]. Linear CRFs are of interest due to their efficiency in sequence labeling problems.

Semi-Markov CRFs

Semi-Markov CRF (semi-CRFs) is the generalization of sequential CRFs which models variable length segmentation of labels Y [65]. Label-transitions are not modelled by CRFs, but they are modeled by semi-CRFs. A first-order sequential CRF is that which captures label dependencies between the adjacent sequence elements. The semi-CRFs are more powerful than first-order sequential CRFs and they are conditionally trained versions of semi-Markov chains [61]. Let $s = \{s_1, ..., s_p\}$ denote a segmentation of input observation x. Each *segment* contains a start position t_j, an end position u_j, and a label $y_j \epsilon Y$. By definition, all x_i's between the positions $i = t_j$ and $i = u_j$ have the tag y_j. For example, consider the sentence "I am going to Spain with Anna Morgan.". For a NER task, the segmentation of the above example might be $s = \{(1,1,O), (2,2,O), (3,3,O), (4,4,O), (5,5,I), (6,6,O), (7,8,I)\}$, where (1,1) denotes the start and end positions of a segment. The segmentation is with respect to the label sequence $y = \{O, O, O, O, I, O, I, I\}$, where the label "O" denotes a word outside the named entity and "I" denotes a word inside the named entity. Let us consider g denotes the *segment feature function* where $g = \{g^1, ..., g^k\}$ and g^k represents $g^k(j, x, s) \epsilon R$. Then the vector of these measurements is given as $G(x, s) = \sum_j^{|s|} g(j, x, s)$. A semi-CRF takes the form,

$$P(s|x, W) = \frac{1}{Z(x)} e^{W.G(x,s)} \tag{2.6}$$

where W is the weight vector and $Z(x)$ is a normalization function of the form, $Z(x) = \sum_{s'} e^W.G(x, s')$.

Semi-CRFs capture the label dependencies between adjacent segments while a first-order CRF captures label dependencies between adjacent sequence elements alone. This type of CRFs requires different types of features than other standard sequential CRFs.

Neural CRFs

Neural CRFs extend traditional CRFs by introducing a module based on articial neural networks [25]. This type of CRF relies on neural networks for learning high-level features. These learned features can be later used as inputs to a linear CRF. This module consists of a hidden layer of activation gates between the input/observations and the states/labels. This layer acts like an auto-encoder and helps accommodate non-linearities in the sequence data.

Features that are learned by the neural networks are passed through the hidden

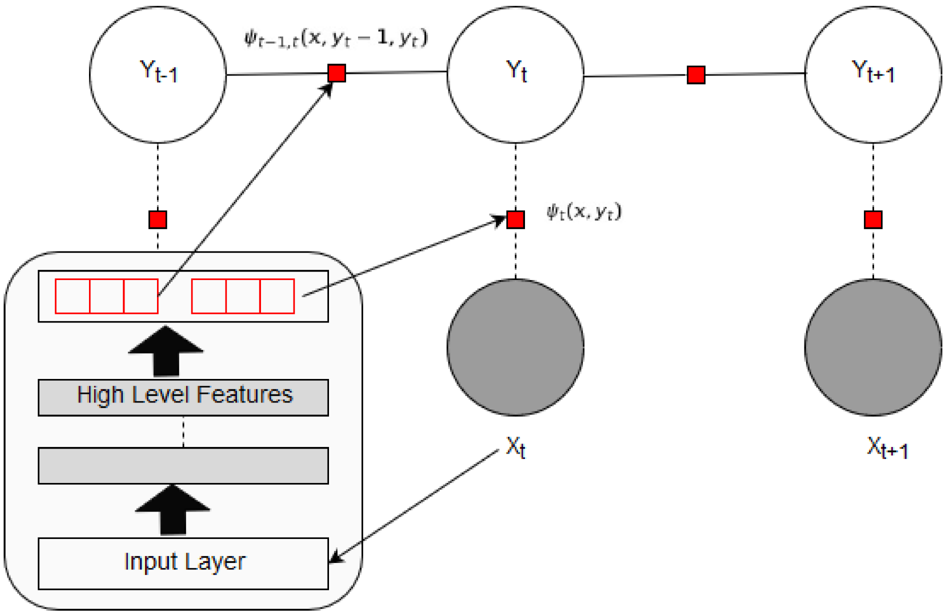


Figure 2.4: Figure of a linear-chain neuroCRF from [25]. X_t represents input sequences at time t and Y_t denotes the label sequences at time t. ψ_t denotes the potential functions.

layers followed by the output layer. Considering $w_c^{y_c}$ as the neural net weights of output layer and w_{NN} as the weights of other layers, the conditional probability implemented by a neuroCRF takes the form,

$$p(y|x) \propto \prod_{c \epsilon C} e^{-E_c(x,y_c,w)} = \prod_{c \epsilon C} e^{\{w_c^{y_c}, \phi_c(x,w_{NN})\}} \tag{2.7}$$

where $\phi_c(x, w_{NN})$ is the high-level representation of the input x at clique c.

Another form of neuroCRF is the linear chain neuroCRF by Do et al., and it is shown in Fig 2.4 which is based on first-order Markov chain structure. There are two potential functions specified in this neuroCRF based on the cliques, $\psi_t(x, y_t)$ is the potential function of *local cliques* at each position t and $\psi_{t-1,t}(x, y_{t-1}, y_t)$ is the potential function of *transition cliques* between two successive positions at $t - 1$ and t respectively. Some implementations in sequence tagging tasks include LSTM-based neuroCRFs in an NER task [60], sequence classification and structured prediction performed using linear-chain neuroCRFs in [12].

2.2.4 CRF Models and Applications

CRFs are successfully applied in many applications of NLP. Some of them include a supervised CRF to identify disorder named entities in medical records [16]. CRFs are also used in context based NER system [54]. In bio-informatics, they are used in variety of tasks like Critical Assessment of Information Extraction Systems in Biology (BioCreAtIve) [34] that involves identification of biomedical entities such as genes and proteins in medical databases like PubMed [6], FlyBase [2], Mouse Genome Informatics [4], Saccharomyces Genome Database [7], and Swiss-Prot [8].

2.3 De-identification and Related Models

De-identification, as mentioned earlier, involves redacting or masking the PHI data in a medical record/EHR such that the patient's identity is protected. Any data belonging to any of the categories specified in Table 1.3 must be removed or replaced by annotations/labels. De-identification facilitates research and sharing of medical records without any compromise to the confidential patient health information. Initially, de-identification involved human annotators but later with an increase in the volume of medical records and adoption of EHRs, human annotation has become a tedious task resulting in automated de-identification systems.

De-identification and its implementation approaches have been an important research topic for over twenty years. Traditional approaches to automated systems are similar to rule-based NER systems that use dictionaries, gazetteers of names, cities, and hospitals. In the year 1996, Sweeney et al. proposed the first automated de-identification system, a rule-based approach, for identifying twenty five types of PHI entities in EHRs [75]. The algorithms in this model relied on lists, that comprised a list of first names, area codes, and cities, for identifying the entities assigned to each algorithm. The system used a scrubbed subset of pediatric medical record system as its data-set. PhysioNet's DEID tool [27, 57] is a rule-based de-identification system which uses dictionaries and regular expressions for de-identifying the records. As a result of the limitations in a rule-based approach, mentioned in Section 2.1.2, systems based on machine learning techniques became popular. Many challenges have been conducted by the informatics and computer science research groups that encouraged the usage of ML-based text-processing techniques for implementing de-identification systems.

The *Informatics for Integrating Biology and the Beside* (I2b2) conducted its first de-identification challenge in the year 2006 [79], which marks the first attempt for de-identifying eight PHI entities (Patients, Doctors, Hospitals, IDs, Dates, Locations, Phone numbers, and Age). The challenging task reviewed the submitted systems: Aramaki et al. proposed a CRF based implementation [13] that uses sentential features which prove to be more important for categories like dates, IDs, and patients; Guillen [29] developed a rule-based model which utilizes global features (sentence positions), local features (lexical cues, special characters and format patterns) and syntactic features for de-identification; Support Vector Machine (SVM) based implementation [31] specifying the importance of context-based learning for patient names; a hybrid system based on rules with SVM [33]; an iterative approach using decision trees with local features and dictionaries [76]; Wellner [83] proposed three variants namely Wellner 2 using LingPipe [3] which is based on an implementation of HMM, Wellner 1 and Wellner 3 using Carafe engine from *MITRE Identification Scrubber Toolkit* (MIST) tool [11]. Following this, I2b2 conducted a similar challenge task in the year 2014 for the de-identification of longitudinal clinical narratives [70, 71]. Following this implementation, a wide range of automated de-identification systems have been introduced based on the type of medical records to be de-identified like discharge summaries [50, 78], nursing progress notes [57], pathology reports [77] and mental health records [26]. This thesis discusses two open source de-identification systems: MITRE MIST [11] and PhysioNet DE-ID tool [27, 57]. The design, implementation and training modes of these models are discussed in Chapter 3.

Chapter 3

Models

This thesis reviews the performance of statistical models and rule-based systems for de-identifying medical records. In my work, I have used a machine learning based model, from The MITRE Corporation. The model is suitable for research because it provides a platform for exploring various experimental variants of its core model for the task of de-identification and it is an open source system. In contrast to a machine learning model, I have also evaluated a rule-based model from PhysioNet, which is a popular de-identification tool that has been used by several works for a comparative study. I present the results and challenges of all mentioned models on three different types of medical records. The rest of the chapter is organized to discuss the design and implementation of the models evaluated in my work.

3.1 MITRE MIST

MITRE Identification Scrubber Toolkit (MIST) is a result of combined research from the MITRE Corporation at the I2b2 medical data de-identification challenge along with Vanderbilt University Medical Center and the University of Michigan. Their research focused on lowering the cost of healthcare practitioners to perform de-identification of medical records. MIST has become a free and open-source tool which can be used for identifying and redacting PHI information in free-text medical records. It provides a platform to build tools that can be used to de-identify records. The MIST toolkit contains five modules. A web-based graphical annotation tool that provides the user interface to load and hand annotate the records. The second is a training module which allows training a model on given annotated data. The third is a tagging module which performs automatic tagging with the help of the trained model. The fourth module is the redaction and resynthesis module that allows redaction of the annotated files. The fifth module is the experiment engine which evaluates the model on new data. The workflow in MIST is described in Fig 3.1.

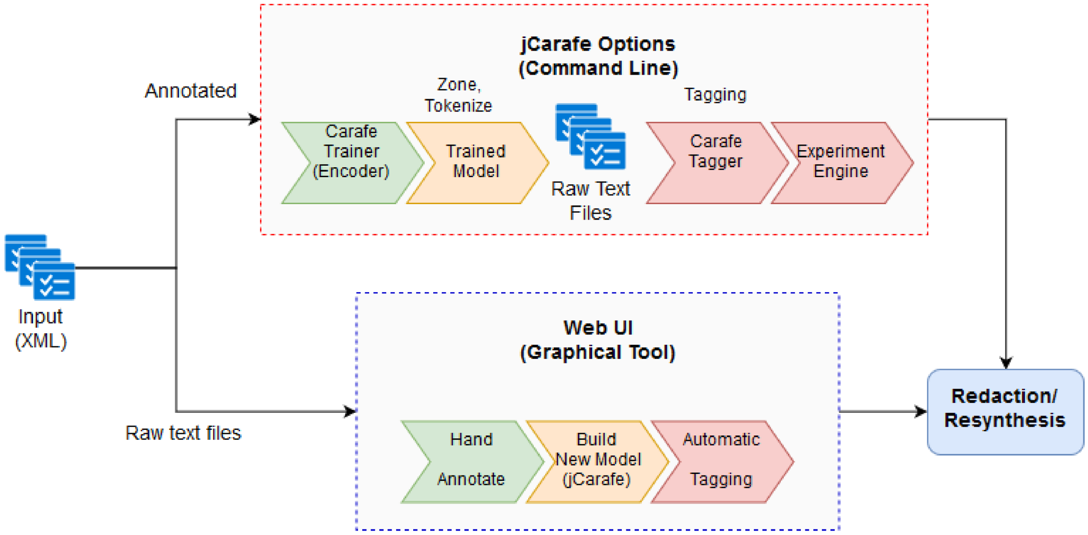


Figure 3.1: Workflow and tasks as per the modules available in MIST. Encoder performs training of model and Decoder (tagger) performs automatic tagging.

My work focuses on building CRF models by customizing MIST carafe engine through the command line options. Building a de-identification model in MIST requires determining the mode of training, the type of input records for the model to learn, and extracting features required for sequence labeling. MIST implements de-identification as a task in the *MIST Annotation Toolkit* (MAT). MAT is the suite of tools in MIST, that is responsible for automatic tagging of span annotations in medical records. There are two main subtasks associated with the de-identification task:

1. **Annotation subtask** responsible for training the models to identify the PHI phrases like “NAME” or “HOSPITAL”, which are named entities representing the PHI categories.

2. **Replacement subtask** replaces the PHI phrases with obscuring fillers like <NAME> and <HOSPITAL> based on the PHI categories used in the records. For example, the sentence, “John was diagnosed with a brain tumor.” becomes “<PATIENT> was diagnosed with a brain tumor.” after de-identification.

The annotation subtask is carried out with the help of MAT toolkit, which consists of highly customizable NLP based tools such as Carafe CRF system (encoder and decoder), web-based hand annotation tool and command-line-based process workflow engine for multi-step tagging. MAT accepts input records in XML/SGML format, for training, where the annotations are represented as XML elements wrapped around the PHI phrases. In Fig 3.1, the jCarafe Options block describes the step by step task responsible for de-identification task. The *Trained Model* initiates the zoning and tokenize activities. Following this, the *Carafe Tagger* performs the tagging activity (annotating the documents). The commonly used step activities in a de-identification workflow are:

1. **zone**: it is a concatenation of zoning and tokenizing, in this step the word boundaries are identified from the region to be processed.

2. **tokenize**: generate tokens (words) of the documents for the model to annotate. jCarafe accepts these tokens as features to train the MIST models.

3. **tag**: add the PHI annotations.

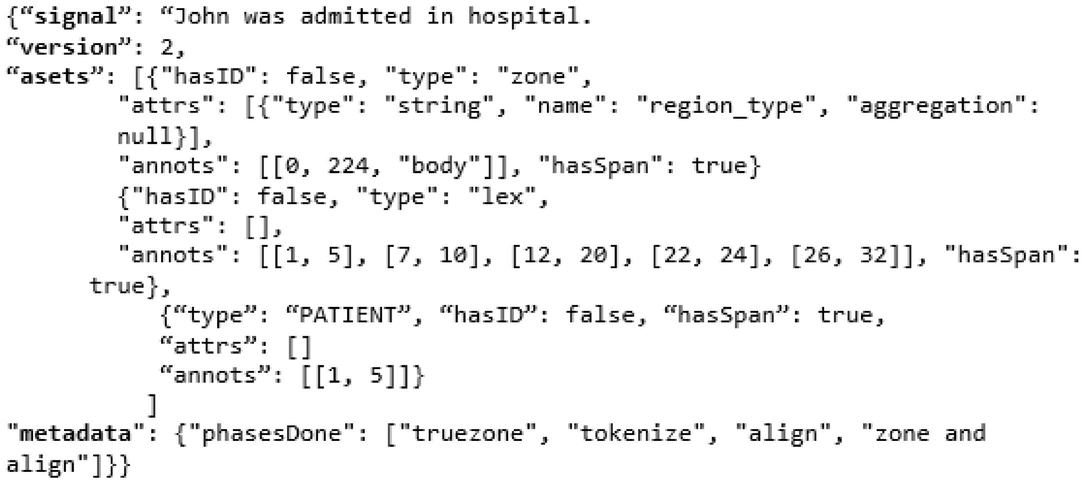
4. **nominate**: choose appropriate replacements or fillers for the PHI phrases.

5. **transform**: the transformation of medical records to new de-identified records.

For example, in the sentence, "<PATIENT>John</PATIENT> was admitted to hospital.", the XML element <PATIENT> is an annotation that specifies the PHI category. Once the input files are pre-processed to find the sentence and word boundaries by zoning, then tokenization produces the tokens to which the tags are added. The result looks like:

> <PATIENT>John</PATIENT> <lex>was</lex> <lex>admitted</lex>
> <lex>in</lex> <lex>hospital</lex>

The input XML files are converted to JavaScript Object Notation (JSON) format which is depicted in Fig 3.2. The annotations are anchored to a particular span of the document. That is, annotations are marked in the region of the document to be de-identified and the region itself is specified as a type of annotation. Each annotation in the JSON specifies the beginning and ending character position of every token in the text grouped according to the type of annotation. In the Fig 3.2, the annotation index value [1, 5] indicates the start and end positions of the token "John" with annotation type as *PATIENT*. The tokens along with other extracted features are used for training the MIST model. The learned model is then used to annotate/tag new records by nominating the appropriate label with the maximum score to each token.



```
{“signal”: “John was admitted in hospital.
“version”: 2,
“asets”: [{"hasID": false, "type": "zone",
        "attrs": [{"type": "string", "name": "region_type", "aggregation":
        null}],
        "annots": [[0, 224, "body"]], "hasSpan": true}
        {"hasID": false, "type": "lex",
        "attrs": [],
        "annots": [[1, 5], [7, 10], [12, 20], [22, 24], [26, 32]], "hasSpan":
      true},
           {“type”: “PATIENT”, “hasID”: false, “hasSpan”: true,
           “attrs”: []
           “annots”: [[1, 5]]}
          ]
"metadata": {"phasesDone": ["truezone", "tokenize", "align", "zone and
align"]}}
```

Figure 3.2: MAT-JSON format as described in [11].

The de-identification system in MIST supports two specializations of annotation tasks namely HIPAA and AMIA de-identification tasks. HIPAA task is the default general-purpose de-identification task provided by MIST. The tags covered by this task includes all PHI categories as governed by the HIPAA Privacy Rule mentioned in Table 1.3. AMIA task is a simple variant of the general HIPAA task, introduced as part of the medical data de-identification challenge conducted by I2b2. Throughout this thesis, AMIA de-identification task is used for evaluations. The choice of the task to be used is configured using a task file (XML file) available in the MAT toolkit. More about the task and its annotations are discussed under Chapter 4. The rest of this section is organized to discuss the core engine in MIST, feature extractions, the supported training modes and finally the performance metrics of the models.

3.1.1 jCarafe

The model at the core of MIST, which performs automatic de-identification, is the jCarafe [82]. jCarafe implements a CRF classifier for facilitating sequence modeling. CRFs provide the machine learning framework for learning model parameters from annotated data, for tasks like entity extraction, de-identification and other text-processing applications. As seen earlier in Chapter 2, CRFs are particularly well-suited to de-identification because they can incorporate sequential information when making decisions about multi-word instances of PHI.

Table 3.1: Default feature functions in AMIA De-identification task

Feature Function	**Description**
wdFn	word at a current position
caselessWdFn	lowercase word at a current position
prefixFn(integer)	prefixes from length 1 to specified range
suffixFn(integer)	suffixes from length 1 to specified range
regexpFn(name, regexp)	name of the regex which is satisfied by the token
lexFn	name of lexicon if the token is present in the lexicon

MIST uses jCarafe[1], the latest version of carafe engine which runs on Java Virtual Machine (JVM), for de-identification. The underlying algorithms used in the jCarafe engine supports a variety of tasks including part-of-speech tagging [44], shallow parsing [67], discourse parsing [72], summarization [68] and other text processing techniques. In the MIST Annotation Toolkit, the carafe engine provides a CRF-based sequence tagger, which is the default tagging and training module in MIST. The jCarafe engine can be configured through command line flags to utilize various training modes supported by MIST. Based on the training mode, jCarafe takes three different forms. First is the MIST CRF which is the default model that uses the generic Linear Chain CRFs [47]. Second is the MIST neural-CRF which implements a neuroCRF [25]. The third is the MIST semi-Markov CRF which implements a semi-CRF [65]. The training results of the jCarafe are discussed under upcoming subsections.

3.1.2 Feature Extraction

By default, the tokens with tags <lex> have associated attribute value pair and MIST uses the attribute value pairs as features for the trainer. The basic input format of

[1] MITRE jCarafe: `http://mist-deid.sourceforge.net/current_docs/html/index.html`

features to the jCarafe is line based. The following is a sample representation of the attribute value pairs.

<sequence label> <feature1> <feature2>.... <featureN>

where each <feature_i> takes the form:<name>:<value>. The <name> denotes the feature name and <value> is the value associated with the feature. Most often the values are 1 or 0 based on whether the feature is present or not. If the features are not binary, a scalar value is supplied to the features. These features are sent to the CRF learning framework within the jCarafe engine of MIST for training purposes. Specifying features in the MIST is done manually through a feature specification file (fspec). The features are often driven by linguistic intuition and domain knowledge. In the jCarafe engine, the features are extracted by applying each feature specification at every position within the sequence of text. MIST supports a handpicked list of built-in feature functions that are responsible for extraction of the features.

MIST provides a default feature specification for each form of the de-identification task. The default list of feature functions used in AMIA de-identification task is given in Table 3.1. All the feature functions specified in Table 3.1 are atomic features. These features provide low-level building blocks for more complicated features. A list of regular expressions used as feature functions in my work is included in Appendix A.1. The atomic feature functions can be used along with *higher-order* feature functions to provide more features related to the context of a word. Functions like *over*, *ngram* and *cross* are higher-order feature specifications in MIST. Two higher-order functions namely *over* and *ngram* are used in my work. The *over* function is specified by passing the offset values as arguments along with an atomic feature function. In the following example, the atomic function *wdFn* is used with a higher-order feature function *over*.

word_unigrams as wdFn over (-1, 1);

The offset values that are passed as arguments specify that the atomic function must be applied to the previous position -1 and the subsequent position 1. The atomic feature function *wdFn* returns all words at the relative positions within the offset range and uses these values along with the offset as features. The *ngram* function conjoins all the features extracted from the relative positions in offsets to form a single feature. When a wdFn ngram(1, 2) is applied at the position of the word “Scott” in the sentence “Audrey Scott said hello”, then the function returns said_hello(1, 2) as the feature. The features specified in the Table 3.1 are common to all variants of CRF used in jCarafe. There are also feature functions that are specific to particular variants of CRFs such as semi-CRF and neural features. In every variant of CRF, the features are explicitly extracted using the feature specification file. The specification file contains the list of feature functions used for every variant and these functions extract the list of features for the given input data-set. The extracted features are temporarily saved as attribute value pairs as discussed earlier. Most of these features functions were provided by the MIST toolkit.

Semi-CRF Features

Similar to the neural features, semi-CRF features are used in the models in which the jCarafe engine is trained using the semi-CRF variant. MIST supports only a few built-in semi-CRF features and they are mentioned in Table 3.2. These features are experimental in MIST and have not yet been discussed widely. The efficiency of these features and the performance of the models trained on these features are discussed in this thesis.

Table 3.2: Semi-CRF feature functions.

Feature Function	**Description**
semiEdgeFn	Bias feature for individual segment
semiNodeFn	Bias feature for label pair of current and previous segment
phraseFn	n-gram/word is added as feature
phraseWds	For every term appearing in a current segment, single feature is added

Neural Features

Neural features are used when MIST uses the neural CRF[2]. Neural features are fed as input to the hidden nodes and other features are fed to the CRF as standard features. MIST recommends using of lexical and part-of-speech attributes as neural features because they tend to be smaller in number and therefore the number of model

[2] MITRE jCarafe, Neural CRF: http://mist-deid.sourceforge.net/current_docs/html/index.html

parameters required is also smaller. The only neural feature used in my experiments is the lexicon-based feature function and others remain as standard CRF input features generated by the atomic feature functions. Features are designated as neural by adding the keyword 'NEURAL' by the side of every feature function:

word_Neural as lexFn NEURAL;

According to the MIST documentation, neural features are experimental and my work has evaluated models with neural features to demonstrate the importance of Artificial Neural Network (ANN) layers in a sequence labeling problem like de-identification. The neural network layers help in discovering the non-linear properties of the input, unlike traditional CRFs.

3.1.3 Training

As mentioned earlier, MIST provides the platform to train a model using the command line options. Based on the variant used in jCarafe, MIST models can be trained to use different variants of CRF, giving new models and new parameters for every variant. Training a model involves estimating the parameters for their optimal values. The number of parameters, which can be manually altered, vary based on the variant of CRF used for training. The list of parameters that can be manually altered in each variant of CRF is tabulated in Table 3.3.

Table 3.3: Parameters of each variant of CRF.

MIST CRF	**MIST neural CRF**	**MIST semi-CRF**
Learning rate	Learning rate	Learning rate
No of iterations	Hidden gates	No of iterations
	Hidden learning rate	
	No of iterations	

The default training method in jCarafe tries to maximize the conditional log-likelihood (CLL) of the data. Gradient-based algorithms are usually used for maximizing the log-likelihood. MIST uses one of the gradient-based methods, stochastic gradient descent (SGD) [18] for estimating the parameters. SGD computes gradients for a small set of training data instead of computing the CLL and its exact gradient on the entire data-set. Based on the computed local gradient for a small part of the training data, the parameters are updated according to a learning rate that decays over a time. The limitation here is that SGD can take much a longer time to converge to optimal values with an initial learning rate. MIST overcomes this problem with the help of Periodic Step-size Adaptation (PSA) [38]. In PSA the learning rates are adjusted for individual parameters based on the history of updates for each parameter. Adopting SGD with PSA is optional during training. In my work, I used PSA based SGD training because it speeds up the training and uses fewer iterations for training a model.

3.1.4 Performance Metrics

MIST provides estimated variants of Precision, Recall and F1-measure to assess the performance of a trained model. These measures are calculated using the MAT Scorer which uses a common scoring algorithm. The scoring algorithm in MAT makes use of similarity profiles to create the metrics for annotations. Similarity profiles are declarative descriptions on how to compare labels based on the dimensions of the annotations and the way in which the dimensions are compared. MAT scorer uses Kuhn-Munkres bipartite set algorithm [51], in a stratified manner to compare and calculate the similarity score of the annotations. For example, in the reference record, “Steven was a physician at ABC Hospital.”, there is a *DOCTOR* annotation over characters 0 - 6 with nomtype = PRO. But in the corresponding test record it is labelled as *PATIENT* over the characters 0 - 5 with nomtype = PRO. Consider the following is the similarity profile used for calculating the similarity score for the token “Steven”.

```
<similarity_profile>
<tag_profile true_labels=“PATIENT,DOCTOR”>
<dimension name=“_label” weight=“2”/>
<dimension name=“_span” method=“overlap” weight=“8” overlap_low=“0.8”/>
<dimension name=“nomtype” weight=“1”/>
</tag_profile>
</similarity_profile>
```

The dimensions, specified in the similarity profile, are checked for a match and based on that the weights are added. In the above example profile, the dimension *_label* contributes a 0 out of a maximum weight of 2 as the labels don't match. Similarly, the *_span* contributes the maximum weight 0.8 as there is a span overlap in 5 out of 6 characters and the *_nomtype* dimension contributes a 1 out of 1. Based on the collected weights the similarity score is calculated as $(0*2)+(8*8)+(1*1)/(2+8+1)$, which equals to 0.59. The default similarity profile used throughout the scoring used in this thesis is:

```
<dimension name=”_label” weight=”0.1” true_residue=”0.5”/>
<dimension name=”_span” weight=”0.9”/>
<dimension name=”_nonannotation_attribute_remainder” weight=”0.1”/>
<dimension name=”_annotation_attribute_remainder” weight=”0.1”/>
```

In MAT Scorer, the true positives or a match is the set of annotations with true labels whose pairs (occurring in test and reference) have a perfect similarity score of 1.0. False negatives or clashes are those annotations whose labels do not have a perfect similarity score or less than 1.0. And any annotations that cannot be paired are considered missing or spurious. The following measures are reported in this thesis for every model:

- **Precision (P)**: Number of true positives divided by the number of tokens labeled;
- **Recall (R)**: Number of true positives divided by the number of true positives and false negatives;
- **F1 Measure**: F1 measure is the harmonic mean of precision and recall computed as $2 * ((P * R)/(P + R))$.

3.2 DE-ID

PhysioNet DE-ID [27, 57] is a rule-based de-identification system, produced as part of a study at the Harvard-MIT Division of Health Science and Technology. DE-ID is written in Perl and uses lexical look-up tables, regular expressions, and simple heuristics to locate HIPAA-defined PHI terms and other extended PHI instances such as dates. The software supports de-identification of names, locations, hospital names, town/city names, street addresses, zip codes, PO Boxes, dates, telephone/fax/pager, patient and doctor names, identification numbers, email addresses, URLs and ages above 89 years of age. There are four types of look-up dictionaries in DE-ID [57]:

1. *Known PHI* look-up tables for known names.
2. *Potential PHI* look-up table for common names or women and men, last names, locations, and states.
3. *PHI indicator* look-up tables which contain terms which are prefixes or suffixes of PHI terms. Examples include ”Dr.”, ”hospital”, ”age”, ”street” and so on.
4. *Non-PHI* look-up tables that have common English words which are taken from Atkinson's Spell Checking Oriented Word Lists [1] which are non-PHI terms.

Apart from the dictionaries mentioned above, additional dictionaries can be included or removed according to the domain of the data-set involved. The list of regular expressions, look-up tables and their descriptions used in this software are mentioned under Appendix A.2. DE-ID comes with nursing notes drawn from the MIMIC II clinical database [49], intensive care unit (ICU) nursing notes for evaluating the software. The structure and format details about the data-sets are described in Chapter 4.

3.2.1 Algorithm Overview

The algorithm behind the DE-ID software scans the input records line by line and then divides them into individual words separated by whitespaces. Then the following steps are carried out on the words/tokens:

1. The algorithm first performs a lexical match with all the non-numerical tokens to identify known and potential PHI terms.
2. It performs a pattern match using the regular expressions to find the named entities.

3. It determines the ambiguity in the identified entities using simple heuristics.
4. Finally, the algorithm replaces identified PHI terms with tags that denote the category of PHI.

3.2.2 Performance Metrics

The performance metrics in DE-ID is calculated by comparing the locations of the PHI terms in both the test and the PhysioNet documents for the corresponding test records. The results are reported as the following measures:

- Positive Predictive Value (PPV)/ Precision: Number of true positives divided by the total number of tokens labeled.
- Sensitivity/Recall: Number of true positives divided by sum of true positives and false negatives.

The above mentioned standard metrics in NLP: Precision, Recall, and F1 measure are used to report the statistics of the models at a token level for MIST and DE-ID. In a problem like de-identification, recall is of more importance. This is because labeling a PHI term as 'OTHER' (false negatives) might lead to unauthorized disclosure of PHI terms. Therefore it is important that a model yields a better recall. On the other hand, for a medical research, it is important that the model does not label any medical term as PHI (false positives), in which case there is a loss of medical information. My work discusses the important aspects in dealing with such models and their parameters, which is a difficult task from a practitioner's perspective, to attain better results that generates de-identified samples that are most suitable for research purposes.

Chapter 4

Data and Annotations

Data-sets used in this thesis, for evaluation, are manually deidentified samples of clinical notes, suitable for research. Medical records are available in various formats based on the type and purpose of the medical records. My work deals with free-text medical notes. There are different types of medical notes such as discharge summaries, nursing notes, laboratory reports and physician notes. A sample format of discharge summary notes is shown in Fig 4.1. It can be observed that these notes follow a certain template like <Heading>:<Content>. Some clinical notes do not have any such templates but are simple, continuous free text. In a sequence tagging problem, each term in the text is considered as a lexical token.

Table 4.1: Distribution of records in the data-sets. A record denotes a single file that contains medical details of a patient.

Data-set	**Total**	**Training**	**Test**
I2b2 2006	889	668	221
I2b2 2014	1304	790	514
PhysioNet Corpus	2434	1703	731

A de-identification model should be able to perform well on any given type of medical notes. Evaluating models on different types of clinical notes will help in analyzing the features required for training models. I have used three types of medical notes: discharge summaries, longitudinal medical records and nursing notes for the evaluations. Two data-sets are taken from I2b2 NLP data-sets[1]. The third data-set (nursing notes) is from PhysioNet DE-ID[2] package. Each data-set contains individual files of a patient. There are multiple files for each patient that has information about the patient's medical history. Records in all three data-sets are divided into training

[1] I2b2 NLP Data-sets: `https://www.I2b2.org/NLP/DataSets/Main.php`

[2] PhysioNet Corpus: `https://www.physionet.org/physiotools/deid/#data`

and test subsets. The composition of the subsets from each data-set is tabulated and shown in Table 4.1. The records have been divided into ten batches of the train (70%) and test (30%) subsets for performing 10-fold cross-validation[3]. The revised split up of training and test subsets (shuffled randomly) is shown in Table 4.2.

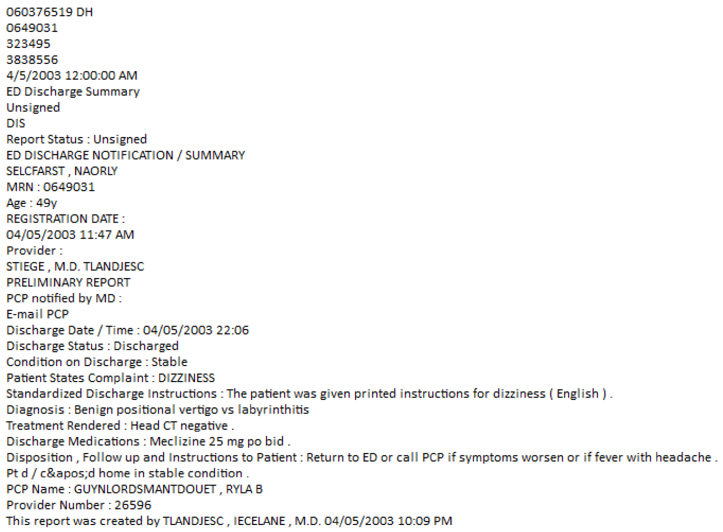
Table 4.2: Distribution of PHI and Non-PHI instances across each data-set. Records are grouped as training and test subsets.

Data-set	**Training**			**Test**		
	Records	**PHI**	**Non-PHI**	**Records**	**PHI**	**Non-PHI**
I2b2 2006	620	13648	425371	267	5795	176345
I2b2 2014	912	18796	728937	392	8674	311200
PhysioNet Corpus	1703	1243	311541	731	525	131853

4.1 Informatics for Integrating Biology and the Beside (I2b2) Data

As mentioned above, two major data-sets are taken from I2b2 NLP DataSets. I2b2 has provided a number of fully deidentified clinical records from the Research Patient Data Repository at Partners HealthCare for enhancing the ability of NLP tools to extract useful information from clinical records. Dr.Ozlem Uzuner [79] conducted many NLP challenges using a wide range of deidentified medical records and these deidentified notes are available to the research community for research purposes. The I2b2 2006 De-identification and Smoking challenge data-set [79] comprises a total of 889 de-identified discharge summaries. Of this total records, two records have been removed due to their inconsistent behaviour during preprocessing. Table 4.2 shows the training and test split made on these records. The records are XML files with inline attributes, as shown in Fig 4.2, defining the tags (e.g.<PHI TYPE="PATIENT">Susan Smith</PHI>) where $PATIENT$ denotes the annotation class label. Each term in the name "Susan Smith" is a lexical token. Every such occurrence is marked as a PHI instance and the total number of PHI instances present in this data-set is shown in Table 4.2 under the PHI column.

[3] Cross-validation: `https://en.wikipedia.org/w/index.php?title=Cross-validation_(statistics)&oldid=846520045`



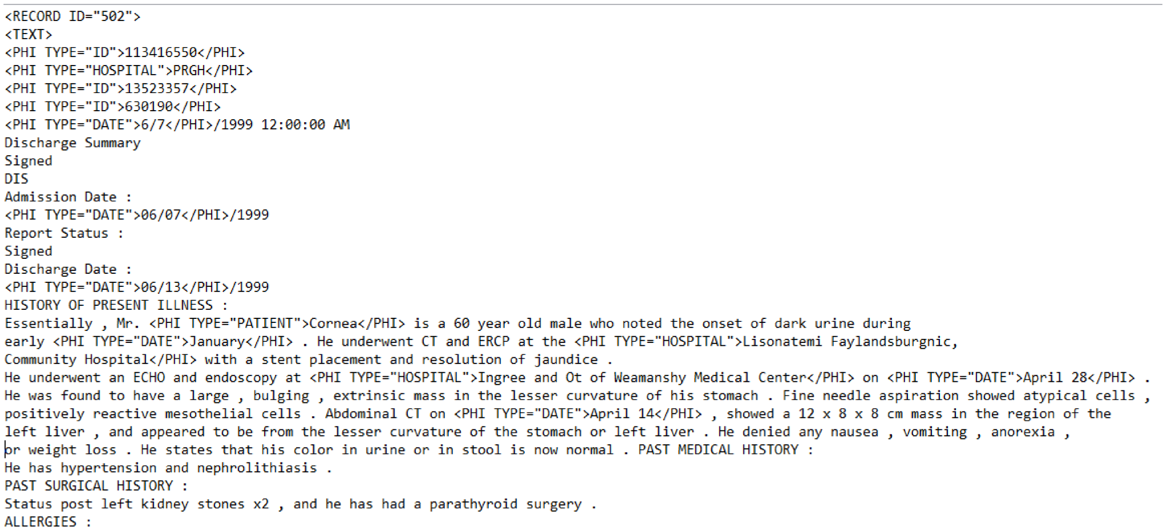
060376519 DH
0649031
323495
3838556
4/5/2003 12:00:00 AM
ED Discharge Summary
Unsigned
DIS
Report Status : Unsigned
ED DISCHARGE NOTIFICATION / SUMMARY
SELCFARST , NAORLY
MRN : 0649031
Age : 49y
REGISTRATION DATE :
04/05/2003 11:47 AM
Provider :
STIEGE , M.D. TLANDJESC
PRELIMINARY REPORT
PCP notified by MD :
E-mail PCP
Discharge Date / Time : 04/05/2003 22:06
Discharge Status : Discharged
Condition on Discharge : Stable
Patient States Complaint : DIZZINESS
Standardized Discharge Instructions : The patient was given printed instructions for dizziness (English) .
Diagnosis : Benign positional vertigo vs labyrinthitis
Treatment Rendered : Head CT negative .
Discharge Medications : Meclizine 25 mg po bid .
Disposition , Follow up and Instructions to Patient : Return to ED or call PCP if symptoms worsen or if fever with headache .
Pt d / c&apos;d home in stable condition .
PCP Name : GUYNLORDSMANTDOUET , RYLA B
Provider Number : 26596
This report was created by TLANDJESC , IECELANE , M.D. 04/05/2003 10:09 PM

Figure 4.1: Discharge Summary Record. Notes are free-text with PHI information.

The second data-set is the I2b2 2014 De-identification and Heart Disease Risk Factors Challenge [70, 71]. It contains longitudinal records for 301 diabetic patients, with two to five records for each patient. There are a total of 1,304 records which are randomly split into 912 training and 392 test records. The training records have 18,796 PHI instances and 8,674 test PHI instances. The format of this data-set is shown in Fig 4.3, unlike I2b2 2006, in every record, the annotation labels are provided separately under the XML element <TAGS>. An initial preparation script from the AMIA package of the MIST tool [11] is run on the I2b2 data-set to make it conform to the requirements of MIST AMIA de-identification task. For evaluation purposes, all records are converted to a single standard document format (MAT-JSON) [11] taken from the MIST toolkit. All clinical notes are converted to JSON files like the example given in Fig 3.2.

4.2 PhysioNet - The Gold-standard Corpus

The third data-set used in this thesis is taken from the *PhysioNet De-Identification Software* (DE-ID) PhysioNet Corpus of deidentified medical texts [27]. It contains a total of 2,434 nursing notes, deidentified and reviewed by at least three experts and have been used in developing a variety of automated methods. The count of training and test samples is shown in Table 4.2. It can be observed that this corpus has a fairly



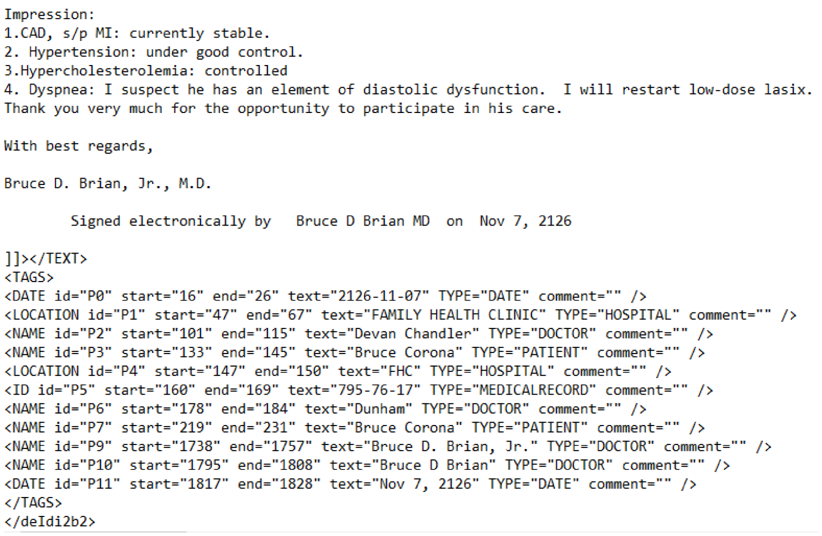
```
<RECORD ID="502">
<TEXT>
<PHI TYPE="ID">113416550</PHI>
<PHI TYPE="HOSPITAL">PRGH</PHI>
<PHI TYPE="ID">13523357</PHI>
<PHI TYPE="ID">630190</PHI>
<PHI TYPE="DATE">6/7</PHI>/1999 12:00:00 AM
Discharge Summary
Signed
DIS
Admission Date :
<PHI TYPE="DATE">06/07</PHI>/1999
Report Status :
Signed
Discharge Date :
<PHI TYPE="DATE">06/13</PHI>/1999
HISTORY OF PRESENT ILLNESS :
Essentially , Mr. <PHI TYPE="PATIENT">Cornea</PHI> is a 60 year old male who noted the onset of dark urine during
early <PHI TYPE="DATE">January</PHI> . He underwent CT and ERCP at the <PHI TYPE="HOSPITAL">Lisonatemi Faylandsburgnic,
Community Hospital</PHI> with a stent placement and resolution of jaundice .
He underwent an ECHO and endoscopy at <PHI TYPE="HOSPITAL">Ingree and Ot of Weamanshy Medical Center</PHI> on <PHI TYPE="DATE">April 28</PHI> .
He was found to have a large , bulging , extrinsic mass in the lesser curvature of his stomach . Fine needle aspiration showed atypical cells ,
positively reactive mesothelial cells . Abdominal CT on <PHI TYPE="DATE">April 14</PHI> , showed a 12 x 8 x 8 cm mass in the region of the
left liver , and appeared to be from the lesser curvature of the stomach or left liver . He denied any nausea , vomiting , anorexia ,
or weight loss . He states that his color in urine or in stool is now normal . PAST MEDICAL HISTORY :
He has hypertension and nephrolithiasis .
PAST SURGICAL HISTORY :
Status post left kidney stones x2 , and he has had a parathyroid surgery .
ALLERGIES :
```

Figure 4.2: Format of an annotated discharge summary record taken from I2b2 2006 Data-set [79].

large number of training records, but note that among the three data-sets, PhysioNet corpus has fewer PHI instances. As specified in [27], all instances of PHI found in these clinical records are replaced by realistic surrogate data, ensuring the privacy of the patient's identity while providing realistic input to a deidentification model. Fig 4.4 shows the format of a single record from the unannotated clinical notes. In this data-set, the PHI annotations are provided as a separate text file that contains the index positions of every token followed by its respective PHI label. The files are preprocessed to match the MIST requirements for training the model and performing automatic tagging.

4.3 Annotations

Annotations represent the labels (PHI categories) that differentiate each token in a clinical note, whether it is a PHI or non-PHI token. The standard annotations correspond to the name of the PHI category as mentioned in Table 1.3. But none of the three data-sets contain all 18 categories of PHI. The I2b2 data used for the challenge task [79] contained only 8 PHI categories and so MIST created the AMIA De-identification task based on these 8 major PHI categories. The set of PHI categories used in AMIA is a subset of PHI categories mentioned in Table 1.3. For the evaluation, the base reference annotations are taken from the AMIA De-identification



```
Impression:
1.CAD, s/p MI: currently stable.
2. Hypertension: under good control.
3.Hypercholesterolemia: controlled
4. Dyspnea: I suspect he has an element of diastolic dysfunction.  I will restart low-dose lasix.
Thank you very much for the opportunity to participate in his care.

With best regards,

Bruce D. Brian, Jr., M.D.

        Signed electronically by   Bruce D Brian MD  on  Nov 7, 2126

]]></TEXT>
<TAGS>
<DATE id="P0" start="16" end="26" text="2126-11-07" TYPE="DATE" comment="" />
<LOCATION id="P1" start="47" end="67" text="FAMILY HEALTH CLINIC" TYPE="HOSPITAL" comment="" />
<NAME id="P2" start="101" end="115" text="Devan Chandler" TYPE="DOCTOR" comment="" />
<NAME id="P3" start="133" end="145" text="Bruce Corona" TYPE="PATIENT" comment="" />
<LOCATION id="P4" start="147" end="150" text="FHC" TYPE="HOSPITAL" comment="" />
<ID id="P5" start="160" end="169" text="795-76-17" TYPE="MEDICALRECORD" comment="" />
<NAME id="P6" start="178" end="184" text="Dunham" TYPE="DOCTOR" comment="" />
<NAME id="P7" start="219" end="231" text="Bruce Corona" TYPE="PATIENT" comment="" />
<NAME id="P9" start="1738" end="1757" text="Bruce D. Brian, Jr." TYPE="DOCTOR" comment="" />
<NAME id="P10" start="1795" end="1808" text="Bruce D Brian" TYPE="DOCTOR" comment="" />
<DATE id="P11" start="1817" end="1828" text="Nov 7, 2126" TYPE="DATE" comment="" />
</TAGS>
</deIdi2b2>
```

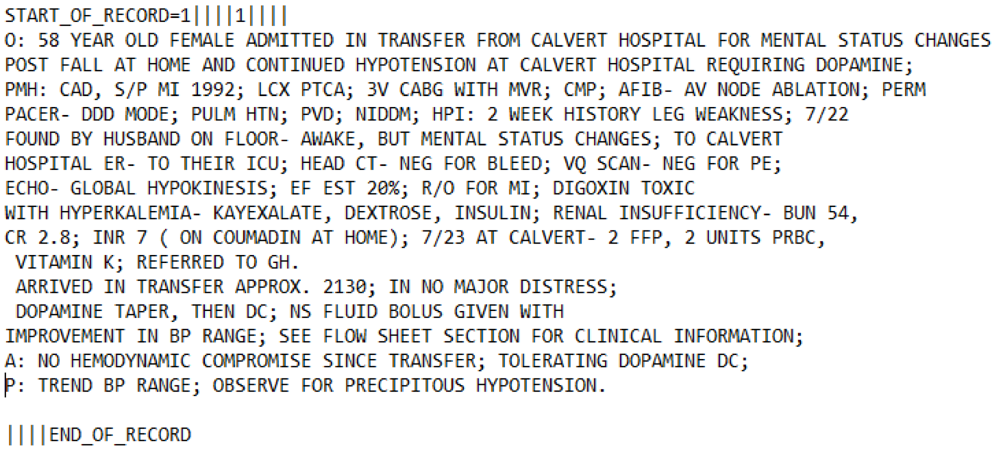
Figure 4.3: Format of an annotated longitudinal medical narrative record from the I2b2 2014 Challenge Data-set [70, 71].

Task in MIST. The task covers the list of PHI categories that all three data-sets contain. As specified in Chapter 3, I have used the AMIA De-identification task for all evaluations. Task specific configurations are made through an XML task file from the command line. Table 4.3 shows the list of annotations present in the AMIA task and the three data-sets.

Table 4.3: Annotations across three data-sets with reference to the labels used in AMIA De-Identification Task in MIST. PHI categories used in AMIA is a subset of the main table 1.3. Labels are represented in the format as per the respective data-set.

AMIA Task	**I2b2 2006**	**I2b2 2014**	**PhysioNet Corpus**
ID	<PHI TYPE=ID>	<ID>	NA
Age	<PHI TYPE=AGE>	<AGE>	Age
Date	<PHI TYPE=DATE>	<DATE>	Date
Hospital	<PHI TYPE=HOSPITAL>	<LOCATION TYPE=HOSPITAL>	Location
Patient	<PHI TYPE=PATIENT>	<NAME TYPE=PATIENT>	Patient
Doctor	<PHI TYPE=DOCTOR>	<NAME TYPE=DOCTOR>	Doctor
Location	<PHI TYPE=LOCATION>	<LOCATION>	Location
Phone	<PHI TYPE=PHONE>	<PHONE>	Phone

With reference to the annotations in the AMIA task, the annotations in I2b2 2014 and PhysioNet corpus are preprocessed to match the labels. By Table 4.3, the PhysioNet corpus has no tokens associated with the label ID and hence this label is



```
START_OF_RECORD=1||||1||||
O: 58 YEAR OLD FEMALE ADMITTED IN TRANSFER FROM CALVERT HOSPITAL FOR MENTAL STATUS CHANGES
POST FALL AT HOME AND CONTINUED HYPOTENSION AT CALVERT HOSPITAL REQUIRING DOPAMINE;
PMH: CAD, S/P MI 1992; LCX PTCA; 3V CABG WITH MVR; CMP; AFIB- AV NODE ABLATION; PERM
PACER- DDD MODE; PULM HTN; PVD; NIDDM; HPI: 2 WEEK HISTORY LEG WEAKNESS; 7/22
FOUND BY HUSBAND ON FLOOR- AWAKE, BUT MENTAL STATUS CHANGES; TO CALVERT
HOSPITAL ER- TO THEIR ICU; HEAD CT- NEG FOR BLEED; VQ SCAN- NEG FOR PE;
ECHO- GLOBAL HYPOKINESIS; EF EST 20%; R/O FOR MI; DIGOXIN TOXIC
WITH HYPERKALEMIA- KAYEXALATE, DEXTROSE, INSULIN; RENAL INSUFFICIENCY- BUN 54,
CR 2.8; INR 7 ( ON COUMADIN AT HOME); 7/23 AT CALVERT- 2 FFP, 2 UNITS PRBC,
 VITAMIN K; REFERRED TO GH.
 ARRIVED IN TRANSFER APPROX. 2130; IN NO MAJOR DISTRESS;
 DOPAMINE TAPER, THEN DC; NS FLUID BOLUS GIVEN WITH
IMPROVEMENT IN BP RANGE; SEE FLOW SHEET SECTION FOR CLINICAL INFORMATION;
A: NO HEMODYNAMIC COMPROMISE SINCE TRANSFER; TOLERATING DOPAMINE DC;
P: TREND BP RANGE; OBSERVE FOR PRECIPITOUS HYPOTENSION.

||||END_OF_RECORD
```

Figure 4.4: Format of an unannotated nursing note from the PhysioNet - Corpus [27].

discarded when calculating performance metrics for this data-set. Similarly, tokens in the PhysioNet corpus corresponding to the HOSPITAL category are mapped to LOCATION tag, unlike other data-sets where the tokens are mapped to HOSPITAL tag. The difference in labels is considered while calculating the performance of every model. The overall performance of the MIST and DE-ID models are evaluated based on the metrics specified in sections 3.1.4 and 3.2.2.

Table 4.4: Count of instances for each PHI category across all three data-sets split into Training and Test subsets.

Data		**AGE**	**DATE**	**DOCTOR**	**HOSPITAL**	**ID**	**LOCATION**	**PATIENT**	**PHONE**
I2b2 2006	Train	12	4976	2621	1686	3358	184	644	167
	Test	4	2103	1117	709	1438	78	281	65
I2b2 2014	Train	0	9574	3248	1570	1022	1575	1518	289
	Test	0	4711	1367	684	435	687	664	126
PhysioNet Corpus	Train	2	367	417	0	0	257	163	37
	Test	1	158	174	0	0	110	66	16

Table 4.4 shows the count of various categories of PHI in each data-set. From the detailed count, the least frequent category is AGE. In the PhysioNet corpus, there is only a single tag LOCATION common for both HOSPITAL and LOCATION and so the count of instances for HOSPITAL is zero. Also, there are no instances of ID present in the PhysioNet corpus. In Table 4.4 most frequently occurring PHI category instances are denoted in bold. In both data-sets from I2b2, the maximum occurring PHI tokens belong to DATE tag. The distribution of the PHI instances is not uniform

across different categories. The results of training the models on these instances and the performances of the models on test data are discussed in Chapter 5.

Chapter 5

Experiments and Results

In this chapter, I present experiments and results of evaluating the models on three different data-sets. The chapter begins with the discussion of what a perfect model would look like and methods for evaluating models. Later I describe and discuss the learning trends of training MIST CRF on every data-set followed by the performance and challenges of the trained models in identifying the true positives (PHI tokens) in test data. Following this, two other variants of CRFs are discussed. A comparison is made among the models for each data-set to identify the best performing model in terms of precision, recall and F1 measure.

5.1 The Approach

Evaluating machine learning model involves assessing the predictive performance of the trained model on an unobserved data-set otherwise known as test data. Cross-validation is an evaluation technique by which a model's predictive performance is validated. This technique helps in evaluating a model's performance on new data, other than the training corpus, in terms of its accuracy in predicting the true labels. In cross-validation, a model is trained on subsets of input data and validated on the complementary subset of the input data. The process repeated for k times with k-1 subsets for training and the left-out subset for validation is known as k-fold cross-validation. I have used 10-fold cross-validation in my work.

It is important to know what contributes to a perfect model and evaluate the models accordingly. In a sequence labeling task like de-identification, it is important to look at the perfection of a model in terms of the model's ability to not mislabel a PHI term as "OTHER" (false negatives). As discussed earlier in chapter 3 regarding the performance metrics of a de-identification system, a model's recall is more important than precision. But it is to be noted that the trade off between recall and precision is specific to the application. Therefore defining a perfect model depends on the model's fit to the application. In other words, to preserve the privacy, the model should have a higher recall, while to preserve the readability and retain the medical information content of the medical records, precision should be high. For my evaluations, a near-perfect model is the one that shows high recall values.

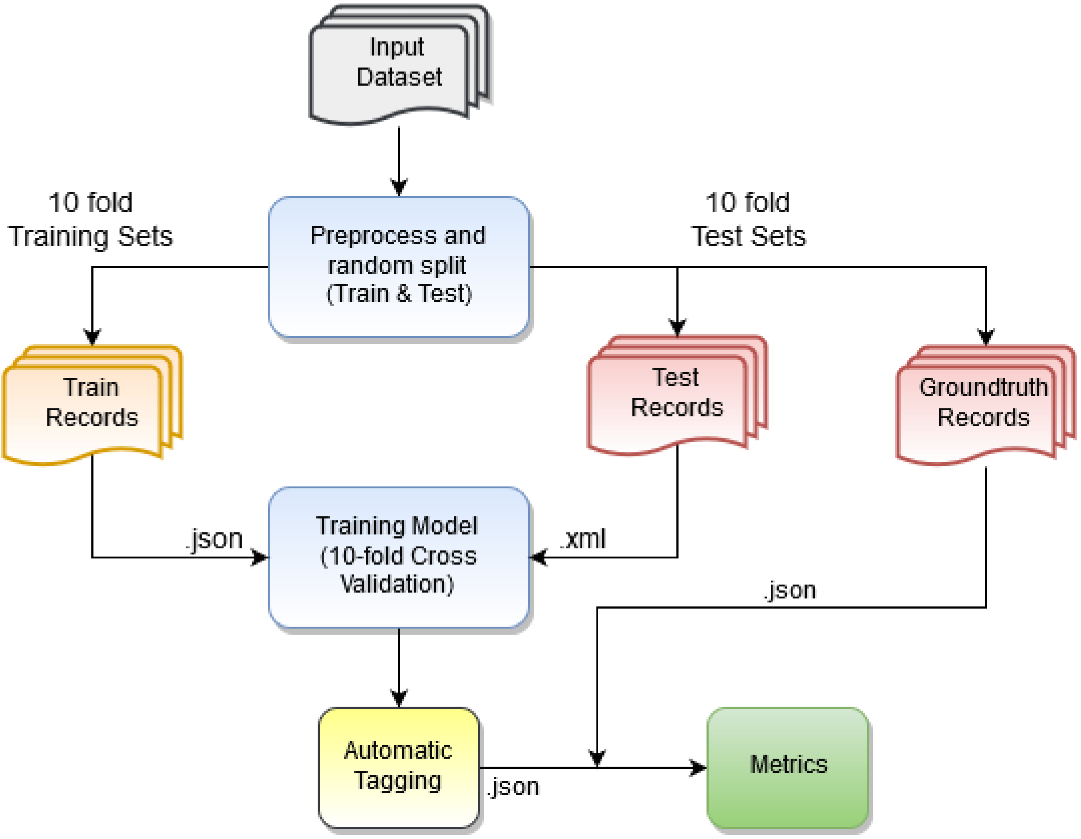


Figure 5.1: Workflow of the evaluation approach. The process begins with random split of preprocessed input records (XML) into 10-folds of training (annotated .json files) and test (unannotated XML files) subsets. Performance is measured for every fold and the average of 10-folds is reported as the final score.

Some papers discuss the ways by which an individual metric can be improved, like in [83], where a bias parameter is introduced to adjust the model weight of the *prior probability* of a token being labeled as "OTHER". This way, the system is biased to improve the recall measure alone. In the experiments here, the model is assessed based on default configuration, so there is no bias to any of the measures in the reported metrics. MIST has a parameter which can help handle the trade off between the precision and recall. But this feature is not included in my experiments. As stated earlier, 10-fold cross-validation was performed to evaluate the models' performance.

For each run, training and test subset are generated by applying a random split on the preprocessed input data. The average values of the model's performance over 10-folds is reported as metrics.

The workflow of my experiment is described in Fig 5.1: the input records are first preprocessed and then divided into two subsets (training and test). The training subset is used for building the MIST models during training. The trained models are made to automatically tag records in the test subset. As a final step, the accuracy of the model's performance is calculated by applying a scoring algorithm on the test documents, annotated by the model, with the ground truth records as the reference. The procedure is repeated for all ten folds and the average of the accuracy is reported in terms of the standard metrics mentioned in section 3.1.4.

5.2 Evaluation Criteria

A general evaluation criterion is followed in calculating the performance metrics of all models. Each model is trained and made to tag records in the test data. The models are evaluated based on their ability to recognize each PHI category in test records. Initially, all test records are preprocessed and made into tokens. The jCarafe engine is responsible for labeling these tokens with suitable PHI categories. MIST determines the number of true positives, false positives, true negatives and false negatives based on a similarity score calculated using the default similarity profile mentioned earlier in *Performance Metrics* section of chapter 3. Measures like a *match*, *missing* and *spurious* are calculated as part of the scoring algorithm. A token is considered to be a *match*, only if the span (start and end position) of the token matches completely in both test and reference documents. For example, a HOSPITAL type PHI token in the sentence, "Dr.Renlan Fyfezeis treated him..", has a span (248-263) in a test record of length 3000, then this token is said to be a *match*, only if there is a corresponding span (248-263) in the reference document with the same label HOSPITAL. For any PHI token, in the reference document, that does not have a corresponding entry in the de-identified test record is known as *missing*. For any PHI token, in the test record, does not have a corresponding entry in the reference record is known as *spurious*. Based on these counts, precision, recall and F1 measures are recorded and used to determine the best performing model for a given data-set.

5.3 MIST CRF

Training

It is necessary to understand the nature of how the models are learned. The models based on CRF are trained using PSA based SGD training method, described in chapter 3, with varying learning rates (α) and a number of iterations. To understand the convergence of the training curves, the average negative conditional log-likelihood (CLL) value at each iteration from ten runs is plotted. Fig 5.2 shows the curves plotted for each model across three data-sets. Each graph consists of three curves, each curve representing an individual model with different learning rate. Each connecting point in the curve is the average negative CLL obtained from 10 runs for every n^{th} iteration.

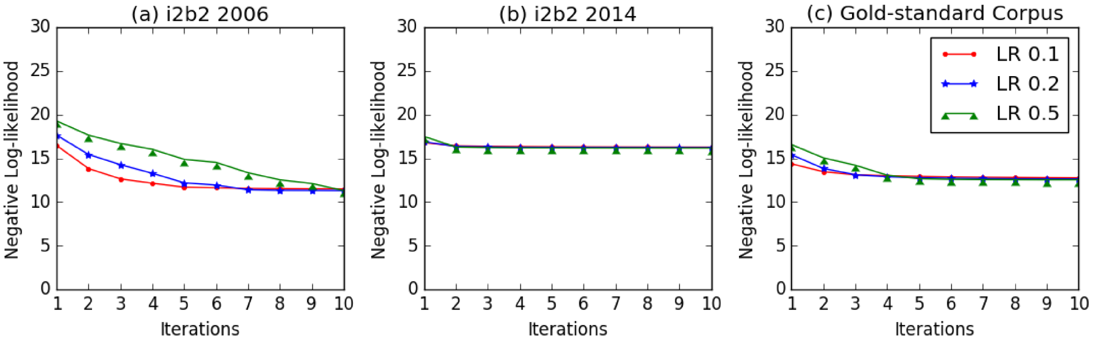


Figure 5.2: Training graphs: (a) I2b2 2006, (b) I2b2 2014, (c) PhysioNet Corpus. Negative CLL value from training the models by PSA based SGD training for different α (learning-rate).

The training curves for I2b2 2006 data-set are plotted in Fig 5.2 (a). All the models are trained on a server having a multi-core processor with 40 cores CPU. The models $\alpha = 0.1$, $\alpha = 0.2$ and $\alpha = 0.5$ are trained on I2b2 2006 training subset (620 records) for a maximum of 20 iterations. A CRF model trained for 20 iterations took an average of 7.5 minutes to complete the training. It is seen that the curves are significantly different from each other for every α. Model with $\alpha = 0.1$ reached the convergence point well before other two models $\alpha = 0.2$ and $\alpha = 0.5$. And the model with $\alpha = 0.5$ does not show any significant convergence but the curve is steadily dropping. During training, the jCarafe engine extracts the features for every run and they are learned by the model. The number of features extracted from the training sub-sets differ based on the length and count of records involved in each sub-set. For example, the number of features extracted from the training subset used in the first run is 287,323 and from second subset it is 292,593 features. The list of feature functions involved in feature extraction is shown in Table 3.1. The trained models are then analyzed based on their tagging capabilities and the performance is discussed under the Performance subsection.

Table 5.1: Metrics of the best performing MIST CRF Model on three data-sets

MIST CRF Model Metrics			
	I2b2 2006	**I2b2 2014**	**PhysioNet Corpus**
	(Discharge Summaries)	(Longitudinal Records)	(Nursing Notes)
Training Docs	620	912	1703
Test Docs	267	392	731
MIST CRF - LR Parameter	0.1	0.2	0.5
Precision	0.9659	0.8711	0.7501
Recall	0.9544	0.8036	0.4287
F1 Measure	0.9598	0.835	0.5897

Table 5.2: Average tag-level performance of best performing MIST CRF model on three data-sets.

	I2b2 2006			**I2b2 2014**			**PhysioNet Corpus**		
	(Discharge Summaries)			**(Longitudinal Records)**			**(Nursing Notes)**		
PHI Type	**Precision**	**Recall**	**F1 Measure**	**Precision**	**Recall**	**F1 Measure**	**Precision**	**Recall**	**F1 Measure**
AGE	1	0.1226	0.1817	*NA*	*NA*	*NA*	0.75	0.8125	0.5833
DATE	0.9646	0.9807	0.9727	0.8827	0.8148	0.8437	0.7334	0.5169	0.6056
DOCTOR	0.9583	0.9488	0.9536	0.8605	0.8398	0.8501	0.8112	0.4712	0.5953
HOSPITAL	0.9630	0.9312	0.9469	0.869	0.8066	0.8365	*NA*	*NA*	*NA*
ID	0.9781	0.9826	0.9804	0.8366	0.8194	0.8277	*NA*	*NA*	*NA*
LOCATION	0.4175	0.2917	0.4175	0.8879	0.7693	0.8236	0.7469	0.4306	0.5453
PATIENT	0.9851	0.9446	0.9643	0.8254	0.7448	0.7826	0.8594	0.1817	0.2962
PHONE	0.7918	0.7061	0.7918	0.9	0.708	0.7897	0.6333	0.0763	0.1176
ALL	0.9659	0.9544	0.9598	0.8711	0.8036	0.8350	0.7695	0.4287	0.5502

In Fig 5.2 (b) the models are trained on the I2b2 2014 training subset (790 records) for a maximum of 30 iterations. It took an average of 24 minutes to train a model on I2b2 2014 training subset. Three models ($\alpha = 0.1$, 0.2 and 0.5) showed a similarity in converging to the optimal values of negative CLL. The number of features extracted from the first training subset is 466655 and from the second subset is 484452. In the Fig 5.2 (c) the curves represent models trained using PhysioNet corpus. The models are trained on 1,703 training records for a maximum of 30 iterations. The average training time taken by the model is 4 minutes. The curves show similar values beyond the convergence but are different in their initial values. The average number of features extracted from each training sub-set is 265,400.

In all three data-sets, models with $\alpha = 0.1$ were consistent with their training curves. The behavior of models with $\alpha = 0.5$ showed difference in each data-set. For I2b2 2006 and PhysioNet corpus, the models reached a similar value at the end of all iterations. All trained models have been evaluated by making them tag the records in the test subset and corresponding performances measures (precision, recall, and F1 measure) have been calculated and analyzed.

Performance

Table 5.3 shows measures of three models trained with learning rates of 0.1, 0.2 and 0.5 respectively, for all three data-sets. Best performing model is chosen by considering a balanced precision and recall measures. Model $\alpha = 0.1$ achieved high precision, recall and F1 measure for I2b2 2006 data-set. In longitudinal records (I2b2 2014), model with $\alpha = 0.2$ outperformed the other two models $\alpha = 0.1$ and 0.5. Model $\alpha = 0.2$ scored maximum precision values for DATE, DOCTOR, HOSPITAL, LOCATION and PATIENT categories when compared to other two models $\alpha = 0.1$ and 0.5. Overall model $\alpha = 0.2$ recorded highest recall values for all PHI categories.

Performance of MIST CRF on the PhysioNet corpus recorded comparatively lower F1 score than other data-sets. The reason for this is that the number of PHI categories are fewer compared to other two data-sets (discharge summaries and longitudinal records). Model ($\alpha = 0.5$) scored a maximum precision of 0.769 and model ($\alpha = 0.2$) achieved maximum recall of 0.485 and F1 measure of 0.589. This data-set does not have HOSPITAL nor ID categories. The score for the tag LOCATION is the score for identifying locations and hospitals in a given text. Metrics of best performing MIST CRF model is shown in Table 5.1 which also has information about the number of training and test records used for evaluating the performance of the model on each type of medical record.

Table 5.3: Average tag-level performance of MIST CRF models for different learning rate α on all three data-sets. Models with best scores for each PHI tag is shown in bold.

		I2b2 2006			**I2b2 2014**			**PhysioNet Corpus**		
Metric	**PHI Type**	α**=0.1**	α **= 0.2**	α **= 0.5**	α **= 0.1**	α **= 0.2**	α **= 0.5**	α **= 0.1**	α **= 0.2**	α **= 0.5**
Precision	AGE	1	1	1	NA	NA	NA	0.75	0.75	0.75
	DATE	0.9646	**0.9654**	0.943	0.8674	**0.8827**	0.8774	0.7047	0.72	**0.7334**
	DOCTOR	0.9583	**0.9629**	0.9327	0.8487	**0.8605**	0.8188	0.8011	**0.8216**	0.8112
	HOSPITAL	**0.963**	0.9629	0.9483	0.8445	**0.869**	0.8543	NA	NA	NA
	ID	**0.9781**	0.9721	0.9221	0.7978	0.8366	**0.8505**	NA	NA	NA
	LOCATION	0.4175	**0.6989**	0.4533	0.8819	**0.8879**	0.8734	0.7322	**0.751**	0.7469
	PATIENT	**0.9851**	0.9834	0.9476	0.8172	**0.8254**	0.7694	**0.8716**	0.86	0.8594
	PHONE	0.7918	0.918	**0.9536**	**0.934**	0.9	0.8745	**0.6617**	0.52	0.6333
	ALL	**0.9659**	0.9656	0.9348	0.8559	**0.8711**	0.8547	0.7501	0.7675	**0.7695**
Recall	AGE	0.1226	**0.2276**	0.1202	NA	NA	NA	0.625	0.625	**0.8125**
	DATE	0.9807	**0.9817**	0.9747	0.808	**0.8148**	0.8064	**0.6081**	0.5774	0.5169
	DOCTOR	0.9488	**0.9525**	0.9038	0.8135	**0.8398**	0.792	**0.533**	0.5192	0.4712
	HOSPITAL	0.9312	**0.9323**	0.9088	0.7804	**0.8066**	0.7627	NA	NA	NA
	ID	**0.9826**	0.9758	0.9274	0.7963	**0.8194**	0.7912	NA	NA	NA
	LOCATION	**0.2917**	0.254	0.124	0.7171	**0.7693**	0.7189	**0.444**	0.4397	0.4306
	PATIENT	0.9446	**0.9494**	0.849	0.7116	**0.7448**	0.6717	**0.2278**	0.2064	0.1817
	PHONE	**0.7061**	0.6407	0.5334	0.6077	**0.708**	0.689	0.073	0.0525	**0.0763**
	ALL	**0.9544**	0.9528	0.9181	0.7826	**0.8036**	0.7747	**0.4858**	0.4678	0.4287
F1 Measure	AGE	0.1817	**0.3106**	0.1844	NA	NA	NA	0.375	0.375	**0.583375**
	DATE	0.9727	**0.9737**	0.9585	0.8326	**0.8437**	0.8364	**0.6522**	0.6404	0.6056
	DOCTOR	0.9536	**0.9578**	0.918	0.8306	**0.8501**	0.805	**0.6399**	0.6358	0.5953
	HOSPITAL	0.9469	**0.9474**	0.9283	0.8112	**0.8365**	0.8061	NA	NA	NA
	ID	**0.9804**	0.9737	0.9247	0.797	**0.8277**	0.8198	NA	NA	NA
	LOCATION	**0.4175**	0.3626	0.1932	0.7906	**0.8236**	0.7884	0.552	**0.5539**	0.5453
	PATIENT	0.9643	**0.9657**	0.8955	0.7606	**0.7826**	0.7169	**0.3597**	0.33	0.2962
	PHONE	**0.7918**	0.75	0.6813	0.7339	**0.7897**	0.7693	0.116	0.0882	**0.1176**
	ALL	**0.9598**	0.9592	0.9262	0.8168	**0.835**	0.8118	**0.5897**	0.5811	0.5502

The tag-level metrics for these models is shown in Table 5.2. It can be seen that the DATE category shows a high recall score in all three data-sets. From Table 4.4, the number of training instances for DATE category is higher than other categories and so the model is well learned for this category. Name-related PHI categories like DOCTOR and PATIENT show better and consistent results across all data-sets when compared to other PHI categories. MIST CRF performs poorly in identifying LOCATION in discharge summaries (0.4175) compared to that of longitudinal records (0.889) and nursing notes (0.7469). The model achieved poor performance in identifying PHONE category tokens in nursing notes. From the values in the clinical notes, it is observed that the PHONE values took different formats (xxx-xxx-xxxx, xxx xxx xxxx, xxx xxxx, xxx/xxx/xxxx, xxx xxx-xxxx) but did not have sufficient training instances for each format to make the model to learn. From Table 4.4 the model had only 36 instances of PHONE category tokens for training. The model fairly performed well on ID category with a precision of 0.9781 in discharge summaries and 0.8366 in longitudinal records. Overall, MIST CRF showed a good performance on DATE, DOCTOR, PATIENT, HOSPITAL and ID categories in all types of medical records. The performance on AGE, PHONE and LOCATION seemed to be specific to the type of input records and the number of training instances involved.

5.4 MIST semi-CRF

Training

In this variant, models are trained as semi-CRF using the semi-CRF training mode. As specified in Table 3.2, the models extract special features by considering the segments in the text. For example, a *phraseFn* extracts a current segment (n-gram) as a single feature but this type of segment features are not supported by MIST CRF. The models are trained for different *alpha* values (0.1, 0.2 and 0.5) like MIST CRFs. The learning results are surprising for semi-CRF models as the values for each learning rate are statistically insignificant from each other. The graphs in Fig 5.3 show a single curve that represents $alpha = 0.1$ and the same learning rate is used for all data-sets. The curves show a similarity in the way they converge for all three data-sets. The number of iterations, required to reach an optimal CLL that yields good performance, is higher for the semi-CRF models. To achieve the optimal level of performance, the model required 60 iterations on I2b2 2006 data-set and took an average of 17 minutes for training. For I2b2 2014 it took an average of 60 minutes and for the PhysioNet corpus, it took 9 minutes. For later data-sets, the model required 70 iterations to produce good results.

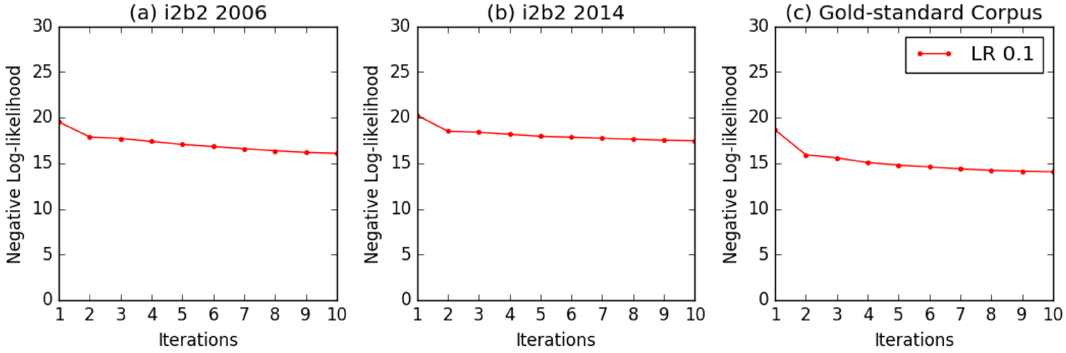


Figure 5.3: Semi-CRF training curves for three data-sets: (a) I2b2 2006, (b) I2b2 2014, and (c) PhysioNet Corpus.

Table 5.4: Metrics of the best performing MIST semi-CRF model on three data-sets

MIST semi-CRF Model Metrics			
	I2b2 2006	**I2b2 2014**	**PhysioNet Corpus**
	(Discharge Summaries)	(Longitudinal Records)	(Nursing Notes)
Training Docs	620	912	1703
Test Docs	267	392	731
Learning Rate (α)	0.1	0.1	0.1
Precision	0.765	0.741	0.648
Recall	0.830	0.696	0.410
F1 Measure	0.796	0.717	0.502

Performance

As discussed earlier, experiments showed that the models generated statistically indistinguishable results for $alpha = 0.1$, 0.2 and 0.5. Thus, the metrics for one model with $alpha = 0.1$ is shown in Table 5.4. The table shows the count of records for training, testing and the metrics for each data-set. Semi-CRF performed well on discharge summaries scoring F1 measure of 0.796 compared to longitudinal records (0.717) and PhysioNet corpus (0.502). As recorded in Table 5.5, semi-CRF models showed higher recall for DATE, DOCTOR, HOSPITAL and ID categories in discharge summaries and longitudinal medical records but performed poorly on PhysioNet corpus. In discharge summaries, the model showed a poor recall for LOCATION and AGE tags. With the PhysioNet corpus, the model achieved poor recall values for most categories.

Table 5.5: Average tag level metrics of MIST semi-CRF models on all three data-sets

MIST semi-CRF Tag Level Metrics										
Metrics	**Data-set**	**AGE**	**DATE**	**DOCTOR**	**HOSPITAL**	**ID**	**LOCATION**	**PATIENT**	**PHONE**	**ALL**
Precision	I2b2 2006	1.000	0.819	0.617	0.789	0.883	0.195	0.474	0.893	0.765
	I2b2 2014	0	0.803	0.676	0.728	0.772	0.694	0.535	0.791	0.741
	PhysioNet Corpus	0.667	0.469	0.758	NA	NA	0.773	0.818	0.333	0.648
Recall	I2b2 2006	0.014	0.901	0.731	0.826	0.925	0.115	0.492	0.538	0.830
	I2b2 2014	0	0.729	0.685	0.736	0.760	0.628	0.539	0.719	0.696
	PhysioNet Corpus	0.667	0.377	0.497	NA	NA	0.422	0.262	0.167	0.410
F1 Measure	I2b2 2006	0.025	0.858	0.669	0.789	0.903	0.130	0.483	0.658	0.796
	I2b2 2014	0	0.761	0.681	0.732	0.765	0.659	0.536	0.752	0.717
	PhysioNet Corpus	0.333	0.418	0.600	NA	NA	0.545	0.396	0.216	0.502

5.5 MIST neural-CRF

Training

Training neural-CRF model involves reaching the optimized values for the model parameters: the learning rate (α), the hidden learning rate and the number of neural/hidden gates that should be introduced at each position in a input sequence. In this training, the features that are allowed to pass through the neural gates are extracted from the lexicon function. If the word "FIH" in the sentence "admitted in FIH.." appears in the lexicon category "hospitals", then the category name "hospitals" is included as a neural feature to the model. This helps the model learn the non-linear properties of the input. In case of neural models also, the training results seemed to be statistically insignificant for different hidden learning rates (0.1, 0.2 and 0.5) on all three data-sets. The curves in Fig 5.4 are representations of model with learning rate $\alpha = 0.1$, hidden learning rate $= 0.1$ and number of neural gates $= 75$. Same parameter values are used for all three data-sets but the number of iterations was configured based on the data-set.

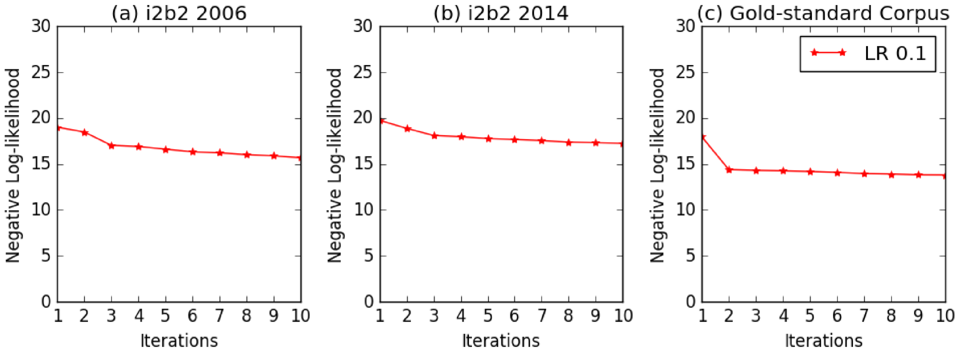


Figure 5.4: Neural-CRF training curves for three datasets: (a) I2b2 2006, (b) I2b2 2014, and (c) PhysioNet corpus. Model parameter values: $\alpha = 0.1$, hidden learning rate = 0.1, number of neural gates = 75.

With a few preliminary experiments, it was found that to increase the performance, the model has to be trained for a certain number of iterations. Observations show that the model required a minimum of 50 iterations of training to reach a good performance. For evaluation purposes, the value of the iterations and the number of hidden gates are made consistent for all three data-sets. Time taken to train the neural models with the above parameter values on I2b2 2006 data-set is 14 minutes, on I2b2 2014 data-set it is 16.5 minutes and on PhysioNet corpus it is 7.5 minutes. In Fig 5.4 (b) the models retained high values for negative CLL at the end of 10th iteration compared to values in other data-sets.

Performance

Table 5.6 shows the overall metrics of neural-CRF model on the three data-sets. From the overall performances, it can be seen that the neural-CRF model showed outstanding results on discharge summaries among the three data-sets. The model recorded the highest precision (0.959) and recall (0.956) on discharge summaries. There is a perfect balance in the precision and recall scores of discharge summaries while in case of longitudinal medical records, the model showed a higher precision (0.747) but low recall (0.632). For the PhysioNet corpus, the model achieved nearly balanced scores for precision (0.688) and recall (0.617).

The tag-level performance is shown in Table 5.7. The model achieved high precision scores of (0.955, 0.776 and 0.755) for DOCTOR and (0.969, 0.645 and 0.731) for

PATIENT categories on all three data-sets. These two PHI tags are considered main categories that a model needs to identify and neural-CRF performed well in identifying these tokens in all three data-sets. The neural-CRFs scored well on pattern-based tokens like DATE and PHONE for all three data-sets. It can be seen that the metrics of the AGE category for longitudinal records and nursing notes are marked 0 as the model did not encounter enough training exemplars thus failing to identify the tags in test records. As shown in Table 4.4, the PhysioNet corpus had no separate tags for HOSPITAL and ID fields in the training records and so the tags are not included in the metrics.

Table 5.6: Metrics of the best performing MIST neural-CRF model on three data-sets.

MIST neural-CRF Model Metrics			
	I2b2 2006	**I2b2 2014**	**PhysioNet Corpus**
	(Discharge Summaries)	(Longitudinal Records)	(Nursing Notes)
Training Docs	620	912	1703
Test Docs	267	392	731
Learning Rate (α)	0.1	0.1	0.1
Hidden Learning Rate	0.1	0.1	0.1
Number of Gates	75	75	75
Precision	0.959	0.747	0.688
Recall	0.956	0.632	0.56
F1 Measure	0.957	0.685	0.617

Table 5.7: Average tag level metrics of MIST neural-CRF models for the optimal parameter values ($\alpha = 0.1$, hidden learning rate - 0.1, number of neural gates - 75) on all three data-sets.

MIST neural - CRF Tag Level Metrics										
Metrics	**Data-set**	**AGE**	**DATE**	**DOCTOR**	**HOSPITAL**	**ID**	**LOCATION**	**PATIENT**	**PHONE**	**ALL**
Precision	I2b2 2006	1	0.968	0.955	0.94	0.983	0.6	0.969	0.845	0.959
	I2b2 2014	0	0.831	0.776	0.68	0.498	0.502	0.645	0.833	0.747
	PhysioNet Corpus	0	0.674	0.755	NA	NA	0.6	0.731	0.5	0.688
Recall	I2b2 2006	0	0.982	0.942	0.918	0.989	0.64	0.899	0.778	0.956
	I2b2 2014	0	0.65	0.747	0.63	0.597	0.362	0.618	0.485	0.632
	PhysioNet Corpus	0	0.722	0.621	NA	NA	0.395	0.279	0.571	0.56
F1 Measure	I2b2 2006	0	0.975	0.948	0.929	0.986	0.62	0.932	0.81	0.957
	I2b2 2014	0	0.729	0.761	0.654	0.543	0.42	0.631	0.613	0.685
	PhysioNet Corpus	0	0.697	0.682	NA	NA	0.476	0.404	0.533	0.617

5.6 PhysioNet DE-ID

Performance

Performance of a rule-based system like PhysioNet DE-ID is based on dictionaries, patterns and look-up tables used by the system. The system achieved a precision of 96.7 and recall of 74.8 on the PhysioNet corpus. DE-ID identified a total of 1,703 true positives out of 1,768 total PHI instances. For I2b2 2006 and I2b2 2014, the input files were converted to fit the model's format requirements. The performance of the model as calculated by the tool reported a low precision of 0.092 for I2b2 2014 data-set. Upon analysis, the spanning of the tokens in the resulting de-identified file did not fit the spanning of the actual ground-truth files of I2b2 2006 and I2b2 2014. As such, I used the "Output" mode to run the model for discharge summaries and longitudinal medical records. Running in this mode does not produce any performance statistics, but only generates the de-identified file as output, which was then scored manually. The model showed better results for DATE category and identified holidays (e.g., "thanksgiving") which was a miss in the other two variants.

DE-ID showed surprising results for the AGE tag compared to that of the statistical models which required training instances for learning. But DE-ID performed by pattern matching technique like in the case where the pattern is an integer value of age followed by suffix years (e.g., "John is 90y and he is from Rome."). As DE-ID is a rule-based system, the performance of this system is restricted to the scope of the lists, patterns, lexical tokens and dictionaries that are used for the de-identification task.

5.7 Discussion

In a task like medical record anonymization, it is difficult to achieve a near-perfect model. From the individual results discussed in previous sections, in most cases, the models achieved high precision when ran on a default configuration without any bias parameters for recall. From the combined results of each variant of MIST on three different data-sets, the 95% confidence intervals are plotted for precision and recall. They are shown in Fig 5.5 and Fig 5.6 respectively. CRF models from Fig 5.5 (a) and Fig 5.6 (a) have very smaller intervals making the average values of these models certain. Similarly, neural-CRF model in Fig 5.5 (b) and Fig 5.6 (b) showed certainty in the performance of the models. The neural models are statistically significant from other two variants (CRF and semi-CRF), with P < 0.05, for discharge summaries and the PhysioNet corpus. From these observations, we see that neural-CRF produced near-perfect results by scoring a high recall of 0.956 for discharge summaries and 0.56 for nursing notes. For longitudinal medical records, CRF scored higher recall of 0.875 compared to semi-CRF 0.771 and neural-CRF 0.627.

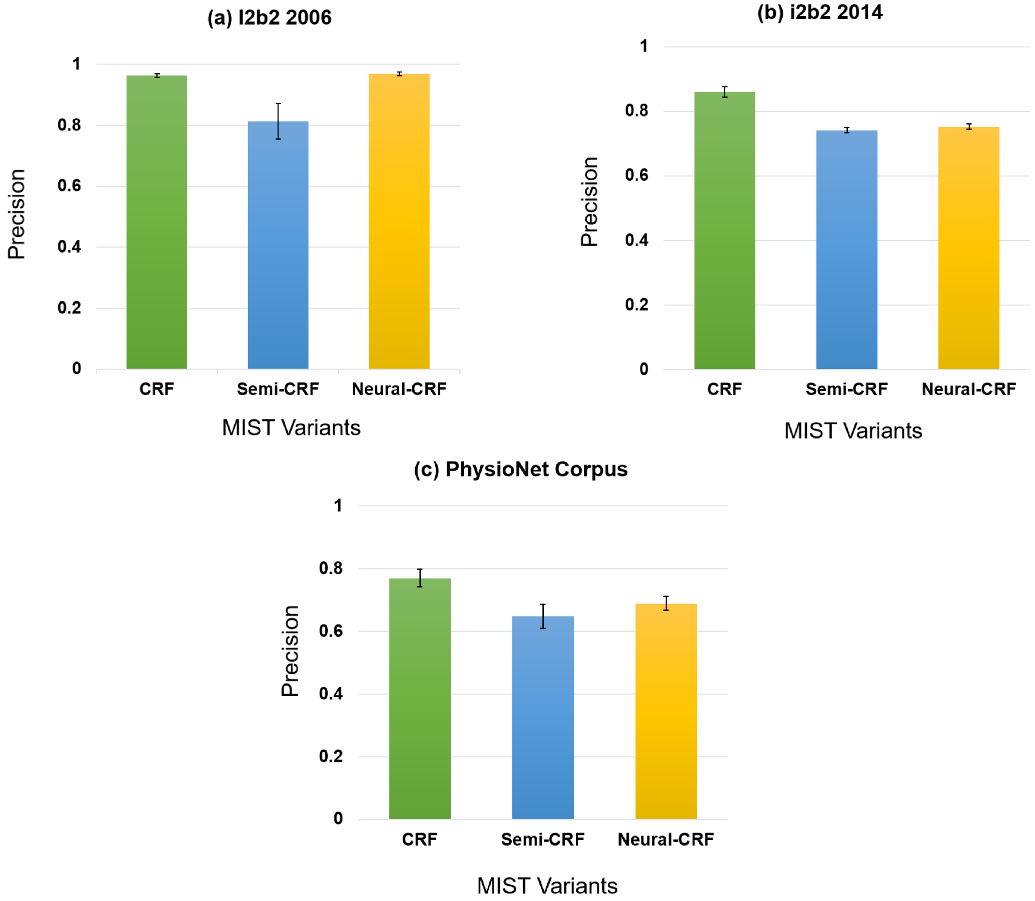


Figure 5.5: 95% confidence intervals for precision on three data-sets: (a) I2b2 2006, (b) I2b2 2014, (c) PhysioNet corpus.

Considering a model's ability to identify the true positives, neural-CRFs scored higher precision of 0.969 (average of 5,688 true positives out of 5795) for discharge summaries. The CRF models scored a maximum number of true positives (indicates high precision) for longitudinal medical records with an average of 6,489 matches out of 8,674 PHI instances and for nursing notes: an average of 223 matches out of 525 total PHI instances. Neural-CRF showed consistent results by achieving a balanced precision (0.959) and recall (0.956) for discharge summaries, making it the best model with an F1 score of (0.957). While it is straightforward that the neural-CRF outperformed other two models for a high recall on the PhysioNet corpus, it scored low precision (0.688) compared to MIST CRF (0.769). For longitudinal medical records, high recall (0.771) and precision (0.86) were by MIST CRF. At a tag level, neural-CRF models showed surprising results which even the core CRF could not produce. Example, for the LOCATION tag, neural-CRF achieved the highest precision of 0.6, which is twice as high as the value achieved by the other two models (0.367 and 0.376). Also neural-CRF model outperformed other two variants in identifying PHONE tags with a precision of 0.845 while it is only 0.785 for MIST CRF and 0.746 for semi-CRF.

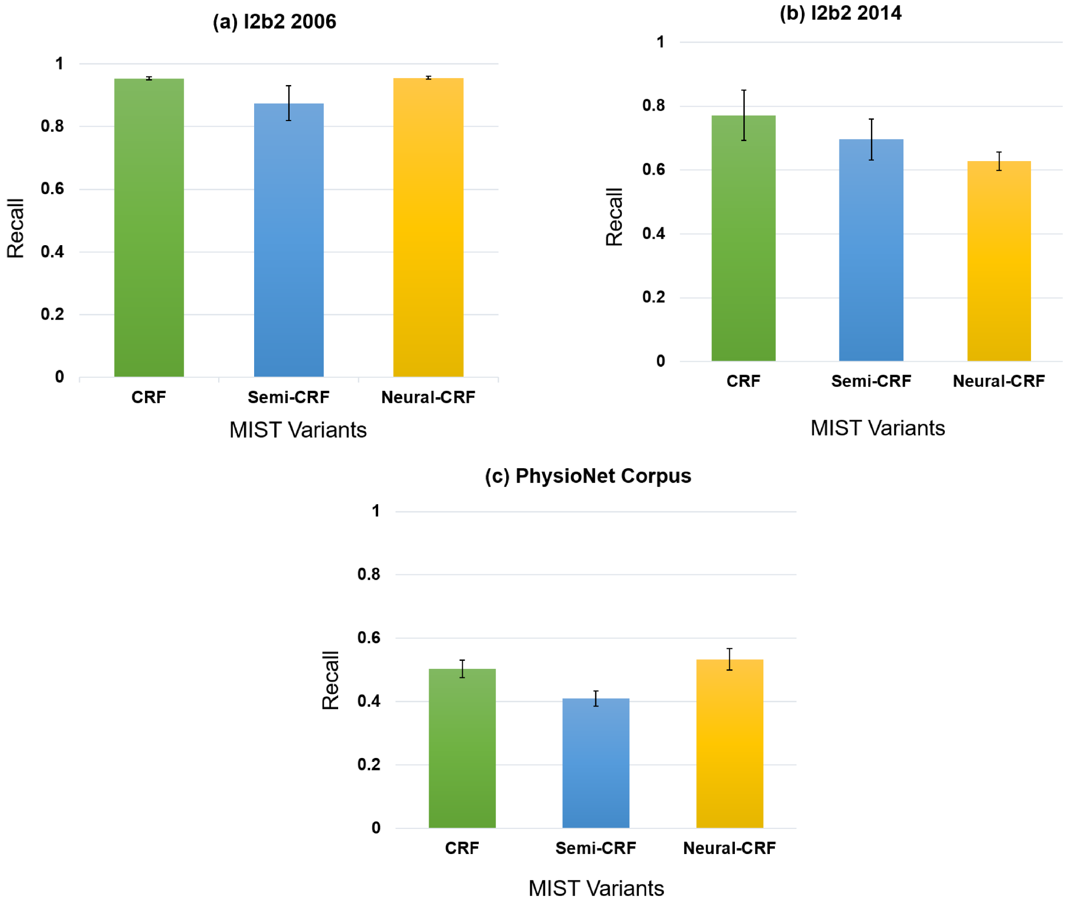


Figure 5.6: 95% confidence intervals for recall on three data-sets: (a) I2b2 2006, (b) I2b2 2014, (c) PhysioNet corpus. Models are statistically significant ($P < 0.05$) from each other.

Error Analysis

It is important that the models are also assessed based on the errors encountered for each type of medical record. Confusion matrices for each model are discussed here along with examples for discharge summaries and PhysioNet corpus. As the list of tags available in input data is similar in case of I2b2 2006 and I2b2 2014 data-sets, one of the data-set is ruled out for confusion matrix discussion. The confusion matrix is designed to list the number of true labels (bold values), *missing* terms, *spurious* terms and type errors for every tag. Table 5.8 refers to the confusion matrix for the three variants of MIST on I2b2 2006 data-set.

Missing tags are harmful as they denote exposure of a portion of PHI tags in the de-identified records. Considering the importance of tags like PATIENT, DOCTOR, ID, and PHONE, the rest of this section discusses the error analysis of these tags. From the de-identified samples, it was observed that all three models failed to identify a set of patient first names (e.g. “Walking” and “Straight”). Upon further analysis, neural-CRFs seemed to miss only the first names that are mostly ambiguous to generic tokens (e.g. “Walking”) compared to semi-CRFs and traditional CRFs which missed some unambiguous full names (e.g. “Rora Dose” and “Daie Elms”). Also, neural-CRF produced only 6 type errors (incorrect labeling) and all six occurrences PATIENT names got incorrectly labeled as DOCTOR. Traditional CRF and semi-CRF produced 17 and 10 type errors (incorrect labels include DOCTOR, LOCATION, and HOSPITAL).

In the case of DOCTOR category, semi-CRFs showed better results in the count of true labels (2,551) compared to neural CRF (2,485) and traditional CRF (2,237). As seen for PATIENT, neural-CRFs produced less number of type errors when compared to the other two models. Traditional CRFs recorded frequent type errors with label PATIENT in cases where the names of DOCTOR are long and case-sensitive (e.g. “FIGEBREUNQINKITH , GI” and “GLOMNEBRED , JAHOLL”). Semi-CRF model showed a large *spurious* count of 243 tokens for DOCTOR tag which is as low as 20 by neural CRF and 13 by traditional CRF. Categories like ID and PHONE showed a less count of *spurious* tags by neural and semi-CRF, which in most cases are 2 digit laboratory values with special symbols (e.g. “-20”, “0-24”, “134/12”). But traditional CRF encountered 0 spurious tags for DOCTOR category. The numerical category AGE did not have many training instances for I2b2 2006 data-set and results reflect the same. Addition of feature functions that extract properties related to AGE tag is recommended for all models.

Table 5.8: Confusion matrix for all variants of CRF on I2b2 2006 Data-set. The measures denote count of PHI tokens.

(a) CRF - I2b2 2006

Labels	**AGE**	**DATE**	**DOCTOR**	**HOSPITAL**	**ID**	**LOCATION**	**PATIENT**	**PHONE**	*MISSING*	**TOTAL**
AGE	**1**	0	0	0	0	0	0	0	*3*	4
DATE	0	**2360**	2	2	30	1	0	1	*23*	2419
DOCTOR	0	3	**2237**	11	3	7	41	0	*30*	2332
HOSPITAL	0	5	5	**1377**	2	10	1	0	*28*	1428
ID	0	3	1	1	**1507**	5	0	25	*12*	1554
LOCATION	0	1	1	3	0	**31**	0	0	*44*	80
PATIENT	0	0	14	2	0	1	**528**	0	*13*	558
PHONE	0	0	0	0	0	0	0	**40**	*5*	45
SPURIOUS	0	*54*	*13*	*5*	*16*	*2*	*0*	*0*		90
TOTAL	1	4783	2273	1401	1558	57	1140	66	158	8510

(b) semiCRF - I2b2 2006

Labels	**AGE**	**DATE**	**DOCTOR**	**HOSPITAL**	**ID**	**LOCATION**	**PATIENT**	**PHONE**	*MISSING*	**TOTAL**
AGE	**1**	0	0	0	0	0	0	0	*4*	5
DATE	0	**2498**	0	1	11	1	0	1	*44*	2556
DOCTOR	0	0	**2551**	14	1	21	4	0	*47*	2738
HOSPITAL	0	0	1	**1454**	1	5	0	0	*29*	1490
ID	0	0	0	1	**1507**	4	0	11	*20*	1544
LOCATION	0	0	0	1	0	**80**	0	0	*42*	123
PATIENT	0	0	10	0	0	0	**676**	0	*39*	725
PHONE	0	0	0	0	0	0	0	**56**	*3*	59
SPURIOUS	0	*260*	*243*	*88*	*66*	*10*	*51*	*1*		719
TOTAL	1	2758	2805	1559	1586	121	731	69	228	9959

(c) neural CRF - I2b2 2006

Labels	**AGE**	**DATE**	**DOCTOR**	**HOSPITAL**	**ID**	**LOCATION**	**PATIENT**	**PHONE**	*MISSING*	**TOTAL**
AGE	0	0	0	0	0	0	0	0	*7*	7
DATE	0	2696	0	0	0	0	0	0	*22*	2718
DOCTOR	0	0	2485	5	0	1	32	0	*38*	2561
HOSPITAL	0	0	4	1575	1	5	0	8	*25*	1618
ID	0	0	0	4	1514	1	0	4	*7*	1530
LOCATION	0	9	11	6	0	156	2	0	*23*	207
PATIENT	0	0	6	0	0	0	590	0	*17*	613
PHONE	0	0	0	0	0	0	0	74	*4*	78
SPURIOUS	0	*27*	*20*	*12*	*10*	*18*	*5*	*3*		95
TOTAL	0	2732	2526	1602	1525	181	629	89	143	9427

Table 5.9 shows the confusion matrix for the PhysioNet corpus (nursing notes). In this data-set, the concentration of PHI terms is much lower than the other two data-sets. MIST CRFs showed poor recall for the PATIENT tag in PhysioNet corpus. In most cases, the model wrongly labeled PATIENT names and missed most occurrences

of PATIENT names. On exploring the de-identified samples, most of the names that the model failed to identify seemed to be frequently occurring names (e.g. "John", "Emily", "Charlie"). In some rare cases, the model could not identify capitalized forms of names that showed up as a *match* for lower case names (e.g. "Veronica" and "VERONICA").

But the model showed a single occurrence of type error, where the sentence "<PATIENT> John <PATIENT> states, his sister is living in <LOCATION> Rome <LOCATION>" got de-identified as "<DOCTOR> John <DOCTOR> states his sister is living in <LOCATION> Rome <LOCATION>". Semi-CRFs and MIST CRFs produced similar errors for PATIENT and DOCTOR tags. Semi-CRFs and MIST CRFs models suffer overmarking errors where the reference content had characters in excess to the documents de-identified. For example, DOCTOR names in reference had suffix "Dr." to most DOCTOR names which the model failed to tag in the test documents. This type of error is more prominent with DOCTOR, LOCATION and DATE tags.

Robustness of De-identification Models

One of the important aspects of a de-identification system is its impact on clinical information. The impact of a system is assessed manually as in [9], where the impact of a system is measured using the *interpretability score* calculated for a subset of input records. The score determined the percentage of medical information retained in the de-identified samples. Another approach [23] discusses the impact of the system by applying information extraction techniques on raw medical notes and compare the results with those obtained from the de-identified notes. In my work, I followed a manual approach in which, the de-identified clinical notes resulted from a single run were assessed for medical terms that were mislabeled as PHI token. Considering the research motive of dealing with unstructured text, importance is given to non-numerical medical terms which may be disease names, symptoms of a disease, parts of human body and drug names. Non-numerical values like laboratory readings were mostly confused with DATE and PHONE categories. They are already discussed as part of the *missing* and *spurious* tokens.

Table 5.9: Confusion matrix for all variants of MIST on PhysioNet Corpus. All measures are count of tokens.

(a) CRF - PhysioNet Corpus

Labels	AGE	DATE	DOCTOR	LOCATION	PATIENT	PHONE	*MISSING*	Total
AGE	1	0	0	0	0	0	*1*	2
DATE	0	9	0	0	0	0	*50*	82
DOCTOR	0	0	91	0	1	0	*70*	172
LOCATION	0	0	0	59	0	0	*51*	110
PATIENT	0	0	1	0	12	0	*50*	63
PHONE	0	0	0	0	0	1	*13*	14
SPURIOUS	*0*	*29*	*13*	*8*	*0*	*1*		51
TOTAL	1	38	105	67	13	2	235	494

(b) semi-CRF - PhysioNet Corpus

Labels	AGE	DATE	DOCTOR	LOCATION	PATIENT	PHONE	*MISSING*	TOTAL
AGE	1	0	0	0	0	0	1	2
DATE	0	108	0	0	0	0	57	165
DOCTOR	0	0	94	0	5	0	76	175
LOCATION	0	0	0	60	0	0	58	108
PATIENT	0	0	1	0	17	0	42	60
PHONE	0	0	0	0	0	2	6	8
SPURIOUS	0	27	13	6	2	1		49
TOTAL	1	135	108	66	24	3	240	567

(c) neural-CRF - PhysioNet Corpus

Labels	AGE	DATE	DOCTOR	LOCATION	PATIENT	PHONE	*MISSING*	TOTAL
AGE	0	0	0	0	0	0	0	0
DATE	0	107	0	0	0	2	36	145
DOCTOR	0	0	66	0	0	0	50	106
LOCATION	0	0	2	52	1	0	54	109
PATIENT	0	0	11	5	16	0	46	78
PHONE	0	0	0	0	0	3	1	4
SPURIOUS	0	47	21	17	1	3		89
TOTAL	0	154	100	74	18	8	187	531

MIST identified a list of *spurious* tags, which are tokens with PHI labels in the de-identified notes but these tokens do not bear any PHI labels in the actual reference documents. Upon analyzing the de-identified samples resulted from the first run using MIST CRF, 15 occurrences out of a total 118 *spurious* occurrences (12.7%)

were non-numerical medical terms. Some of the short terms, as reported in the de-identified notes, are “Scattered Kaposi” (disease name), “RIGHT ADNEXAL MASS” (symptoms of a disease) and “Sloan-caer/viewbonse” (symptoms). Out of the total 15 medical related terms, 8 occurrences had full segments that described symptoms and names of diseases (e.g., “HE HAS DYSPNEA ON EXERTION , HOWEVER , DENIES ORTHOPNEA”, “WHEN HE BECAME FEBRILE”). From the de-identified notes generated by semi-CRF, 47 out of 546 *spurious* occurrences (8.6%) were non-numerical terms related to medical content. Unlike the traditional CRF, semi-CRF models were more inclined to short words which were either names of disease (“Kaposi”, “Jaundice”), parts of human body (“Lung”, “Hand”), common health terms (e.g., “Infectious”, “Trauma”, “Medicine”, “Anticoagulation”, “Fluids”). And finally with the results produced by neural-CRF, 51 occurrences out of 176 (28.9%) were non-numerical medical content-related terms. There were total 7 instances where the model mislabeled drug names (e.g., “Nasalide 600 Meigs q 12”, “Tamoxifen”) as PHI categories. Overall all three models showed an impact on loss of clinical information. But comparatively, the semi CRFs can be considered less impactful as they often deal with short and common medical terms, unlike the other two models.

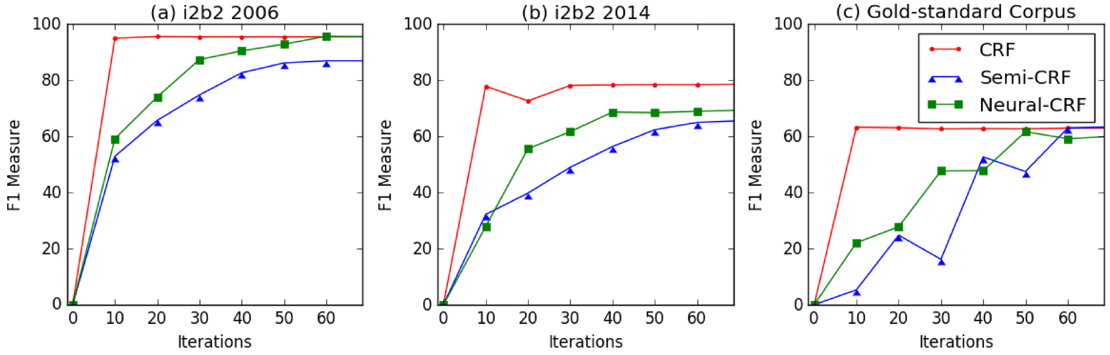


Figure 5.7: F1 trend in three variants of CRF on three data-sets: (a) I2b2 2006, (b) I2b2 2014, (c) PhysioNet corpus

Overall Performance

Overall comparing the models based on their trend in F1 scores of each model, shown in Fig 5.7, traditional CRF model reached a high F1 score of 0.955 (discharge summaries), 0.78 (longitudinal medical records) and 0.63 (nursing notes) for no less than

10 iterations. The values do not change with more iterations. But in the case of semi-CRF and neural-CRF models, the F1 score showed a steady increase, unlike the core CRF model which reached optimal performance too early. Though the F1 score is lower compared to MIST CRF, this steady increase in scores motivated me to test the efficiency of the two models (neural-CRF and semi-CRF) beyond 60 iterations. And when tested, the models showed only minor changes in their performance for I2b2 2006 and I2b2 2014 data-sets. But it is seen in Fig 5.7 that neural-CRF models showed an increase in the F1 score for all types of medical notes.

From the observations and discussion, the following conclusions are made regarding the model's performance on each data-set:

1. CRF models showed a consistent behavior by performing well on all three data-sets compared to semi-CRF and neural-CRF because traditional CRFs learned the dependencies among the input words in a sequence with the help of the rich and complex features. This type of model is most suitable for discharge summaries, as it took minimal effort in tuning the model parameters to achieve a F1 of 0.96.

2. Semi-CRF model is recommended for discharge summaries. The model in general can be recommended to any data-set, as it has the facility to deal with segment level features unlike the traditional CRF but requires large training iterations to reach an optimal performance.

3. Neural-CRF is more likely for all types of medical records. The model could produce better recall (0.56) with the given training exemplars for PhysioNet, which CRF (0.486) and semi-CRF (0.502) could not produce. The model could perform better given enough exemplars for each label. Neural-CRF models are more likely to replace the traditional CRFs because of their neural architecture which yields better performance.

It is apparent that neural-CRF models could perform well with an increase in the number of neural features and the number of neural gates for each model. Achieving such surprising results with an experimental variant of MIST motivates the research towards using deep learning models for the task of de-identification. There have been some groundbreaking implementations like LSTM-based NeuroCRFs [60] for NER and the use of recurrent neural networks to de-identify medical notes [24]. All these provide scope for better feature engineering and better results. But in a standard RNN there are data flexibility issues where the input data is fixed and the input information for a future state cannot be reached from the current state. An alter nate to these limitations is Long Short-Term Memory [35] type RNN models. From the findings, I implemented a bidirectional LSTM based de-identification model. Bi-directional LSTMs have two hidden layers, in opposite directions, connected to the output layer that allows information from past and future states. This makes it more suitable for context based learning. LSTMs have the ability to learn and store information about input sequences and are good in dealing with sequence labeling problems. Also they do not require a complex feature extraction system like those used in a normal CRF based model. The preliminary results from the training and test run are positive for the de-identification task. A detailed analysis of this model could be a path for future work.

Chapter 6

Conclusion

In a data-driven era, sharing any form of an individual's data potentially exposes them to privacy risks. Within healthcare industries, compromising an individual's privacy or exposing an individual's personal health information may lead to a uniquely personal kind of privacy breach. To avoid the exposition of patient's identity, the patient identifiable information needs to be removed before the medical records are shared among researchers. Removal / redaction of PHI is one of the important sequence labeling task (de-identification) in NLP. Systems that rely only on patterns and dictionaries, specific to a single domain, lack the ability to perform well on other data-sets. Machine learning techniques aid in building models that can be trained and used on any type of data-sets. This motivated the research presented here on machine learning models for de-identification.

In my thesis, I described an approach to evaluate free-text analysis techniques in natural language processing to de-identify medical records. For this, I evaluated the performance of machine learning models and their efficiency. I used MITRE MIST and its variants: CRF, semi-CRF, and neural-CRF models, for evaluation. The design and implementations of these machine learning models were described, followed by a detailed discussion of types of features and tuning of model parameters. The models were tested on three different types of medical records to assess the consistency in performance. All three data-sets were carefully chosen from de-identification challenges [79, 27] that produced benchmark results with these data-sets. To evaluate the predictive ability of the models, I used 10-fold cross-validation technique. All the data-sets were preprocessed and converted to the format used by the models.

The models were evaluated based on the performance metrics: precision, recall and F1 measures. For this task, the evaluation criteria is that recall carries more weight than precision because labeling any PHI token as OTHER token might lead to a compromise in the patient's identity. But in cases like medical research where precision is also important, F1 measure over-weights recall. Results in chapter 5 suggest that the core CRF models showed better performances on discharge summaries and longitudinal medical records. The semi-CRF variant produced higher recall measures and lower precision values with the help of segment level features making it a near-perfect model in terms of evaluation criteria. Having said that, the models took a large number of iterations of training to achieve that result. Of the three variants, neural-CRF models showed outstanding results in most cases with fewer training iterations compared to semi-CRF models. The performance mostly relied on the count of training exemplars involved. Though traditional CRFs produced consistent results, the models reached optimal values very soon and did not show any further improvements. But semi-CRF and neural-CRF models showed a gradual increase in the performance, making them suitable for further analysis.

With the exploratory results from various models on different kinds of clinical records, I found that context-based learning models (MIST variants), with large training datasets, produced the best results. The design of CRF allows them to learn the contextual information of a given word from the previous labels, which is important for a task like de-identification. This underscores the importance of using model-driven approaches that can accommodate arbitrary (i.e., not pre-defined) and contextual information. Rule-based models work on the list of rules created by the user while context-driven models or machine learning models learn the rules automatically thus making them more suitable for different types of data-sets. These results help understand various complexities that should be considered when choosing a de-identification system for a dataset.

Advances in artificial neural networks applied particularly to areas of natural language processing might soon take overtake CRFs and related variants in automatic de-identification. Results from neural-CRF directed my research towards deep learning models. The limitations in a standard neural network and that of a recurrent neural network made me choose models that were advanced in overcoming those limitations. Working closely with deep learning models that used long short-term memory based [35] recurrent neural networks helped deal with complex feature engineering in CRFs. Bi-directional LSTMs learn the input information from past and future states thus improving the context-based learning for a given input sequence. The experimental results produced by the LSTM model on discharge summaries show positive signs for further improvements and analysis on this model. My findings in this thesis could be further used to evaluate such new models.

Bibliography

[1] Atkinson's Spell Checking Oriented Word Lists.

[2] FlyBase.

[3] LingPipe.

[4] Mouse Genome Informatics.

[5] Privacy Rights Clearinghouse's Chronology of Data Security Breaches.

[6] PubMed Central Homepage. National Institutes of Health (NIH).

[7] The Saccharomyces Genome Database (SGD).

[8] UniProtKB/Swiss-Prot. European Molecular Biology Laboratory.

[9] Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics 50* (2014), 142 – 150.

[10] 75health. What Components Constitute an Electronic Health Record?, April, 2017.

[11] Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B., and Hirschman, L. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *International Journal of Medical Informatics*, 12, 849–859.

[12] Abramson, M. Sequence Classification with Neural Conditional Random Fields. *CoRR abs/1602.02123*, 1–6.

[13] Aramaki, E., Imai, T., Miyo, K., and Ohe, K. Automatic deidentification by using sentence features and label consistency. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data* (2006), vol. 2006, p. 1011.

[14] Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mashuichi, H., and Ohe, K. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (2009), BioNLP '09, pp. 185–192.

[15] Berner, E. S., Detmer, D. E., and Simborg, D. Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. *Journal of the American Medical Informatics Association 12*, 1, 3–7.

[16] Bodnari, A., Deléger, L., Lavergne, T., Névéol, A., and Zweigenbaum, P. A supervised named-entity extraction system for medical text. *CEUR Workshop Proceedings 1179* (2013), 1–8.

[17] Bottou, L. *Une Approche théorique de l'Apprentissage Connexionniste: Applications à la Reconnaissance de la Parole*. PhD thesis, Université de Paris XI, Orsay, France, 1991.

[18] Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010* (Heidelberg, 2010), Y. Lechevallier and G. Saporta, Eds., Physica-Verlag HD, pp. 177–186.

[19] Cerf, V. PARRY encounters the DOCTOR. *RFC 439* (1972).

[20] Chamberlain, B. *The Policeman's Beard Is Half Constructed*. UbuWeb, Warner Books, 1984.

[21] Chang, F., and Gupta, N. Progress in electronic medical record adoption in Canada. *Canadian Family Physician 61*, 12 (2015), 1076–1084.

[22] Chowdhury, G. G. Natural language processing. *Annual Review of Information Science and Technology 37*, 1, 51–89.

[23] Deleger, L. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association 20*, 1 (2013), 84–94.

[24] Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association 24*, 3 (2017), 596–606.

[25] Do, T., and Artieres, T. Neural conditional random fields. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010), Y. W. Teh and M. Titterington, Eds., vol. 9 of *Proceedings of Machine Learning Research*, PMLR, pp. 177–184.

[26] Fernandes, A. C. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Medical Informatics and Decision Making 13*, 1 (2013).

[27] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-k., and Stanley, H. E. PhysioBank, PhysioToolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation 101*, 23 (2000), e215–e220.

[28] Grishman, R., and Sundheim, B. Message Understanding Conference-6: a brief history. *COLING '96 Proceedings of the 16th conference on Computational linguistics*, 466–471.

[29] Guillen, R. Automated De-Identification and Categorization of Medical Records. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data.

[30] Gunter TD, Terry NP. The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. *J Med Internet Res 7*, 1, e3.

[31] Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G., and Hepple, M. Identifying Personal Health Information Using Support Vector Machines. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data* (2006), 10–11.

[32] HealthIT.gov. Office of the National Coordinator for Health Information Technology. Breaches of Unsecured Protected Health Information, Health IT Quick-Stat #53, February, 2016.

[33] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their Applications 13*, 4 (July 1998), 18–28.

[34] Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics 6*, SUPPL.1 (2005), 1–10.

[35] Hochreiter, S., and Urgen Schmidhuber, J. Long Short-Term Memory. *Neural Computation 9*, 8 (1997), 1735–1780.

[36] Howard Burde, J. Health law the hitech actan overview. *American Medical Association Journal of Ethics 13*, 3 (2011), 172–175.

[37] Hsiao, C.-J., Hing, E., and Ashman, J. Trends in electronic health record system use among office-based physicians: United States, 2007-2012. *National health statistics reports*, 75, 1–18.

[38] Huang, H. S., Chang, Y. M., and Hsu, C. N. Training Conditional Random Fields by Periodic Step Size Adaptation for Large-Scale Text Mining. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (Oct 2007), pp. 511–516.

[39] Hutchins, J. The first public demonstration of machine translation : the Georgetown-IBM system , 7th January 1954. *MT News International*, 8 (1994), 15–18.

[40] Jensen, P. B., Jensen, L. J., and Brunak, S. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 6, 395–405.

[41] Jones, E. HIPAA Protected Health Information: What Does PHI Include?, September 1, 2009.

[42] Jordan, M. I. Graphical Models. *Statistical Science*, 1, 140–155.

[43] Judea, P. Bayesian Netwroks: A Model of self-activated memory for evidential reasoning. *Annual Conference of Cognitive Science Society*.

[44] Jurafsky, D., and Martin, J. H. *Part-of-Speech Tagging.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[45] Krallinger, M., Erhardt, R. A. A., and Valencia, A. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 6, 439–445.

[46] Kruse, C. S., Goswamy, R., Raval, Y., and Marawi, S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Medical Informatics*, 4, e38.

[47] Lafferty, J., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. No. June, pp. 282–289.

[48] Lavergne, T., and Yvon, F. Learning the Structure of Variable-Order CRFs: a finite-state perspective. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 433–439.

[49] Lehman, L.-w., Moody, G., Heldt, T., and Kyaw, T. H. NIH Public Access. *Critical Care 39*, February 2010 (2011), 952–960.

[50] Liu, Z., Chen, Y., Tang, B., Wang, X., Chen, Q., Li, H., Wang, J., Deng, Q., and Zhu, S. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of Biomedical Informatics 58* (2015), S47–S52.

[51] Luo, X. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (2005), HLT '05, pp. 25–32.

[52] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. Association for Computational Linguistics, pp. 55–60.

[53] McCallum, A., Freitag, D., and Pereira, F. F. Maximum Entropy Markov Models for Information Extraction and Segmentation. *Proceedings of ICML*, 591–598.

[54] McCallum, A., and Li, W. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, 188–191.

[55] McDonald, C. J. The Barriers to Electronic Medical Record Systems and How to Overcome Them. *Journal of the American Medical Informatics Association*, 3, 213–221.

[56] Miller, R. A. Medical diagnostic decision support systemspast, present, and futurea threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association 1*, 1 (1994), 8–27.

[57] Neamatullah, I., Douglass, M. M., Lehman, L. W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making 8*, 1 (2008), 1–16.

[58] Office of Civil Rights Department of Health and Human Services. Summary of the HIPAA Privacy Rule. *OCR Privacy Brief* (2003).

[59] Rollman, B. L., Hanusa, B. H., Gilbert, T., Lowe, H. J., Kapoor, W. N., and Schulberg, H. C. The Electronic Medical Record. *Archives of Internal Medicine*, 2, 189.

[60] Rondeau, M.-A., and Su, Y. LSTM-Based NeuroCRFs for Named Entity Recognition. In *Interspeech 2016* (2016).

[61] Sahner, R., Trivedi, K. S., and Puliafito, A. *Semi-Markov Chains.* Springer US, Boston, MA, 1996, pp. 143–149.

[62] Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjolander, K., Underwood, R. C., and Haussler, D. Stochastic context-free grammars for tRNA modelling. *Nucleic Acids Research 22*, 23 (1994), 5112–5120.

[63] Sang, E. F. T. K., and De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. 142–147.

[64] Sang, E. F. T. K., and Veenstra, J. Representing Text Chunks. *CoRR cs.CL/9907006*.

[65] Sarawagi, S., and Cohen, W. W. Semi-markov conditional random fields for information extraction. In *Proceedings of the 17th International Conference on Neural Information Processing Systems* (2004), NIPS'04, pp. 1185–1192.

[66] Schuster-Böckler, B., and Bateman, A. An Introduction to Hidden Markov Models. *Current Protocols in Bioinformatics 18*, 1, A.3A.1–A.3A.9.

[67] Sha, F., and Pereira, F. Shallow Parsing with Conditional Random Fields. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, June, 134–141.

[68] Shen, D., Sun, J.-t., Li, H., Yang, Q., and Chen, Z. Document Summarization using Conditional Random Fields. *Science*, 2862–2867.

[69] Stephenson, T. A. *IDIAP research report*, Idiap-RR-03-2000.

[70] Stubbs, A., Kotfila, C., and Uzuner, Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics 58* (2015), S11–S19.

[71] Stubbs, A., and Uzuner, Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics 58* (2015), S20–S29.

[72] Subba, R. *Discourse Parsing: A Relational Learning Approach*. PhD thesis, Chicago, IL, USA, 2008. AAI3345494.

[73] Sutton, C., and Mccallum, A. An Introduction to Conditional Random Fields for Relational Learning. *Graphical Models*, 93.

[74] Swain, P. H., and Hauska, H. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 3, 142–147.

[75] Sweeney, L. Replacing personally-identifying information in medical records, the Scrub system. *AMIA Annu Symp Proc*, 333–7.

[76] Szarvas, G., Farkas, R., and Busa-Fekete, R. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association 14*, 5, 574–580.

[77] Thomas, S. M., Mamlin, B., Schadow, G., and McDonald, C. A successful technique for removing names in pathology reports using an augmented search and replace method. *Proceedings of the AMIA Symposium* (2002), 777–781.

[78] Uzuner, Ö., Sibanda, T. C., Luo, Y., and Szolovits, P. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine 42*, 1 (2008), 13–35.

[79] Uzuner, ., Luo, Y., and Szolovits, P. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association 14*, 5 (2007), 550–563.

[80] Velardi, P., and Cucchiarelli, A. A theoretical analysis of context-based learning algorithms for word sense disambiguation. In *Proceedings of the 14th European Conference on Artificial Intelligence* (2000), ECAI'00, pp. 451–455.

[81] Weizenbaum, J. Eliza&mdash;a computer program for the study of natural language communication between man and machine. *Commun. ACM 9*, 1 (Jan. 1966), 36–45.

[82] Wellner, B. *jCarafe*.

[83] Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J., and Hirschman, L. Rapidly Retargetable Approaches to De-identification in Medical Records. *Journal of the American Medical Informatics Association 14*, 5 (2007), 564–573.

[84] Xue, J. H., and Titterington, D. M. Comment on "On Discriminative vs. Generative classifiers: A Comparison of Logistic Regression and Naive Bayes". *Neural Processing Letters 28*, 3, 169–187.

[85] Zhou, G., and Su, J. Named Entity Recognition using an HMM-based Chunk Tagger. *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, July, 473–480.

[86] Zhou, G. D., Zhang, J., Su, J., Shen, D., and Tan, C. L. Recognizing names in biomedical texts: A machine learning approach. *Bioinformatics 20*, 7, 1178–1190.

[87] Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., and Shen, B. Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, 2, 200–211.

Appendix A

A.1 Regular Expressions - MIST

Table A.1: Regular expressions in MIST AMIA De-identification Task.

Function	Regular Expression
Initials Capital letters	([A-Z].*$)
Starts with Upper case	([A-Z][a-z]*$)
Starts with Upper case and alphanumeric	([A-Z]+$)
Lower and Upper case	([A-Za-z]+$)
Contains Digits	(.*[0-9].*$)
Single Digit	([0-9]$)
Double Digit	([0-9][0-9]$)
Four Digits	([0-9][0-9][0-9][0-9]$)
Natural Numbers	([0-9]+$)
Real Numbers	([0-9]\\.[0-9]+$)
Alphanumeric	([0-9A-Za-z]+$)
Roman Numerals	([lcximvLCXIMV]+$)
Contains ’-’	(.*-.*$)
Starts with ’-’	(ˆ-.*$)
Ends with ’-’	(.*-$)
Punctuation	(ˆ[ˆA-Za-z0-9]+$)
Capital Letters	(ˆ[A-Z].*$)

A.2 DE-ID Lists and Dictionaries

Table A.2: Dictionaries and lists used in PhysioNet DE-ID system.

Lists & Dictionaries	**Description**
Unambiguous Country names	Country names unambiguous to PHI terms
Unambiguous Company names	Company/organization names unambiguous to medical terms
Ambiguous Company names	Company/organization names ambiguous to medical terms
Unambiguous Ethinicities	Ethinicity names ambiguous to medical
Unambiguous Locations	Locations unambiguous to medical terms
Ambiguous Locations	Locations ambiguous to medical terms
Unambiguous Local places	Cities and street names unambiguous to medical terms.
Ambiguous Local places	Cities and street names ambiguous to medical terms.
Doctor First names	First names of doctor. Two lists for ambiguous names and unambiguous terms.
Doctor Last names	Last names of doctor. Two lists for ambiguous and unambiguous terms.
US States	List of states in US.
US States Abbreviation	Abbreviations of states in US
Female Names	Female first and last names. Two files for ambiguous and unambiguous names.
Male Names	Male first and last names. Two files for ambiguous and unambiguous names.
Prefixes Unambiguous	Unambiguous prefix terms used in notes.
Hospital Names	List of hospital names in US.