MANCHESTER
1824
The University of Manchester

Extraction of Temporal Information from Clinical Free Text

A dissertation submitted to The University of Manchester for the degree of
Master of Science in Advanced Computer Science
in the Faculty of Science and Engineering

Year of submission
2022

By
Hangyu Tu
10816601

Department of Computer Science

Contents

Contents

Word count: 13578

List of figures

List of tables

Abstract

With the introduction of information technology in the medical domain, the level of automation in the industry has been continuously improved. Clinical text information processing technology is gradually becoming a new research hotspot. Clinical texts, represented in electronic medical records (EMRs), contain rich medical information and are essential for disease prediction, personalised information recommendation, clinical decision support, and medication pattern mining and measurement. Despite the rich medical knowledge in medical texts, it is also more challenging to process, because the medical texts mainly include many the doctor's own input and may lead to spelling errors in the text, abbreviations of medical terms, and idioms of different doctors in different regions. The medical information contained has not yet been effectively utilised by a computer. Therefore, machine learning and natural language processing (NLP) related technologies, on a large scale, play an essential role in analysing and mining medical texts. In order to better explore and utilise medical texts, especially the semi-structured and unstructured information of EMRs, it is very important to standardise and structure the unstructured free texts. Traditional text classification requires a series of preprocessing and feature engineering modelling. Medical texts are analysed and processed, and ultimately valuable information and knowledge need to be generated to assist in decision-making, such as mining patients' medication temporal information patterns from EMRs and recording the timeline for how long the medications were taken to help doctors' diagnosis and medication decisions and even provide personalised clinical pathways. To achieve this, machine learning-based methods (including Bidirectional Long Short-Term Memory Neural Networks, Conditional Random Field and Convolutional Neural Networks) are applied for Medication Named Entity Recognition. A BERT-CNN pre-trained model is addressed to extract the relations between medications, medical Events and time. I have achieved a 0.98 weighted average accuracy and 0.69 of Macro averaging for medication NER task and 64.48% of precision, 67.17% of recall and 65.03% of f1-score for RE task. Besides, a BERT-based CNN has been implemented achieving exact 64.48% of precision, 67.17% of recall and 65.03% of f1-score for relation extraction task. Finally, I have attained a approximate 33% of prediction accuracy for medication temporal information extraction

Keywords: Text Mining, machine learning, NLP, CRF, CNN, BiLSTM, BERT

Declaration of originality

I hereby confirm that no portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

1. The author of this dissertation (including any appendices and/or schedules to this dissertation) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

2. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements in which the University has from time to time. This page must form part of any such copies made.

3. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the dissertation, for example, graphs and tables ("Reproductions"), which may be described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

4. Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses.

1 Introduction

As early as the 1960s, Weed et al. (1968) digitised paper medical records and proposed problem-oriented management of EMRs in order to help medical staff with diagnostic reasoning. With the development of time, clinical decision-making based on EMRs data still attracted much attention in this century.

1.1 Background and Motivation

With the continuous growth of the need for medical text mining, extraction of temporal information (e.g. when the disease was diagnosed, how long some medicine a patient should take) has become a significant field of Natural Language Process (NLP) (Su et al., 2004). Due to its importance for many other NLP tasks, temporal information tagging and annotation from clinical texts have become an active research field over the years. Clinical texts (including electronic health records (EHRs), which consist of “medical history, diagnoses, medications, treatment plans, immunisation dates, allergies, radiology images, and laboratory and test results”) contain a lot of useful information which provides the opportunity to support medical research (e.g. effect of medication time on side effects Fredriksen et al. (2014)), create systems that enable the improvement of medical care quality such as: to help doctors make better decisions for patients having the precise and thorough information including patients medical condition and history of drugs. Furthermore, finding the correlations between drug use and its outcomes can be vital and helpful in digging out the side effects of the drug.

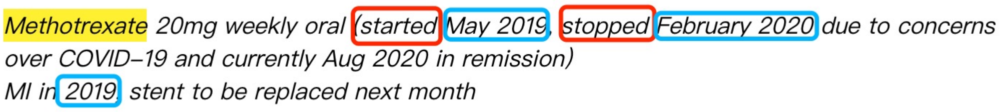
In view of the importance of temporal information in clinical free texts, in order to extract it from unstructured data and convert it to structured features consisting of temporal relations (TLINKs) between events and time expressions, researchers have developed some techniques, tools and workflow. Generally, they can be divided into three main categories: rule-based approaches, machine learning and combined approaches.

Moreover, most machine learning methods rely on feature-engineered supervised learning to predict temporal relations between events, which require tagged datasets that cost a lot of effort and time of experts in the medical domain.

Therefore, this dissertation aims to investigate and explore machine learning (CRF) and deep learning (CNN and BiLSTM) models from a perspective of how a model could be applied to recognise the Medication Entity effectively and extract the relations between the Medication Entity and corresponding time baseline. Two major tasks were a focus. Medication Entity Recognition, which involves finding and identifying named drugs mentioned in unstructured text, is the first step toward automating the extraction of medications from EMRs. This topic is utilised by numerous NLP programs that address use cases such as machine translation, information retrieval, and chatbot.)which is a multi-classification task. A lot of work has been done, and methods have been proposed by (Chang et al., 2013; Hartley et al., 2017; Lin et al., 2013). Second, automatic temporal relation extraction from free text EMRs. This dissertation addressed and applied the BiLSTM+CNN model, BiLSTM+CRF model and BERT-based CNN model to train on the dataset. Additionally, a SparkNLP tool (DateNormalizer) (Kocaman and Talby, 2021) was used to extract & normalize dates from relative date-time phrases. The main achievement of this dissertation that demonstrates that deep learning models perform better than traditional machining learning methods on NLP classification tasks.

1.2 Aims and Objectives

The primary purpose of this dissertation is to extract temporal information related to diagnoses, drugs and symptoms from free clinical texts (e.g. 'Methotrexate 20mg weekly oral (started May 2019, stopped February 2020 due to concerns over COVID-19 and currently Aug 2020 in remission)MI in 2019, stent to be replaced next month') and convert it to structured data (as illustrated below) to a structured table and for those do not appear in the text being tagged 'Unknown').



Methotrexate 20mg weekly oral (started May 2019, stopped February 2020 due to concerns
over COVID–19 and currently Aug 2020 in remission)
MI in 2019, stent to be replaced next month

Fig. 1. Original Text

Methotrexate	
IN USE	OFF
start = May 2019	start = February 2020
stop = February 2020	stop = Unknown

Table 1. Medication Temporal Information Extraction Result Example

This dissertation focuses on building a system which is used to extract temporal information from texts facilitating the work in hospitals (e.g. help doctors attain a better picture of patients' medical history) and research in the clinical text mining domain (e.g. with incorporating extracted temporal information as a feature for any other NLP tasks). To build up this system with the ability to extract temporal information automatically, there are several specific steps:

- Review literature and come up with the annotation schema for adding temporal features with available tools.

- Apply state-of-the-art ruled-based and machining learning-based methods to recognise EVENTs and TIMEx (time expressions) mentions in texts and explore the possibility of improving them.
- Extract features related to a DRUG event and classify these events into appropriate categories.
- Find whether this model works for MEDICATION events.
- Link relations to their corresponding entity.

Finally, with this model I planned to build, a structured temporal feature is extracted for further research and medicinal purposes.

1.3 Report Structure

The remainder of this dissertation is organized as follows. In this dissertation, Chapter 1 gives an introduction of what this dissertation focuses on and what its goal is. In Chapter 2, it presents the relevant background and discusses what significant work has been done regarding clinical information extraction research which includes 3 main techniques including rule-based, machine learning, and hybrid methods. Chapter 3 explains all the techniques, tools, and models involved in this dissertation and clarifies the complexity of classic NER and RE statistic models. The procedure of data pre-processing, model construction and evaluation are demonstrated in Chapter 4. Finally, I conclude this dissertation with future work in Chapter 5.

2 Literature Review

Researchers have done significant studies on clinical text mining in the past few years. These research findings facilitate data-driven projects in hospitals (help doctors make better diagnoses and treat patients, and reduce the unnecessary process of drug use history enquiries). This section discusses the existing methods and research results I have surveyed in the extraction of temporal information. Typically, information extraction consists of Named Entity Recognition (NER) and Relation Extraction (RE). The research methods of NER and RE can be divided into rule-based and statistics-based methods (including traditional machine learning-based and deep learning-based), and the two methods can also be combined. In terms of accuracy, rule-based methods perform better than traditional statistics-based methods. However, with the development of deep learning neural networks, the performance of methods based on deep learning exceed that of rule-based methods Pustejovsky et al. (2003). However, in many cases, the establishment of rules is often related to the language environment and must be completed with the cooperation of experienced language experts. In contrast, statistics-based methods do not require extensive linguistic knowledge, which is acquired through machine learning and has good portability. The model only needs to be retrained during transplantation on a new corpus, but statistics-based methods require a large corpus for learning. So there are advantages and disadvantages to both approaches.

2.1 Rule Based Methods

Ruled-based methods, which require experts with specific specialities to make a set of reasonable rules, are objective and vary from person to person with different opinions towards this domain. Technically, rule-based methods often constitute a set of regular expressions. They are encoded by experts with domain knowledge and experience based on linguistic analysis, which requires great time and effort. For NER rule-based methods, rules often based on the named entity itself and the context of a situation, are generally formulated by linguists. Compared with other languages (such as Chinese), it is relatively easy for English corpora to name entities because English name entities usually start with a capital letter, such as name, organisation name, etc., all start with a capital letter. If the named entity is identified, then the corresponding rules can be used to determine which type the named entity belongs to. In various corpora, the rules are constantly improved through the error analysis of rule-based experimental results until a certain recognition accuracy is achieved.

Strötgen and Gertz (2013) built a rule-based system pattern called HeidelTime that expects part-of-speech(POS) tagged sentences and user custom parameters determining what kinds of temporal expressions are specified to extract temporal expressions and obtained high-quality results, which are 90%, 82% and 86% in Precision, Recall, and F1-score respectively. Reeves et al. (2012) used handcrafted rules to annotate temporal expressions (TIMEx). They considered each TIMEx as a tuple $t_i \in \{date; time; duration\}$. Basically, they made use of regular expressions to construct patterns to match sentences with similar expressions using the designed rules and achieved at least an increase of 5% in the precision score.

In general, rule-based methods have been widely used in the last 2 decades and some of them (Roberts et al., 2013; Lin et al., 2013) achieved remarkable results in Medical terminology (Peek et al., 2013) and TIMEx (Hao et al., 2018) extraction tasks. Hao et al. (2018) first produced a set of rules based on heuristic theory and specific temporal features training on clinical texts to generate structural patterns that work on other pattern-based temporal relations. This system achieved 0.916 of F1 score in extracting temporal expressions on Chinese clinical texts, which continually hold influence over this research field. Rule-based methods can perform well in event recognition; however, we can not ignore that developing a comprehensive and heuristic set of rules is a rather time-consuming and difficult task. These rules often depend on specific languages, domains and text formats, which are difficult to achieve a broad language coverage. Therefore, it is difficult for humans, even experts, to consider the correct rules thoroughly in all aspects. Wang et al. (2012) who generated rules based on syntactic information, including dependency types and conjunctions for information extraction from biomedical literature, achieving an increass by 3%-5% than (Coulet et al., 2010). In order to investigate whether the existing medical terminologies work on Sweden language clinical texts as well as it does on English language texts and different techniques for pre-processing that might affect the performance of the corresponding system, Skeppstedt et al. (2012) proposed a system using different pre-processing methods that evaluated existing systems (including SNOMED-CT (Donnelly et al., 2006)) for Sweden clinical texts. The main rule of this system for selecting annotation categories is to annotate words into categories that perceive them in clinical reality as described in the text. Thus, the same word or expression may be a discovery in one case and a disease in another, and it is annotated according to context. Hypertension, for example, can be seen as a symptom of multiple diseases, but it is also a disease in its own right, with its own underlying pathological processes. With this system, they achieved a precision of 0.74 and a recall of 0.80 on the body structure entity. Uzuner et al. (2010) built a system to extract medical information (body parts, anatomical parts, surgery reason, disease)for automated clinical data processing. Uzuner et al. (2010) proposed two apps designed to select data from Polish-language medical documents: mammogram reports and electronic records of diabetics. For each domain (body part, disease or medication), they defined a set of appropriate rules that enabled them to identify its value. For example, the syntax for the mammogram report contains 190 rules, while the syntax for the diabetes field contains 150 rules. In the case of mammography, the rules cover almost the entire report, whereas, in the case of diabetes, we target only a subset of information. This system obtained significant results in precision and recall achieving more than 80% in most templates.

Another widely applied technical framework that facilitates users to build and develop natural language processing tasks (NER and IE) called GATE (Cunningham H, 2002), developed by the Natural Language Processing Group at the University of Sheffield with funding from the Engineering and Sciences Research Council. It is an open-source natural language processing platform from the UK. The three main purposes of GATE design are: 1. Provide a technical framework for software for various natural language processing tasks. 2. Provide component and class library services, which can be embedded in the natural language processing software. 3. Provide a development platform for a natural language processing system, allowing users to debug in the visual interface.

Despite all the hard work from other researchers, there still exists a vague definition regarding what could be a perfect rule as a system to extract name entities or relations between mentions. Besides some well-done research in the English language clinical free text information extraction, rule-based information extraction is still an intractable challenge. Besides, in general, a set of rules generated from a specific domain are hard to be extended to apply to other domain (Ji et al., 2019); thus, rule-based methods are no longer the first option for temporal information extraction. With the development of computation resources and machining learning, it allows researchers to build more complex models but only costs an acceptable time to run the model that could learn sufficient features and custom configuration to make one domain corpus distinguish from other fields.

2.2 Machine Learning-based Methods

Compared with computer vision and speech recognition, deep learning is a latecomer to natural language processing. On the one hand, because the text does not have digital features, the way of modelling text features is not mature. On the other hand, because the traditional classical network algorithm can not get the temporal features of text, it can not achieve good results in the application of text processing. Through the efforts of many scholars for more than ten years, the application of deep learning to natural language processing has become increasingly mature. Bengio et al. (2000) systematically proposed Neural Network Language Model (NNLM) and conducted in-depth research. One of the core ideas presented by NNLM is word vectors. With the proposal of a word vector, the neural network becomes more and more mature for the modelling. Prior to the invention of word vectors, researchers mainly extracted text properties using statistical techniques like One-hot (Chren, 1998) or TF-IDF (Wu et al., 2008). Although these two approaches can partially represent text, their drawbacks are extremely clear. First of all, One-HOT and TF-IDF are essentially just statistics of how frequently words occur in the text. Since each dimension in the vectors produced by this statistical method has distinct properties, they are unable to convey the text's semantic content. Second, each one-hot and TF-IDF dimension corresponds to a word in the lexicon. It is straightforward to get into excessive dimensions for a little larger corpus. It is unavoidable to sacrifice certain words in order to lessen the effects of excessive dimensionality, which can result in the input statement losing some information.

With the rapid development of artificial intelligence and machine learning, machine learning-based methods have been widely used in NLP. Machine learning generally falls into two categories: supervised learning and unsupervised learning. Supervised learning, of which application is used wider, consists of three main parts:

1. The extraction of feature to represent the data;

2. A model to work as a classifier;

3. Process of tuning the parameters of the model.

Basically, the main part that differs between machine learning-based (Categorized into supervised and unsupervised learning) methods is how the algorithm works for feature engineering. The difference between supervised and unsupervised feature selection is that one works with labelled classes, and the other does not. However, it could be very time-consuming work for people to label the raw data if doing supervised learning classification. The method of named entity recognition based on statistical machine learning has been deeply studied and applied in view of the shortcomings of dictionary and rule-based methods. Among various statistical machine learning methods, the support vector machine (SVM) is widely used. It plays a vital role in solving pattern recognition problems such as small samples, linear inseparability and high dimension. The method based on the Hidden Markov model (HMM) is also one of the common methods of named entity recognition in the medical field. Zhou et al. (2005) combined one SVM classifier and two HMM classifiers and found that different classifiers had different results on the dataset and could complement each other, and the F1 value on the genetic dataset reached 0.83. However, HMM is based on the assumption of independence. Strict independence assumption can not truly describe the information contained in the data sequence and can not understand the semantic relationship between the remote contexts of text sentences. The appearance of conditional random fields (CRFs) solves the independence assumption of the hidden Markov model, considers the context information, and solves the semantic connection between the remote contexts of text statements. Compared with feature-based methods, deep learning-based methods help discover hidden features automatically and automatically learn the required representation from the original input in an end-to-end manner. The application of neural networks in NER was first proposed by (Collobert et al., 2011), who deployed a model architecture based on temporal convolutional neural networks. When common prior knowledge is incorporated, the resulting system outperforms the baseline using only word-level representations. Wu et al. (2015) used convolutional layers to generate global features represented by several global hidden nodes. The local and global features were then fed into a standard affine network to extract named entities in clinical medical records that obtained a high F1-score of 92.8% in Chinese EMRs. Aguilar et al. (2019) proposed a multi-task approach for NER that uses CNN to capture positive spelling features and words at a character level. For the syntactic and contextual information at the word level, such as part of speech and word embedding, the model is implemented through the LSTM model architecture.

Jiang and Chen addressed that, with the CRFs algorithm, this system achieved at least 0.8475 in the F-measure (Jiang et al., 2011). For TLINKs detection, machine learning-based shows a remarkable effect on this classification task (Chang et al., 2013). Focusing on CRFs and support vector machine (SVM), Xu et al. (2013) proposed a system with three sub-systems, including clinical events extraction, temporal information extraction and temporal relation extraction. In terms of prediction tasks, applying temporal information as a feature to improve the performance of the model has been proved a feasible way (Pang et al., 2021). Theoretically, RNN can process any long sequence, but with the continuous accumulation of time series, the gradient will decay exponentially, which makes it difficult for RNN to record long-distance historical information, and its performance is also restricted. In order to solve this problem, Hochreiter and Schmidhuber (1997) proposed the concept of the long short-term memory network(LSTM) and theoretically proved that this structure can well solve the problem of gradient disappearance and explosion.Dang et al. (2018) has addressed a bidirectional Long Short-Term Memory (an advanced deep learning neural network) which achieved 87.62% F1 score for the gene/protein NER and 93.14 and 84.68% for the chemical and disease NER, respectively. In the field of NLP, the most commonly used deep learning algorithm is based on the deep structure of the recurrent neural network. The traditional neural network can not process the continuous input of natural language with time-series characteristics. The improvement of RNN is that it adds a loop pointing to itself so that the network can make use of the sequence characteristics of the input, Wu et al. (2017) compared CNN and RNN with baseline CRF in the i2b2 2010 clinical concept extraction corpus. The results showed that the performance of RNN was better than the best clinical named entity recognition system based on SSVM, and the F1 value was 0.86.

Theoretically, RNN can process arbitrarily long sequences, but with the continuous accumulation of time series, the gradient will be exponentially attenuated, which makes it difficult for RNN to record historical information within a long distance, and its performance is restricted. However, the effect of the BiLSTM and CRFs deep learning method is still limited by the quality of the training set. Therefore, people began to pay attention to how to obtain a priori semantic knowledge from a large number of unlabeled texts to enhance the semantic representation and proposed a transformer-based pre-training model, such as the Bert model. Some researchers(Kim and Lee, 2020; Lee et al., 2020; Jiang et al., 2019) improved accuracy, recall, and F1 value by combining the Bert word embedding model with the conventional bilstm CRF model and taking into account the polysemy of a word in conjunction with the context. Bert, as we can see, has excellent semantic analysis skills. However, there is no separator between Chinese words, so Bert can only cover up words, not words when using the Chinese corpus for pre-training. Thus, the semantic representation generated by the pre-training is only at the word level, and the word level semantic representation cannot be obtained; that is, the word information cannot be obtained during the pre-training process. Recently, BERT, as a pre-trained Deep Bidirectional Transformers model for human language understanding, has been widely used in NLP related tasks, and a BERT-BiLSTM-CRFs (Ma et al., 2022) model was proposed in extracting temporal information from social media messages. As a matter of fact, the latest BERT adaptation (Rasmy et al., 2021) has introduced this adapted model Med-BERT with fine-tuning pre-trained on a structured EHR dataset which showed tremendous improvement in disease prediction tasks. Yuan et al. (2022) has designed the BiLSTM-CRFs model and BiGRU-Dual Attention model which achieved a significant F-1 score of 87.87% and 88.05% in the Archaeological site text domain. Despite the tremendous success of the BiLSTM-CRFs model in recognition performance in NER tasks and its ability to preserve long-term information, still, there is an unavoidable error in long sentence extraction, also called tagging inconsistency in (Luo et al., 2018). In order to cope with this problem, based on the standard BiLSTM-CRFs algorithm, Ji et al. (2019) proposed a system, introducing attention mechanism to BiLSTM-CRFs, and conducted this system on Chinese EMRs, which achieved an F1-score of 90.82% higher than other work on the test dataset.

Compared to traditional deep learning methods that require a large amount of corpus for training, transfer learning can transfer knowledge from the source domain to the target domain by relaxing the condition that training instances and test instances obey independent and identical distribution, which has a tremendous positive impact on many fields that are difficult to make progress due to insufficient training data. Instance-based deep transfer learning refers to adding some samples in the source domain to the target domain as a supplement to the training samples in the target domain through special weight adjustment measures and assigning appropriate weights to these selected instances. It is based on the assumption that although there are differences between the source domain and the target domain, some instances in the source domain can be utilised by the target domain with a specific weight. Ling et al. (2008) proposed a TrAdaBoost system that deletes samples in the source domain that are not similar to the target domain with AdaBoost-based technology. AskTrAdaBoost is a fast algorithm invented by (Yao and Doretto, 2010), which can speed up the retraining of new target domains. Different from TrAdaBoost, it is designed for classification problems.

However, to achieve efficient and outstanding performance, deep learning networks need a lot of reasonable feature engineering work, which is expert-required and time-consuming. With BiLSTM, words or tokens in sentences can be represented by required dimensional vectors (also called word embedding) to be taken as input fed to a Conditional Random Fields (CRFs) model(used for classification). Apart from these embedding vectors, which contain contextual information, Part-of-Speech(POS), Chunks and Dependencies(Dep) are widely used as features to improve model performance (Luo et al., 2018). Nevertheless, the extraction of some temporal phrases (e.g. next week, for 2 weeks) is still a problematic and prerequisite task.

2.3 Hybrid Approaches

Compared to the methods in 2.1 and 2.2, hybrid approaches interweave rule-based and machine learning-based methods to extract temporal information from clinical free texts. Chang, Yung-Chun and Dai addressed that their system performed better by 0.97% than those of single-stage based method on test dataset (Chang et al., 2013). Chang et al. (2013) also found that their prediction system's performance on TLINK AFTER is poorer than the other two categories because of the imbalance of the dataset. Lin et al. (2013) proposed a system of six main steps : (1) data preprocessing, (2) temporal tagging by HeidelTime (Strötgen and Gertz, 2013), (3) TIMEX and EVENT named-entity extraction , (4) classification, (5) clinical temporal normalization strategies, and (6) post-processing.

The rapid development of temporal information extraction methods, with the work of Pustejovsky (Pustejovsky et al., 2003), has prevailed among many researches since 2003. This specification language TimeML (Pustejovsky et al., 2003) takes into account the ordering of dependency existing in narratives that allows some of the temporal expressions to be decided later after interpretation, which makes a great contribution to question answering tasks. Tang et al. (2013)proposed a rule-based and machine learning-based combined method, and their system achieved top rank in The 2012 i2b2 NLP challenge. Their system contains three parts: Event extraction, Temporal expression extraction and time linking(TLink) extraction. Separating a CRFs based classifier combined with SVM into two-stages, this classifier allows for extracting medical tests, problems and treatments in the first stage and the other events are extracted in the second stage achieving 0.9374 of precision 0.9013 of F1-score.

Yang et al. (2015) addressed a hybrid method for adding temporal information as an additional feature to build a model that 'enhances the visualization accuracy of lung motion'. Atluri et al. (2018) has made great summaries on Spatio-Temporal Data Mining problems and methods. However, rule constraints inevitably require a lot of linguistic knowledge, which is not common among computer science researchers.

Different from traditional hybrid methods, Zhang et al. (2018) proposed a hybrid technique that applies CNN combined with traditional RNN, which takes advantage of the ability of CNN and RNN to learn potential features from biomedical literature. Then these learnt features are used to extract biomedical relations between proteins. Generally, this hybrid method, which takes dependency path and dependency word sequence as input, has shown significant performance in extracting relation in LLL (Pyysalo, 2008) corpus dataset that boosts a single CNN or RNN model effectively with 0.852 and 0.961 in F1-score and Recall, respectively.

Benefiting from the acceleration of research and development of computation, it takes less time for researchers to run and debug models. It gives more time for them to focus on what features contribute the most information to the model as well as the structure of Neural Networks. Basically, machine learning-based methods have prevailed in NER and relation extraction tasks. In addition, Cocos et al. (2017) used the advancements in deep learning technology to prove the benefits of deep neural network technology by comparing it to the conventional CRF model, finding that the former requires less intervention from artificial features than the latter does, hence allowing for more accuracy and recall. Deep learning can automatically extract word features, eliminate subjectivity in feature selection, and enhance recognition accuracy. In order to consider the context information, solve the semantic connection between the remote contexts of text sentences, and solve the gradient disappearance and explosion problem of the RNN model in the process of entity recognition, this paper applied the deep learning model BiLSTM-CRF to improve the accuracy of named entity recognition by using the context information.

2.4 Chapter Summary

In this chapter, rule-based methods, machine learning-based methods and hybrid methods (including the combination of rule-based and machine learning based as well as combining different machining learning models and deep learning models) are introduced regarding NER tasks and relation extraction tasks.

3 Research Methodology and Experimental Design

3.1 Dataset and Evaluation Method

Based on previous research direction for this dissertation, this combined temporal information extraction on clinical text system pipeline consists of 5 main elements (which will be demonstrated below). In this section, I propose my system in detail, and the pipeline of all the processes and techniques is designed as shown in figure 2. First, for pre-processing, I apply Sentence Segmentation, Tokenization, POS Tagging and Sysntactic to extract features from the raw data and obtain structured data for modelling input. Then some models are implemented for NER and RE tasks. Finally, with some rules, I generate some structured output as expected.

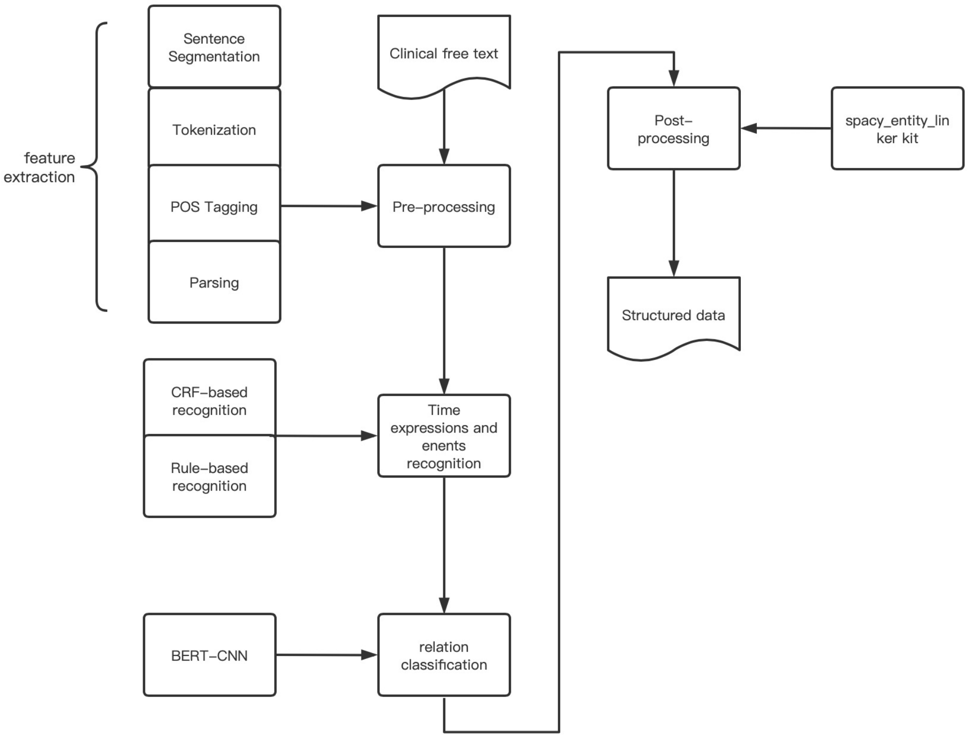


Fig. 2. System Pipeline

3.1.1 Dataset

Despite the popularity and importance of clinical free text mining research, given the sensitivity of patients (Schriever, 2014), there have not been too many public data sets available, and significant challenges are present when researchers are extracting useful information from clinical free text automatically. The EHRs data for this dissertation mostly comes from a National NLP Clinical Challenges (n2c2, formerly called i2b2, which are challenges and workshops designed and led by Özlem Uzuner et al. to find 'the potential of existing clinical records to yield insights that directly impact healthcare improvement') dataset. These free texts contain sensitive information. Therefore, in order to research the data, I will have to ask the agencies for permission. In addition, the data collection is out of my hands so that I have no choice of which data I could use, and some of the data might be useless or meaningless (e.g. full of noise and incomplete). These could be obstacles to this dissertation.

In this dissertation, n2c2 challenge track 2009 and 2012 datasets were used. The objective is to gather details about every drug that the patient is known to take for each patient report that is given. Partners Healthcare's discharge summaries will be the input for the medication challenge (Sun et al., 2012). These files are loosely organized. Many drugs are covered in the narrative text in addition to being listed by name in lists of pharmaceuticals. According to statistic analysis, it contains 447 EMRs, with 252 annotated, which could be used to train and test the model. While the data from the 2012 n2c2 temporal relations challenge has similar content as the challenge track, 2009 does but with a different format, additional entity categories and temporal relations between events. It consists of 760 files including 190 TXT files (texts), 190 XML.EXTENT files (annotation and position in original text), XML files(annotation and text) and TLINK files (relations) (Gurulingappa et al., 2012).

3.1.2 Evaluation Method

To deliver and evaluate the performance of the temporal information extraction system properly, the effectiveness (Accuracy, precision, recall, and F-score included) of the system will be analysed on the training data and test data. To evaluate the machine learning-based classifier, the confusion matrix is widely applied with true positives indicating an outcome where the model correctly predicts the positive class (TP), true negatives indicating an outcome where the model correctly predicts the negative class (TN), false positives indicating an outcome where the model incorrectly predicts the positive class (FP), and false negatives indicating a result where the model incorrectly predicts the negative category (FN) as measures. To make acknowledge whether our system, when finished, would facilitate the work in hospitals, it can be sent to a medical agency to conduct a trial on actual situations and give feedback on it.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

In terms of interpretation, extraction results will be demonstrated in a structured table (mentioned above) or CSV shape file with a comprehensive explanation and figures if necessary. Considering that there are multiple categories of labels, macro average evaluation (the function to compute precision, recall and F1-score, for each label and returns the average without considering the proportion for each label in the dataset) and weighted average evaluation (the function to compute precision, recall and F1-score for each label, and returns the average considering the proportion for each label in the dataset).

3.2 Data Pre-processing

Clinical texts are mixed with different formats. The target of pre-processing is trying to format different data with certain rules and enrich texts with their potential features.

Typically, Pre-processing follows 5 steps: Sentence Segmentation, Tokenization, Parts of Speech (POS, which explains how a word is used in a sentence), Tagging and Parsing (demonstrated in figure 3). For my dissertation, GATE (Cunningham H, 2002), cTAKES (Savova et al., 2010) and MedSpaCy (a new clinical text processing toolkit in Python) shows excellent power in building rule-based modules to retrieve extra features. Additionally, the library 'spacy' is valuable and powerful to link mentions to the associated verb. An example result of this pipeline output is shown as follows (table 2).

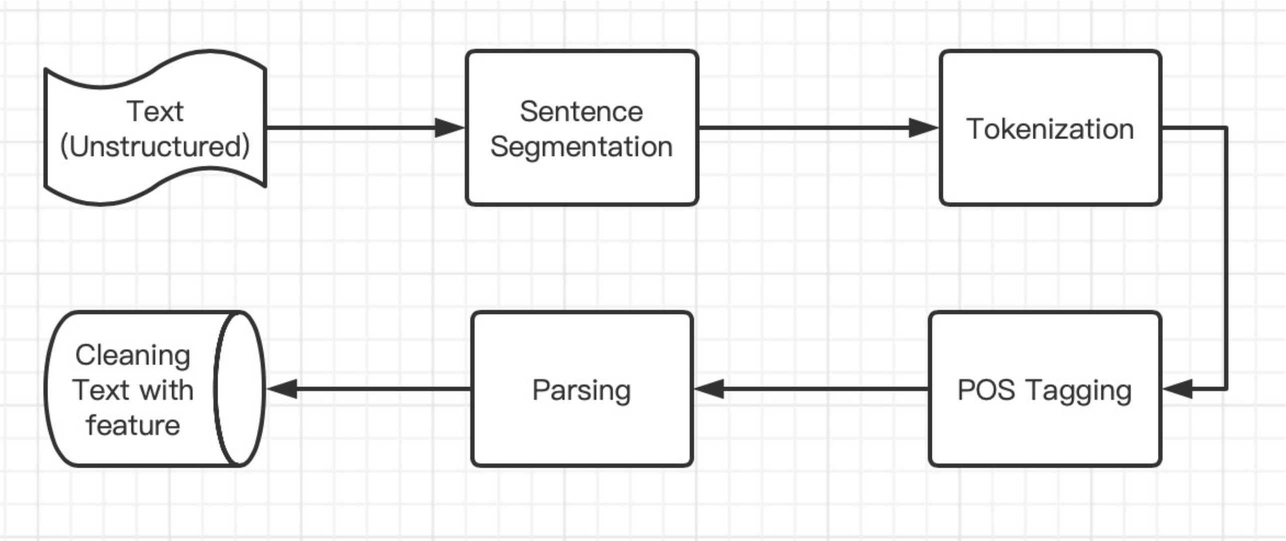


Fig. 3. Pre-processing Workflow

Original word	Lemmatization	POS	Tag	Explain
The	the	DET	DT	det
canning	canning	NOUN	NN	compound
process	process	NOUN	NN	nsubjpass
for	for	ADP	IN	prep
food	food	NOUN	NN	pobj
was	be	AUX	VBD	auxpass
invented	invent	VERB	VBN	ROOT
by	by	ADP	IN	agent
Nicolas	Nicolas	PROPN	NNP	compound
Appert	Appert	PROPN	NNP	pobj
by	by	ADP	IN	prep
1999	1999	NUM	CD	pobj
.	.	PUNCT	.	punct

Table 2. POS Tagging Example (Marcinkiewicz, 1994)

3.2.1 Sentence Boundary Disambiguation

The task of sentence boundary detection is to identify sentences within a text which appears first in an NLP task pipeline, and its performance could decide the following pipeline parts. Typically, the full stop is detected as the end of a sentence, which is obviously ambiguous and not enough since the full stop is not only used as a sentence boundary delimiter but also juxtaposed with abbreviations (e.g. M.r, A.M, Doc.) which appear quite a lot in clinical texts. Besides, full stop occurs in some special expressions including dates(2022.07.22), number(10.0 mm), and web links(https://www.google.com/). Sometimes, other delimiters like “!”, “?” could be the end of sentences as well.

In this dissertation, statistical sentence segmentation of spacy is used and refined with custom rules. For example, given sentence:
The patient ’s bilirubin level at 24 hours of life was 4.6.
Spacy returns three sentences:

1. The patient
2. ’s
3. bilirubin level at 24 hours of life was 4.6

Here I want to make sure that ” ’s ” tokens will never be the start of a sentence. So a rule is added to make sure ” ’s ” never occur at a start of a sentence.

3.2.2 Tokenization

Having paragraphs segmented into sentences, sentences will be split to tokens, also known as Tokenization, that is to decompose long texts such as sentences, paragraphs and articles into word-based data structures for the convenience of subsequent processing and analysis. Text is some unstructured data; we need first to transform these data into structured data, structured data can be converted into mathematical problems, and Tokenization is the first step of transformation.

Compared to other languages(Chinese), tokenization is easier for English corpuses since tokens could be separated by space. However, in order to preserve integral information of original texts, endings of contractions (”’re” in ”we’re”, ”’t” in ”doesn’t”) should receive special treatment (added rules) as mentioned in 3.1.1.

3.2.3 POS Tagging

POS Tagging, also called grammatical-tagging or word category disambiguation (the POS tags from the Penn Treebank project (Marcinkiewicz, 1994), which are widely used in practice), refers to the process of marking a word in the corpus as corresponding to a particular part of speech, based on both its definition and context and the output of POS Tagging is a good feature for supervised learning as an input which makes it an important step:

1. POS tagging is important for word sense disambiguation
2. For example, consider the sentence “They can fish”; One interpretation: They/PRON can/AUX fish/VERB; and another: They/PRON can/VERB fish/NOUN
3. POS tagging is of importance in applications such as machine translation and information retrieval.

3.2.4 Syntactic Parsing

In order to understand sentences, we cannot treat each word in isolation. We need to determine what a word is attached or connected to. Syntactic parsing is a technique by which segmented, tokenised, and part-of-speech tagged text is assigned a structure that reveals the relationships between tokens governed by syntax rules, e.g. by grammars. In NLP, rules of context-free grammar (CFG) or probabilistic context-free grammar (PCFG) are used to analyse sentences. Building a parser from scratch is a very complex task in itself. I would pick a grammar like CFG or PCFG, then decide upon which type of parser we want to build. Based on that, I would implement specific algorithms to develop my very own parser. For example, given the sentence ”The patient was not started on antibiotics”, the result of spacy parsing visualisation (figure 4):

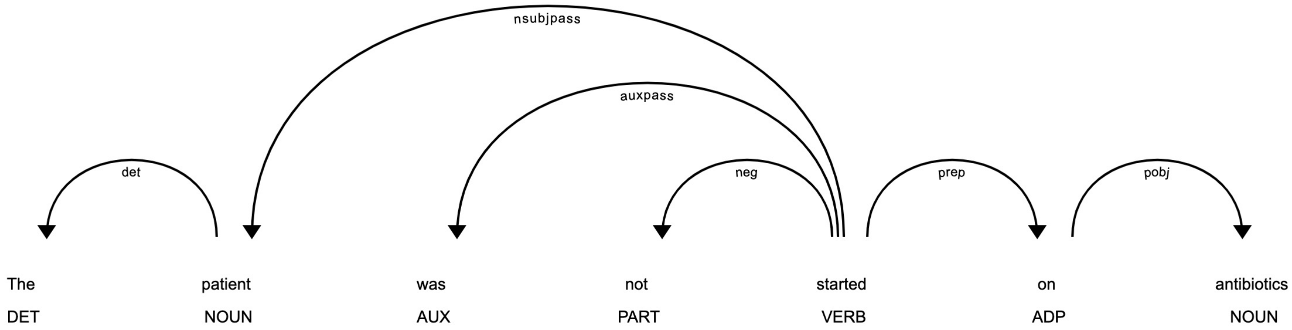


Fig. 4. Syntactic Parsing

3.3 Word Embedding

Word embedding techniques play a role in capturing the meaning, semantic relationship, and context of different words while creating word representations. A word embedding is a technique used to generate a dense vector representation of words that include context words about their own. In addition to one-hot embedding that results in sparse matrix embedding, there are enhanced versions of simple bag-of-words models, such as word counts and frequency counters, that essentially represent sparse vectors.

3.3.1 One-hot Encoding

A popular encoding technique is called ”one hot encoding,” and it allows a word to be represented as a vector in a format that is more amenable to algorithmic processing. This strategy begins by establishing a vocabulary and assigning numbers to each word. The vector’s dimension is the size of the entire vocabulary.For example, ”The patient was not started on antibiotics” encoded as follow (figure 5):

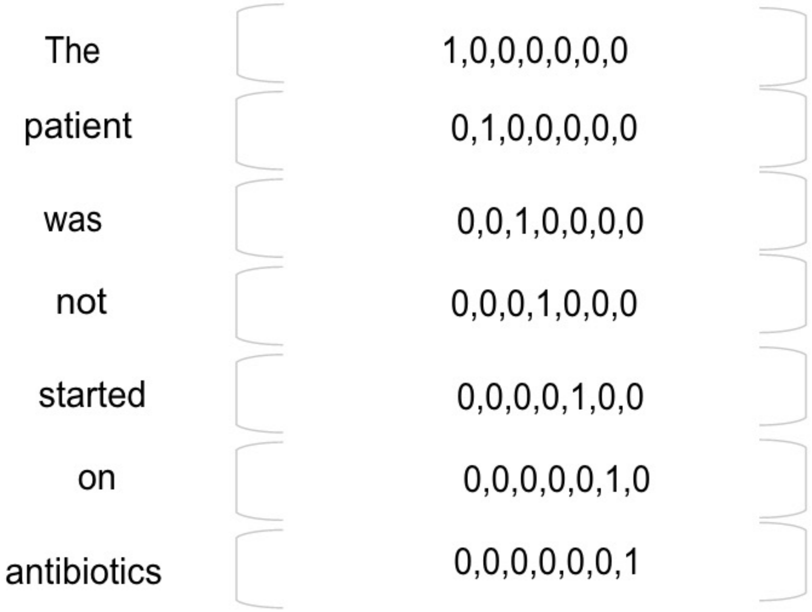


Fig. 5. One-hot Encoding

One hot encoding is useful for data that have no relationship to each other. Machine learning algorithms treat the order of numbers as an attribute of significance. In other words, they will read a higher number as better or more important than a lower number.

3.3.2 Bag-of-words embedding

A bag-of-words model (BoW) is a way of extracting features from the text for use in modelling, such as with machine learning algorithms, which is a popular way of representing documents. Throughout the realm of information retrieval, the bow model is used in information retrieval on the assumption that a text may be reduced to a collection of words without regard to its actual order within the document, its grammar, or its syntax. Each word's presence in the document is autonomous and unrelated to the incidence of other terms. In other words, each word at each place in the document is individually selected without regard to the document's semantics.
For given two sentences,

1. She is admitted from hospital for a coronary artery bypass surgery.

2. She is a female admitted to outside hospital with chest discomfort.

Based on these two sentences, a dictionary which contains 16 words is built: Dictionary = 1:”She”, 2. ”is”, 3. ”admitted”, 4. ”from”, 5. ”hospital”, 6. ”for”, 7. ”a”, 8. ”coronary”, 9. ”bypass”, 10. ”surgery”, 11. ”female”, 12. ”to”, 13. ”outside”, 14. ”with”, 15. ”chest”, 16. ”discomfort” Then the two sentences can be represented as:

1.

$$1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0$$

2.

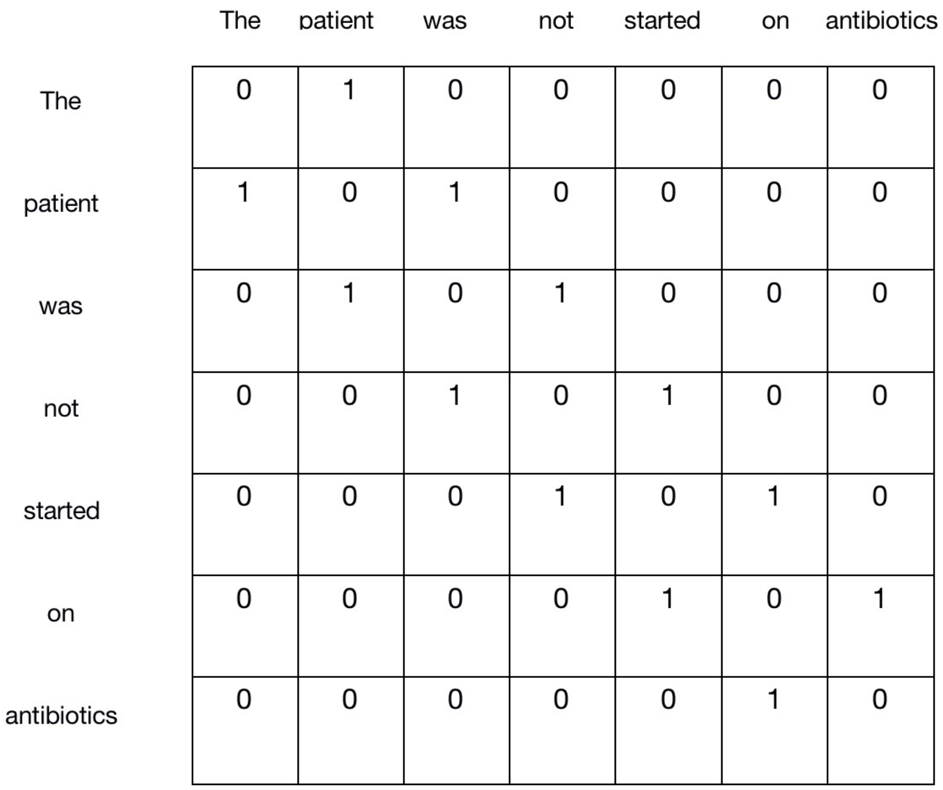
$$1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1$$

Each element in the vector represents the number of times related words in the dictionary appear in the document. However, in the process of constructing document vectors, we can see that we do not express the order in which words appear in the original sentence.

3.3.3 GloVe Embedding

GloVe stands for “Global Vectors” (Pennington et al., 2014). Compared to Word2vec, which only keeps local statistic information, the advantage of GloVe is that it captures both global statistics and local statistics of a corpus, and it is an unsupervised learning pre-trained model that represents words in sentences. It has the ability to extract semantic relationships.

The main idea of the Glove method is that ”semantic relationships between words can be extracted from the co-occurrence matrix (Bullinaria and Levy, 2007)”. Given a document having a sentence of **n** words, the co-occurrence matrix **X** will be a **n*n** matrix (as follows) where X_{ij} indicates how many times word **i** has co-occurred with word **j**.



	The	patient	was	not	started	on	antibiotics
The	0	1	0	0	0	0	0
patient	1	0	1	0	0	0	0
was	0	1	0	1	0	0	0
not	0	0	1	0	1	0	0
started	0	0	0	1	0	1	0
on	0	0	0	0	1	0	1
antibiotics	0	0	0	0	0	1	0

Fig. 6. Co-occurrence Matrix with a Window Size of 1

Let P_{ij} represent the likelihood of seeing the words I and j together, which is calculated by dividing the number of times I and j occurred together (X_{ij}) by the total number of times word I appeared in the corpus (X_i), where $P_{ij} = X_{ij}/X_i$. Suppose there is a function F that takes in word vectors of I j, and k and returns the desired ratio.
$F(w_i, w_j, u_k) = P_i k/P_j k$

3.4 Modeling Introduction

3.4.1 BiLSTM

It is very similar to the structure of the traditional recurrent neural networks and multi-layer feedforward neural networks after expansion, so the problem of gradient disappearance will inevitably occur if there are too many layers during training (Hochreiter, 1998). The schematic diagram of gradient disappearance is shown in Figure 6. The node colour depth in the figure indicates how much input information at the first moment can be retained at the current moment. The darker the colour is, the more information is retained, and the better the model effect is; the lighter the colour is, the less information is retained, and the model is prone to gradient disappearance. From the feedforward process of the model, as time goes by, the information that can be extracted at subsequent moments gradually decreases. As shown in the figure, the information at time 1 that can be obtained when processing the data at time 7 has almost disappeared. From the perspective of the back propagation process in the training process, when the error of the output layer at time 7 propagates forward through the gradient, the model cannot effectively update the corresponding weights at the previous time due to the decrease of the gradient. The problem that the model cannot get the previous moment information due to the vanishing gradient is also called the long-term dependence problem. In order to solve this problem, many researchers at home and abroad have improved the RNN model, including BRNN (Arisoy et al., 2015), GRU (Tang et al., 2016), LSTM, etc., among which the LSTM model is the most widely used.

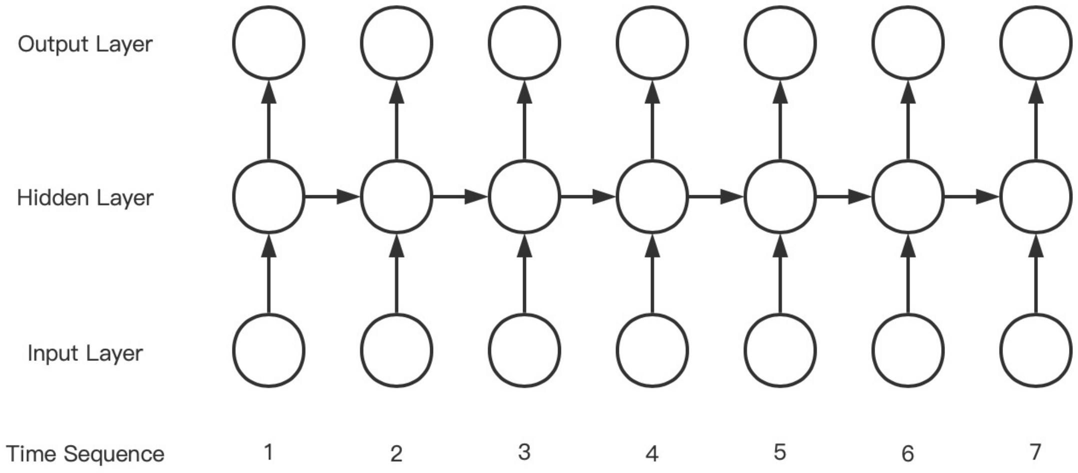


Fig. 7. The Vanishing Gradient Problem in Deep Learning (Hochreiter, 1998)

LSTM is an improved recurrent neural network structure which can effectively avoid the long-term dependence problem that the traditional recurrent neural network cannot solve. LSTM network was first proposed by Hochreiter in 1997. In the later work, many people adjusted and popularised the model, and the modern, widely used LSTM structure came into being. LSTM network has the same chain structure as traditional recurrent neural networks, but the repeating module has a different design. The recurrent neural network only has a single layer of repeated modules. In contrast, the LSTM network has four interacting neural network layers, which makes the LSTM model more complex and sacrifices a certain amount of time, but the model effect becomes better. The standard LSTM model is shown in Figure 8, which mainly includes three groups of adaptive element multiplication gates, namely, the input gate, forgetting gate and output gate.

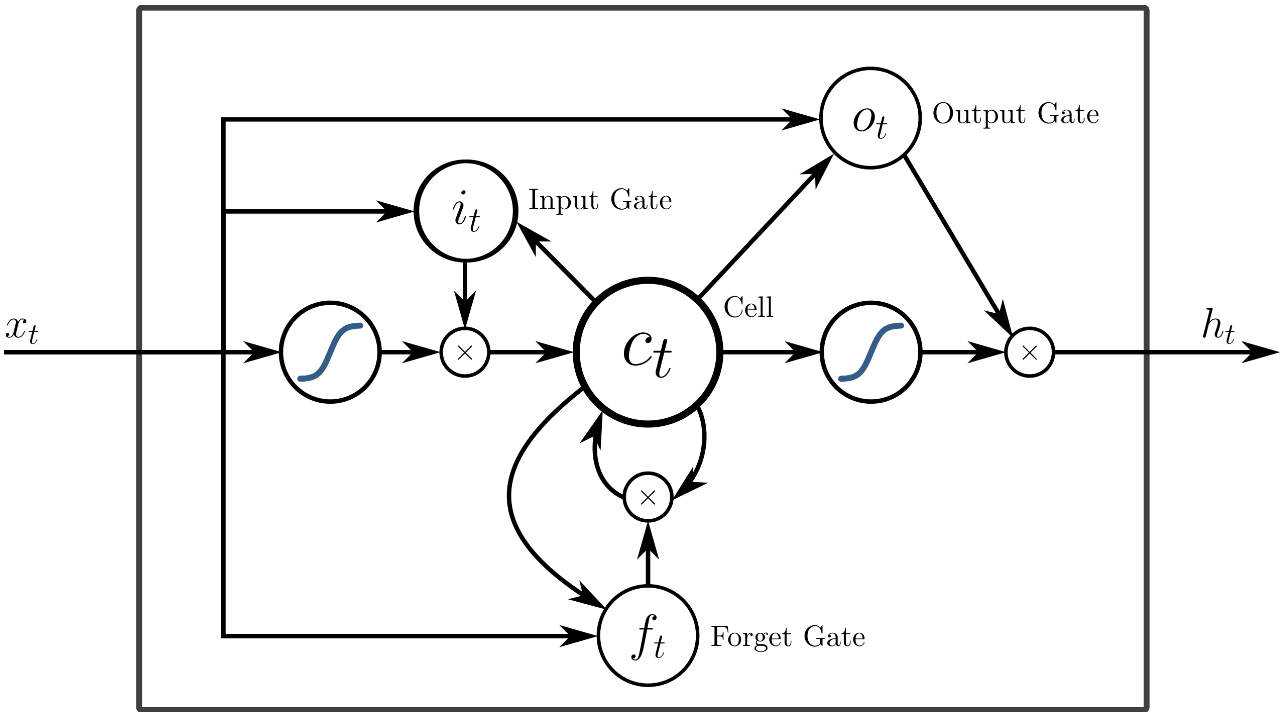


Fig. 8. Standard LSTM Structure (Wikipedia)

In actual text processing applications, the text is characterized by sequential feature correlation, and words in the current moment are not only affected by words in the past moment but also by words in the future moment. In order to make the model contain contextual information at the same time, some scholars put forward a bidirectional long short-term memory model (Zhang et al., 2015; Zhou et al., 2016). BiLSTM contains two LSTM layers, which train the forward and backward sequences, respectively, and both LSTM layers are connected to the output layer. This forward-backward two-layer LSTM structure can provide historical information and future information to the output layer at the same time, which can not only retain the advantages of LSTM to solve long-term dependence but also take into account the context information and effectively deal with sequence problems.

To illustrate the structure, the BiLSTM model is presented in the figure below:

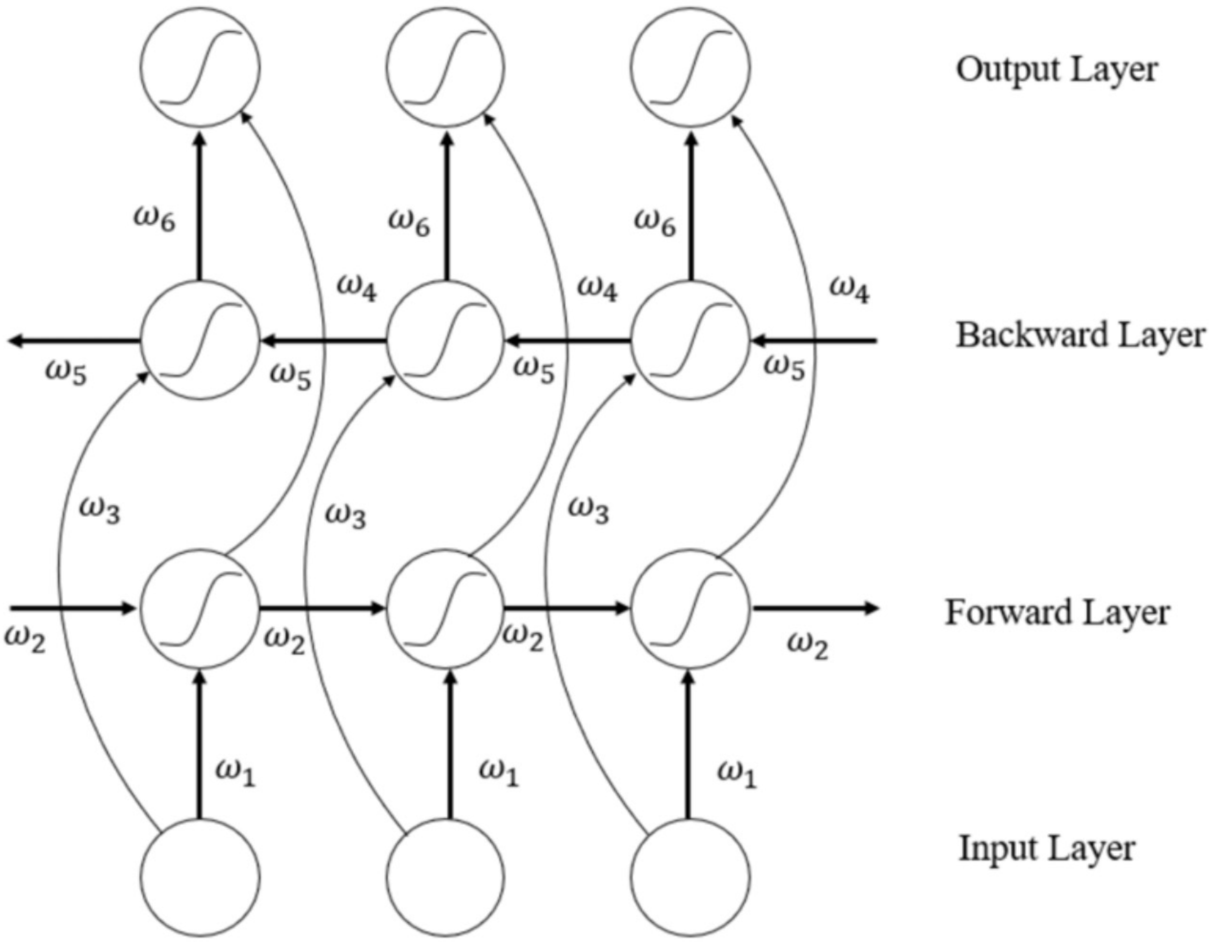


Fig. 9. BiLSTM Structure (Gan et al., 2021)

3.4.2 CRF model

Conditional random fields (CRFs) (Lafferty et al., 2001) are a type of statistical modelling methodology frequently employed in pattern recognition and machine learning, as well as for structured prediction, which is a widely-applied modelling strategy for many NLP tasks. Data having an underlying graph structure can be better modelled using CRFs, a graphical model that can take advantage of structural dependencies between outputs. In contrast to classifiers, which make label predictions for isolated data without considering their ”neighbouring” samples, CRFs are able to account for context. In order to accomplish this goal, the predictions are modelled in the form of a graphical model, which depicts the existence of relationships between the predictions. The choice of the graph to implement is context-specific. In natural language processing, for instance, ”linear chain” CRFs are prevalent because each prediction depends only on its immediate neighbours. Connecting neighbouring and/or analogous regions ensures that all regions within a picture are predicted in the same way by the graph. As an undirected graph model conforming to Markov random field, the conditional distribution $P(Y|X, \lambda)$ of target tag sequence y is calculated based on observation sequence X. Given a clinical text sequence $X =< x_1, x_2, ..., x_n >$, CRFs model aims to make tags on

X with a sequence $X = < y_1, y_2, ..., y_n >$ to get this outcome $t_i \in \{B, I, E, O\}$, where B I E O represent the start, the end, the inner and the outside of a TIMEx corresponding to x_i. Given under the condition of random variable X, if the conditional probability distribution of random variable Y $P(Y|X, \lambda)$ that constitutes a conditional random field, which meets the Markov property(1).

$$P(Y_i|X, Y_1, Y_2, ..., Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1}) \tag{1}$$

The conditional probability can be calculated by the following formula (2):

$$P(Y|X) = \frac{1}{Z(X)} e^{\sum_{i=1}^{I} \sum_{k=1}^{K}} \lambda_k f_k(i, Y_i, Y_{i-1}, X) \tag{2}$$

where Z(X)is the normalization factor(3):

$$Z(X) = \sum_{Y} e^{\sum_{i=1}^{I} \sum_{k=1}^{K}} \lambda_k f_k(i, Y_i, Y_{i-1}, X) \tag{3}$$

Given the location of the input sequence X and i, where $f_k(i, Y_i, Y_{i-1}, X)$ represents the first k eigenvalue values of mark Y_i and mark Y_{i-1} where λ_k represents the feature weights. From the formula, conditional random field model to the algorithm to calculate the sequence through the forward - after the characteristics of different position expectation and conditional probability, then using maximum likelihood estimation of quasi-newton method to find the solution of the model parameters, such as, the use of dynamic programming in the viterbi algorithm for decoding, and the sequence of the output data for testing.

From the formula, conditional random field model to the algorithm to calculate the sequence through the forward - after the characteristics of different position expectation and conditional probability, then using maximum likelihood estimation of quasi-newton method to find the solution of the model parameters, such as, the use of dynamic programming in the viterbi algorithm for decoding, and the sequence of the output data for testing as show in the following picture (figure 10). Additionally, there are other possible NLP systems that can be used to extract various features including (1) MedLEE, (2) KnowledgeMap, (3) a Dictionary-based Semantic Tagger (DST).

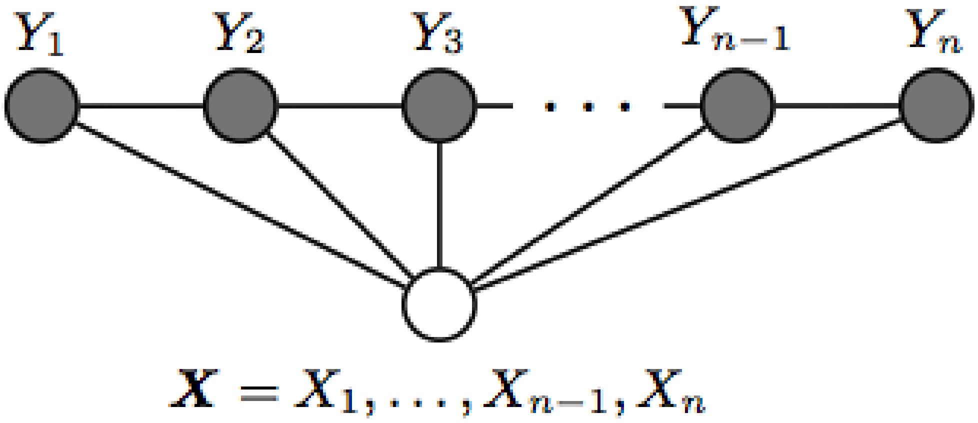


Fig. 10. CRF Structure (Lafferty et al., 2001)

Suppose our NER task uses a BiLSTM model with some word embeddings (illustrated in the graph below). The BiLSTM provides the logits per class for each word, but each prediction for a single word is independent of the predictions for its neighbours. Each token is independently classified without the model knowing the predicted classes of its neighbours.

3.4.3 CNN model

In deep learning, a classic convolutional neural network (CNN) is a class of artificial neural networks (ANN), most commonly applied to analyse visual imagery. CNN mainly consists of these layers: input layer, convolution layer, relu layer, pooling layer and full connection layer (structure shown in figure10).

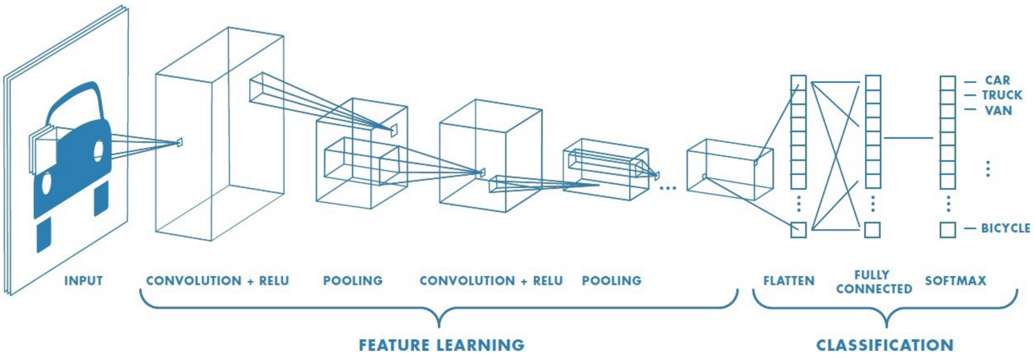


Fig. 11. Convoluted Neural Networks Topology (Saha)

One of CNN's most alluring qualities is its ability to make use of spatial or temporal correlation in data. CNN is divided into multiple learning phases, each of which is made up of a combination of convolutional layers, nonlinear processing units, and subsampling layers. In a CNN, each layer of the network uses a set of convolutional kernels to perform many transformations. The convolution process simplifies extracting useful features from spatially correlated data points.

To the best of my knowledge, CNN performs admirably on NLP-related tasks like semantic parsing, NER, and Relation Extraction(RE), where deep learning has been efficiently and widely used due to the prevailing interests of academics in this area. ConvNets' architecture mimics the human brain's neuronal connectivity pattern and takes design cues from the Visual Cortex's structure. Only some areas of the visual field, referred to as the receptive field, are capable of eliciting responses from individual neurons when they are stimulated. A combination of these fields overlaps to encompass the entire visual field. CNN is able to receive an image as input, assign importance (learnable weights and biases) to distinct aspects/objects in the picture, and differentiate between them. Comparatively, ConvNet requires substantially less pre-processing than other classification techniques. In contrast to primitive approaches, ConvNets are capable of learning these filters/characteristics given sufficient training. This concept works just as well in words and characters as it does in practice.

3.4.4 BERT model

Brief for ”Bidirectional Encoder Representations from Transformers,” BERT (Devlin et al., 2018) is founded on Transformers, a cutting-edge deep learning model in which all output elements are related to all input elements and the weights between them are dynamically computed based on their connection. BERT is unique in that it can be used with text that is read either from left to right or right to left. With this skill, BERT is taught for two NLP tasks: Next Sentence Prediction and Masked Language Modeling.

The input for BERT is comprised of Token Embeddings, Segment Embeddings, and Position Embeddings, followed by pre-training, as indicated in the figureFig. 12. The final output of the [CLS] parameter of the BERT is set as C, the input embedding is designated as E, and the final output of the ith token is designated as T. This will enable effective dynamic feature encoding of polysemous words. In other words, the same word might produce a variety of word vector outputs depending on the language environment. Additionally, the pre-trained model's embedding layer transfers the knowledge from the corpus to make the word vectors more generic, greatly enhancing the model's accuracy.

Instead of using word embedding, the conventional method of language processing, BERT assigns each word to a vector that only conveys a single dimension of its meaning. Extensive quantities of labeled data are required for word embedding models. However, because all words are in some sense bound to a vector or meaning, they have difficulty with the context-heavy, predictive nature of question responding.In order to prevent the target word from ”seeing itself” (i.e., developing a meaning apart from its context), BERT employs a technique called masked language modelling. The advent of pre-trained language models in recent years has marked a significant milestone in developing NLP. Pre-trained language model methods from the past focused on acquiring high-quality word embeddings. Word vectors are context-free, notwithstanding their ability to capture a word's semantics. Since this is the case, individuals are unable to understand the deeper meanings included in the text, such as its syntax and semantics.

As discussed in Chapter 2, for downstream tasks modelled by BERT, especially NLP tasks that include sentence text, firstly, the sentences need to be encoded independently and pre-trained through BERT's self-attention mechanism. It is important to note that in the text classification task, we will need to test the sentence and the label L together and form the text - tag on; this text classification task can be directly into the sentence to the task. Finally, fine-tune the BERT model to find the most suitable parameters for the corresponding task.

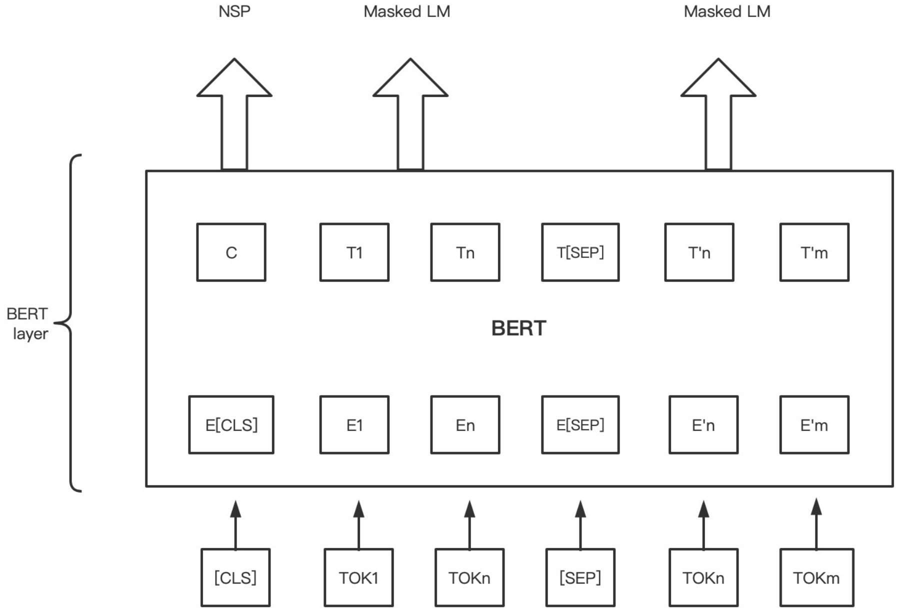


Fig. 12. BERT Pre-trained Process (Shi et al., 2020)

3.5 Time Expressions and Events Named Entity Recognition

The goal of the annotation is to mark up temporal information present in clinical text in order to enable reasoning and relevant events (medications) for each patient. Time expressions identification, medical events identification and time relations identification were performed on 600 clinical notes, and pathological texts from cancer patients (Bethard et al.). Conditional Random Fields (CRFs) have been found to be an effective method to detect Time Expressions and events (Lin et al., 2013), based on which, I plan to act CRFs-based method as the recogniser to extract TIMEX (e.g. 'This February', '24th March, 2020') and EVENT (e.g. 'Hospitalization from...', 'chemotherapy is indicated from... '). CRFs model is a conditional probability model proposed in 2001 (Lafferty et al., 2001). It combines the characteristics of the hidden Markov model (HMM) (Eddy, 2004) and Maximum Entropy Markov model (MEMM) (McCallum et al., 2000), avoids the problem of label bias through global normalisation, and achieves good results in the task of NER.

According to the results discussed in Section 2, statistical model-based techniques, including machine learning and deep learning, contribute to better performance than rule-based techniques.

Table 3. Tags Used in the i2b2 2009 Dataset.

Tag	Meaning	Example
m	medication	Percocet
do	dosage	3.2mg
f	frequency	twice a day
mo	mode (/route of administration)	Mode: “nm” (not “Tablet”)
du	duration	10-day course
r	reasons	Dizziness

Table 4. Tags Used in the i2b2 2012 Dataset.

Tag	Meaning	Example
CLINICAL_DEPT	clinical department	emergency room
EVIDENTIAL	events that have an ‘evidential’ nature	CT **shows**
TEST	clinical tests	CT
PROBLEM	symptoms	sickness
TREATMENT	medications, surgeries and other procedures	Levaquin
OCCURRENCE	the default value for other event types	He was **readmitted** for

In this dissertation, an Inside–outside–beginning (IOB) format, a commonly used format in entity tagging (Ramshaw and Marcus, 1999), is used, which is explained as follows:

1. the B-prefix indicates that the tag is at the beginning of a chunk that follows another chunk without O tags between the two chunks

2. I-prefix indicates that the tag is inside a chunk

3. the O-prefix indicates that the token belongs to no chunk

The entity tags used in the n2c2 challenge track 2009 and 2012 have difference which are shown as follows (Table 3 and Table 4):

In this dissertation, a BiLSTM+CNN model and a BiLSTM+CRF model were deployed to achieve the goal and a comparison of performance was taken to demonstrate the difference, advantages and disadvantages between these two models.

3.6 Temporal Relations and Medications Candidates Classification

Relation extraction is the task of predicting attributes and relations for entities in a sentence. Extracted relationships usually occur between two or more entities of a particular type (e.g. Medication, Dosage) and fall into several semantic categories. In this dissertation, these relations that are discussed are between medications and corresponding dates (e.g. ”He has been treated with simvastatin since 12/04/1998”, which contains the relation ”simvastatin” happened after

”12/04/1998”). Generally, there are eight types of temporal relations, including before, after, simultaneous, overlap, begun_by, ended_by, during, and before_overlap (Sun et al., 2012). However, only before, after, and overlap meet my aim.

According to the tasks in the n2c2 Challenge, medical EVENTs classification requires that Classify medication mentioned in clinical notes as either: (1) Disposition (medication change discussed), (2) NoDisposition (no change discussed), or (3) Undetermined (need more information). For temporal relation, the classification of the contextual information for Disposition events categorises along at least 2 orthogonal dimensions: Action (e.g. start, end), Temporality (e.g. past, present, future).

Given the possible vague description of events, I plan to include contextual characteristics, POS-tagging features and location features of EVENTs and TIMEX as the input to ensure the classifier performs well as classifier models. Furthermore, in order to achieve better performance in the Precision, Recall and F-score, a BERT-based BiLSTM model was applied. To explore the influence of different word embedding on classification models, BERT, one-hot and Glove embedding were applied separately.

3.7 Post-processing

After finishing the former four steps, some results might be incorrect. To make this system more robust on a wide span of texts, additional rules could be applied. The following process is to convert results to a structured shape table or CSV shape (table 3).

However, in a previous 2012 i2b2 NLP challenge on temporal relation extraction, post-processing was considered a set of classification routines for the EVENTs, and TIMEx remained unclassified (Lin et al., 2013). In this dissertation, I applied an open resource tool, SparkNLP, combined with my medication entity recognition system and temporal relation extraction system to perfect the results and generate a structured table (Table 5).

ID	Event	Statues	Start	Stop
134529565	Methotrexate	ON	May 2019	February 2020
134529566	Methotrexate	OFF	February 2020	Unknown

Table 5. Extraction Result Example

3.8 Software and Environment

Considering the computation resources I possess, in order to meet with the computational ability of the models’ needs, the hardware and software used will be discussed in this section. Given the complex progress of this dissertation, I have done the data pre-processing with my personal computer (configurations shown in table 6).

Table 6. NER Evaluation on the 2009 Dataset.

Hardware	Test Environment
Central Processing Unit (CPU)	Intel Core i5 1.6GHz
Random Access Memory (RAM)	16GB 2133MHz DDR3
Graphics Processing Unit (GPU)	Intel UHD Graphics 617 1536 MB
Disk Memory	256GB

This dissertation modelling is a relatively large amount of computing resources consuming process, it will take more than a week if I train the model on my personal computer, in which case, a google virtual machine platform (Google Co-Lab) and Microsoft Azure Machine Learning Studio are used to train and test models. All the codes were run in Ubuntu 18.04, macOS Monterey 12.1, Python version 3.8 and Tensorflow version 2.1.4.

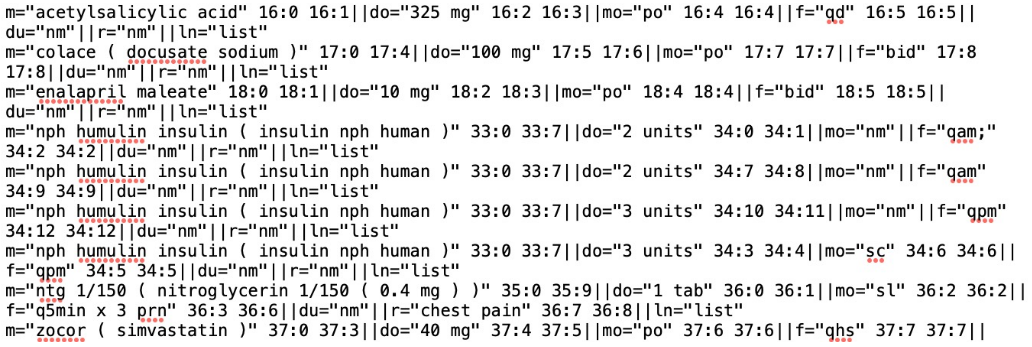
3.9 Chapter Summary

This chapter gives the methodology of this dissertation; I discuss the primary tasks of named entity recognition, including data pre-processing (sentence segmentation, tokenisation, POS tagging and Syntactic parsing) and word embedding (including word vector training and language model). The advantages and disadvantages of different models. Moreover, the definition and explanation of time expressions and temporal relation extraction.

4 Experimental Evaluations

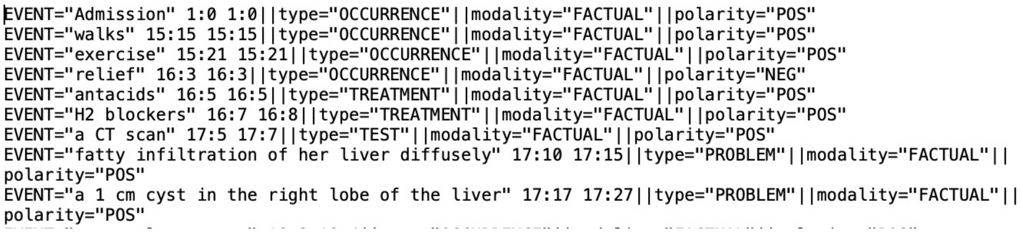
4.1 Data Pre-process and Statistic analysis

The n2c2 dataset (both 2009 and 2012) is composed of clinical notes that have been de-identified. Their annotated format is confusing (Figure 13 and Figure 14):



```
m="acetylsalicylic acid" 16:0 16:1||do="325 mg" 16:2 16:3||mo="po" 16:4 16:4||f="qd" 16:5 16:5||
du="nm"||r="nm"||ln="list"
m="colace ( docusate sodium )" 17:0 17:4||do="100 mg" 17:5 17:6||mo="po" 17:7 17:7||f="bid" 17:8
17:8||du="nm"||r="nm"||ln="list"
m="enalapril maleate" 18:0 18:1||do="10 mg" 18:2 18:3||mo="po" 18:4 18:4||f="bid" 18:5 18:5||
du="nm"||r="nm"||ln="list"
m="nph humulin insulin ( insulin nph human )" 33:0 33:7||do="2 units" 34:0 34:1||mo="nm"||f="qam;"
34:2 34:2||du="nm"||r="nm"||ln="list"
m="nph humulin insulin ( insulin nph human )" 33:0 33:7||do="2 units" 34:7 34:8||mo="nm"||f="qam"
34:9 34:9||du="nm"||r="nm"||ln="list"
m="nph humulin insulin ( insulin nph human )" 33:0 33:7||do="3 units" 34:10 34:11||mo="nm"||f="qpm"
34:12 34:12||du="nm"||r="nm"||ln="list"
m="nph humulin insulin ( insulin nph human )" 33:0 33:7||do="3 units" 34:3 34:4||mo="sc" 34:6 34:6||
f="qpm" 34:5 34:5||du="nm"||r="nm"||ln="list"
m="ntg 1/150 ( nitroglycerin 1/150 ( 0.4 mg ) )" 35:0 35:9||do="1 tab" 36:0 36:1||mo="sl" 36:2 36:2||
f="q5min x 3 prn" 36:3 36:6||du="nm"||r="chest pain" 36:7 36:8||ln="list"
m="zocor ( simvastatin )" 37:0 37:3||do="40 mg" 37:4 37:5||mo="po" 37:6 37:6||f="qhs" 37:7 37:7||
```

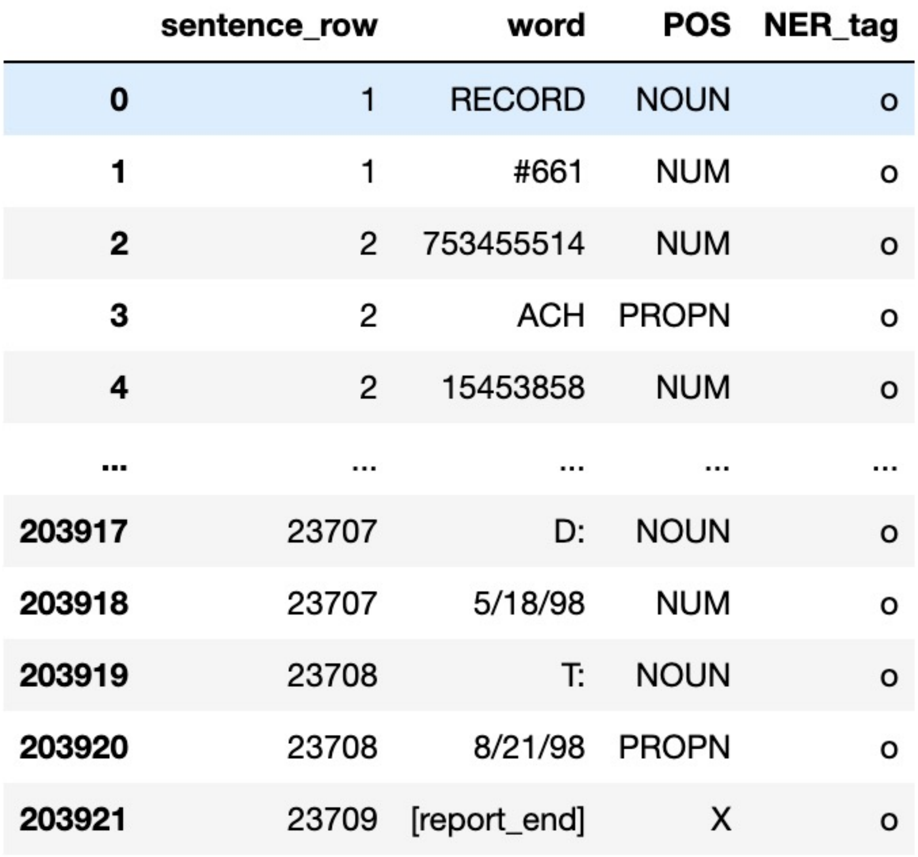
Fig. 13. i2b2 2009 Dataset Example



```
EVENT="Admission" 1:0 1:0||type="OCCURRENCE"||modality="FACTUAL"||polarity="POS"
EVENT="walks" 15:15 15:15||type="OCCURRENCE"||modality="FACTUAL"||polarity="POS"
EVENT="exercise" 15:21 15:21||type="OCCURRENCE"||modality="FACTUAL"||polarity="POS"
EVENT="relief" 16:3 16:3||type="OCCURRENCE"||modality="FACTUAL"||polarity="NEG"
EVENT="antacids" 16:5 16:5||type="TREATMENT"||modality="FACTUAL"||polarity="POS"
EVENT="H2 blockers" 16:7 16:8||type="TREATMENT"||modality="FACTUAL"||polarity="POS"
EVENT="a CT scan" 17:5 17:7||type="TEST"||modality="FACTUAL"||polarity="POS"
EVENT="fatty infiltration of her liver diffusely" 17:10 17:15||type="PROBLEM"||modality="FACTUAL"||
polarity="POS"
EVENT="a 1 cm cyst in the right lobe of the liver" 17:17 17:27||type="PROBLEM"||modality="FACTUAL"||
polarity="POS"
```

Fig. 14. i2b2 2012 Dataset Example

It could be not very pleasant if the chosen model is expecting a different format of training data. A widely used training dataset format is CoNLL (Nivre et al., 2007) for NLP NER tasks. In this data presentation, each token (in this case, an individual word or punctuation mark) sits on its own line. The entire tokenised document is presented in that first column. The entity tag is in the last column. A custom parser is required to transform the data from n2c2's entity-only, offset-based annotation format into CoNLL's all-token, table-based format. I proposed this parser to convert the data from the n2c2 2009 dataset into CoNLL's format, which data structure consists of sentence row number, separate words (tokens), POS tag and label (as shown in Figure 15). However, there could be slight differences among different tracks data; this parser could be used on a different dataset.



	sentence_row	word	POS	NER_tag
0	1	RECORD	NOUN	o
1	1	#661	NUM	o
2	2	753455514	NUM	o
3	2	ACH	PROPN	o
4	2	15453858	NUM	o
...	...	...	...	...
203917	23707	D:	NOUN	o
203918	23707	5/18/98	NUM	o
203919	23708	T:	NOUN	o
203920	23708	8/21/98	PROPN	o
203921	23709	[report_end]	X	o

Fig. 15. CoNLL Format

To better understand the dataset, some statistical analysis was taken. As shown in Figure 16, with BIO tagging format, the i2b2 challenge 2009 track dataset labels consist of 13 categories: B-m, I-m, B-do, I-do, B-f, I-f, B-r, I-r, B-mo, I-mo, B-du, I-du and B-m is in the majority of this dataset. A re-sample was conducted to erase the imbalance of the data sample (cutting labels more than 3000 down to 3000).

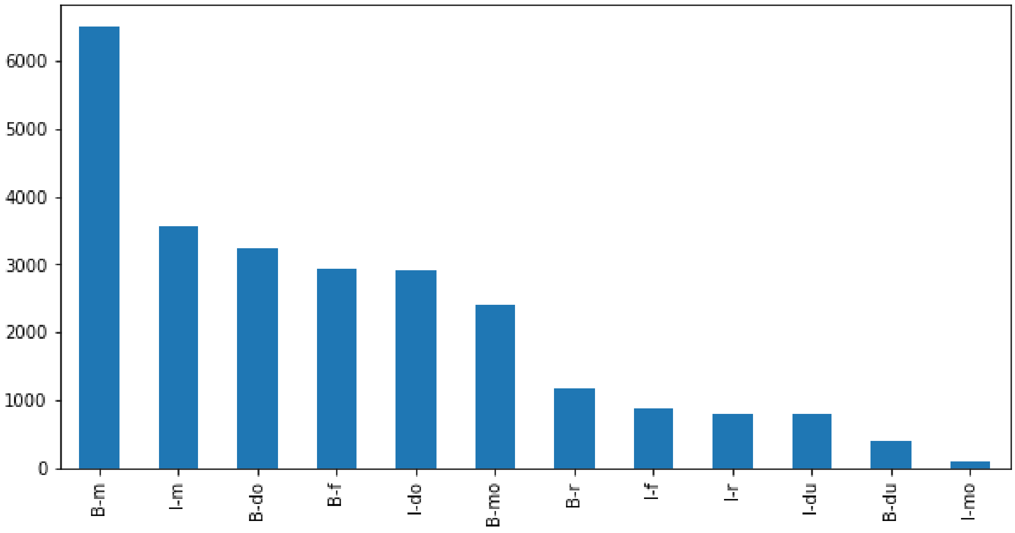


Fig. 16. 2009 Dataset Labels Statistical Analysis

Then, through the statistical analysis of the sentence length of the input text data, the expected value of the normal distribution of the sentence length (it can be seen in Figure 17 that the majority of the sentence length is 10-20 words) is taken as the target length of the sentence input in the subsequent model.

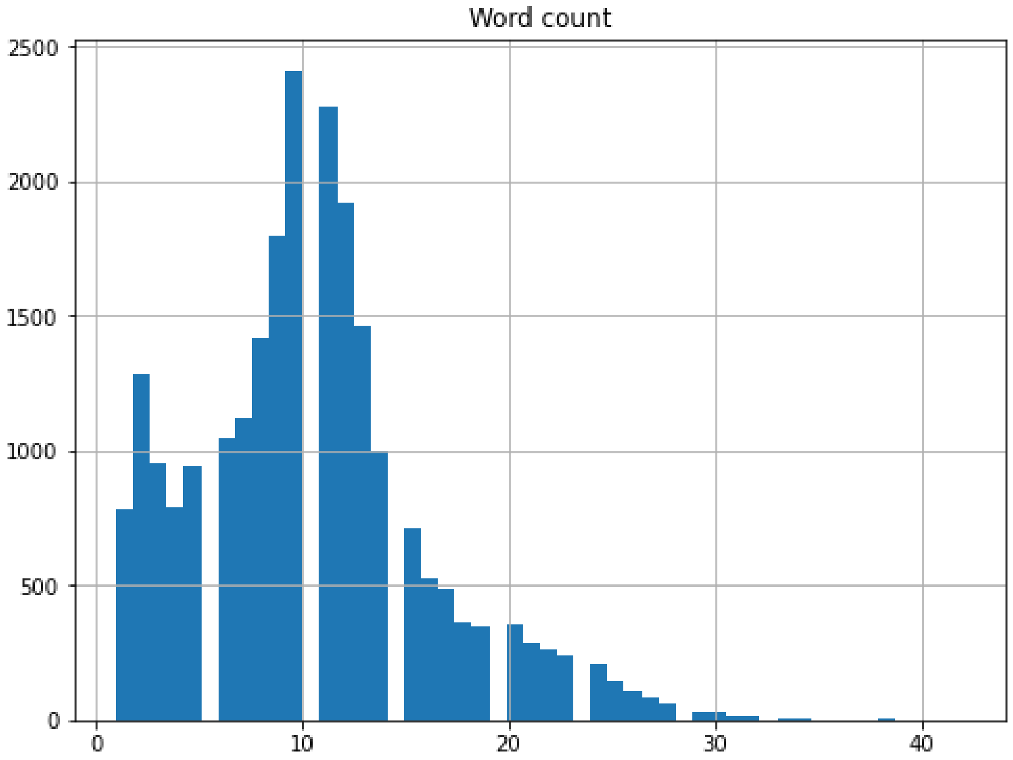


Fig. 17. 2009 Word Count Statistical Analysis

Similarly, the same work was conducted on the n2c2 challenge 2012 dataset. Most sentences are between 5 and 25 words in length with wider range distribution than the 2009 dataset, as seen in Figure 18.

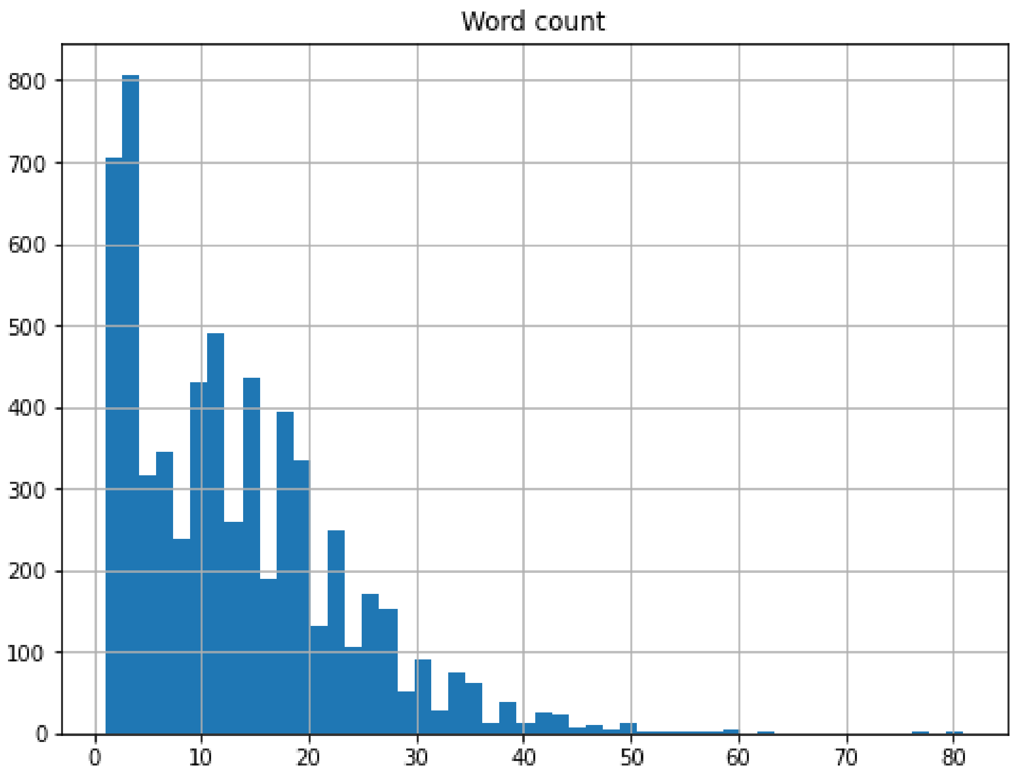


Fig. 18. 2012 Word Count Statistical Analysis

4.2 Model-I: BiLSTM-CRF for NER task

Each classification decision is conditionally independent when a simple BiLSTM is followed by a classifier. The linear-chain CRF explicitly describes dependencies between labels as a table, including transition scores between all possible label pairs.

If the labels adhere to a tight internal syntax, the CRF can acquire this information very quickly. There are numerous methods for encoding the output of NER. Still, they typically encode at least the Beginning, Inside, and Outside of an entity, which must be in a syntactically determined order. CRF will recognise very soon that it is impossible for I-m to follow O, as it must always follow B-m.

As they are conditionally independent, BiLSTM may be uncertain as to whether it should place B-m on a drug or one dosage unit later and, therefore, output both. CRF layer that is aware that this is unlikely and enforces the tags' intrinsic logic would output B-do, I-do.

Based on the discussion in Chapter2, Combined with the above two models, this dissertation conducted a method of electronic disease named entity recognition based on BiLSTM-CRF. The specific work steps are as follows:

1. Apply Glove or bag-of-word model to obtain words semantic representation.
2. After obtaining the semantic representation, the sequence information is captured by the BiLSTM network, which effectively avoids the gradient disappearance and explosion of other deep learning models.

3. Restrict the sequence relationship between tags by using conditional random fields.

4. The experimental results show that the model can effectively improve the accuracy of named entity recognition of electronic medical record.

LSTM is able to add or remove information to the state unit, which is controlled by structures called gates. Followed by a softmax layer, although BiLSTM learns adequate context information in the model training stage, the output probabilities of softmax are independent of each other, and there is no mutual dependence between the output values. Softmax only selects the label corresponding to the current optimal solution as the output at each step, which cannot avoid the output of wrong labels. For example, arguably, I-m always occurs following B-m instead of B-r. However, with softmax, this type of apparent mis-classification happens a lot. To solve this problem, following BiLSTM, CRF as the output layer, take account into the order information between labels, so as to obtain more accurate labels through probability calculation.

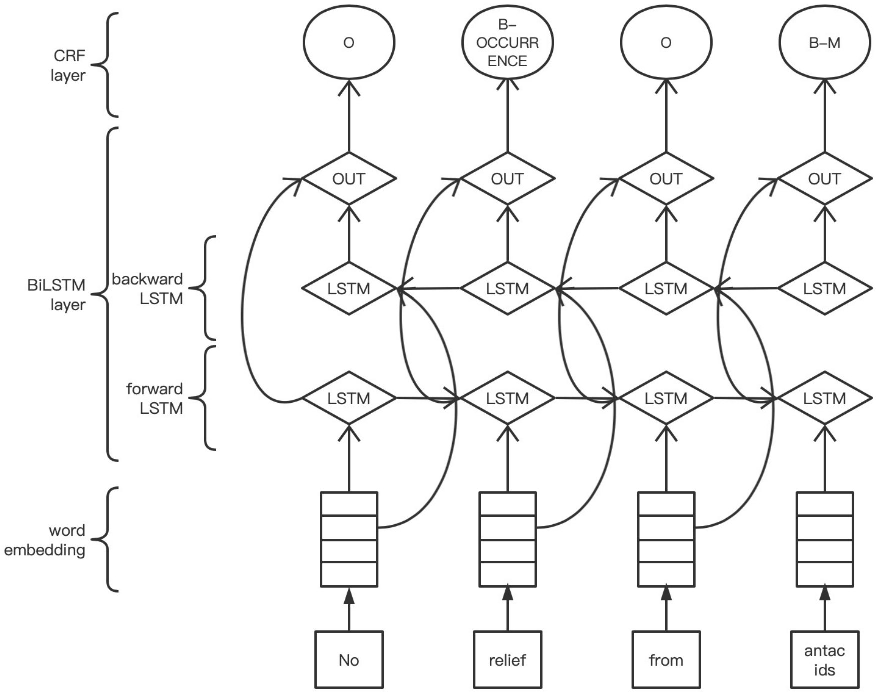
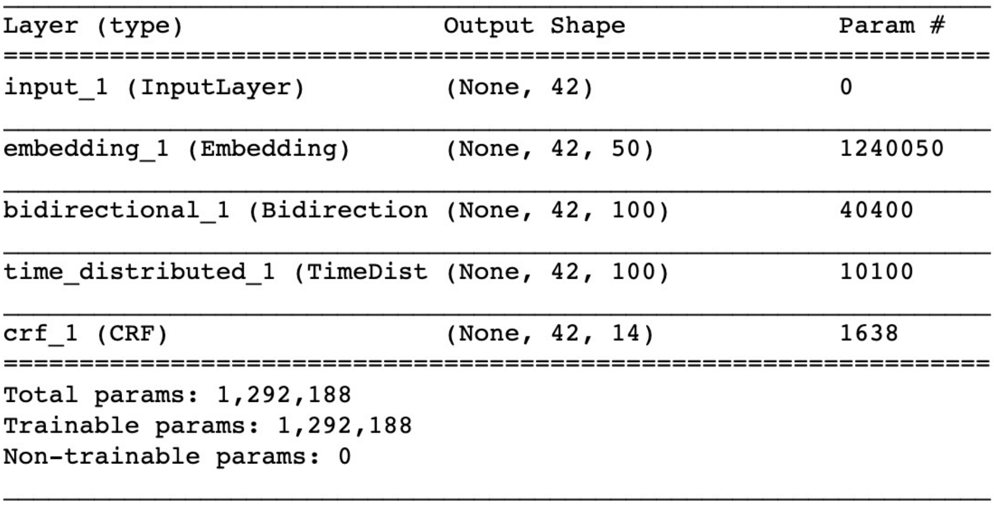


Fig. 19. Model Structure of BiLSTM+CRF

CRF directly models the conditional probability distribution $P(Y|X)$ combining the advantages of graph models to predict multivariate outputs y with a large number of input features x. CRF is the clique potential function φ and c of all input features are a variant of conditional Markov random fields. The clique potential function is usually assumed to be log-linear. The linear chain CRF built on BI-LSTM can effectively model several hard constraints containing the dependencies of output labels.

4.2.1 Model-I Evaluation

The structure of the model for the i2b2 challenge 2009 dataset NER task is shown in Figure 20. Several groups of control experiments are carried out under this model. The main parameters of the final model are shown in Table 7, and the optimal performance on the test set is obtained after 4 epoch training.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 42)                0
_________________________________________________________________
embedding_1 (Embedding)      (None, 42, 50)            1240050
_________________________________________________________________
bidirectional_1 (Bidirection (None, 42, 100)           40400
_________________________________________________________________
time_distributed_1 (TimeDist (None, 42, 100)           10100
_________________________________________________________________
crf_1 (CRF)                  (None, 42, 14)            1638
=================================================================
Total params: 1,292,188
Trainable params: 1,292,188
Non-trainable params: 0
_________________________________________________________________
```

Fig. 20. Network of BiLSTM Model on the 2009 Dataset

parameter	value
DENSE_EMBEDDING	50
LSTM_UNITS	50
LSTM_DROPOUT	0.2
DENSE_UNITS	100
BATCH_SIZE	256
MAX_EPOCHS	10
LEARNING_RATE	0.001

Table 7. Model Parameters of BiLSTM+CRF for the 2009 Dataset

With 70% of training dataset, 15% validation dataset and 15% test dataset, the training process for NER task on the i2b2 2009 dataset is shown below:

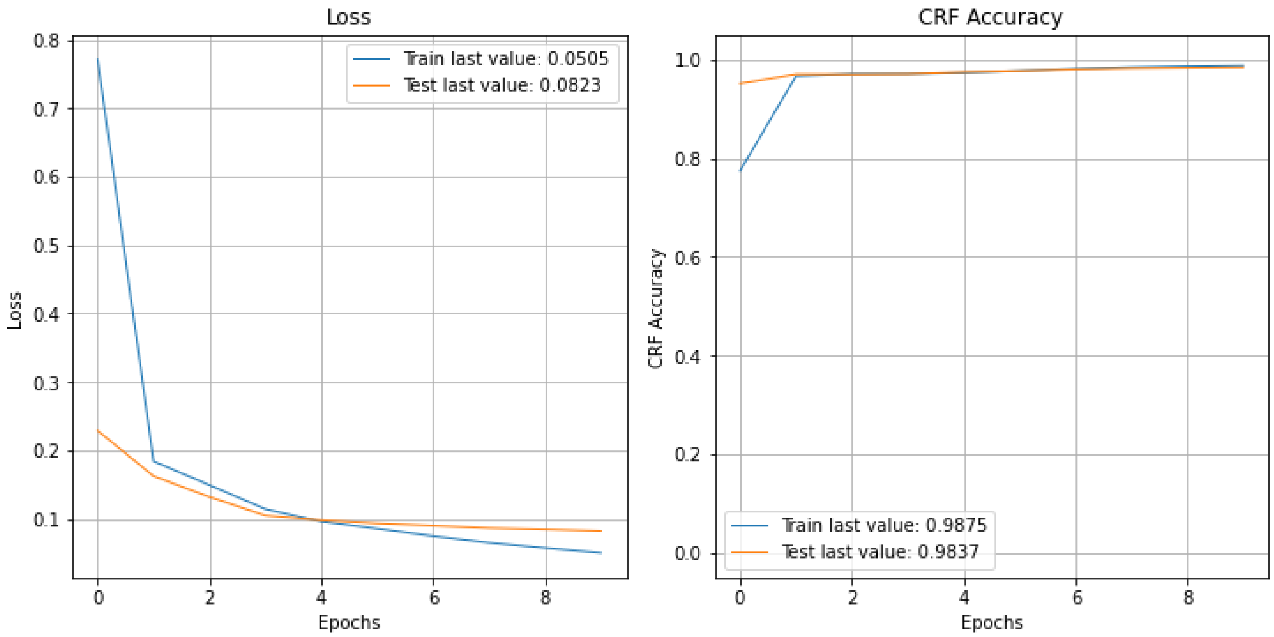
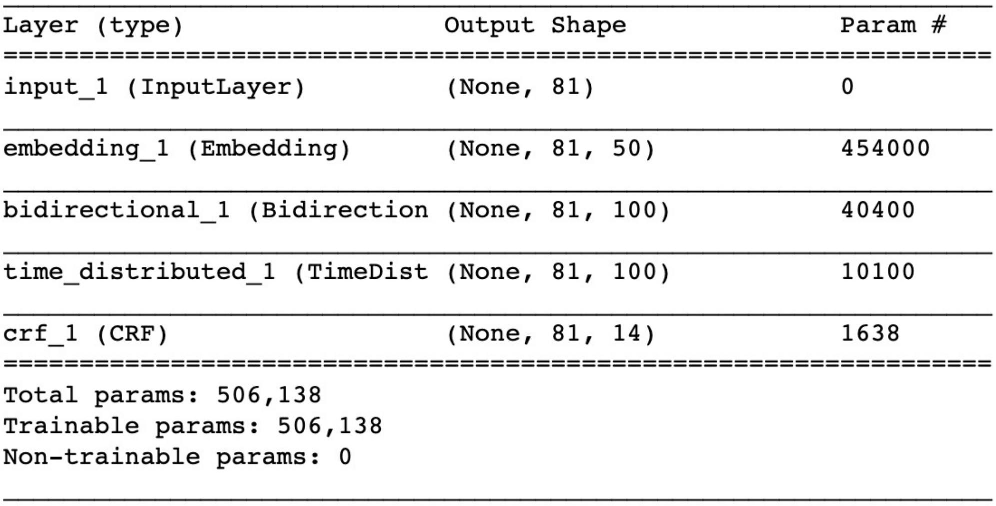


Fig. 21. Training Process on 2009 Dataset

The structure of the model for the i2b2 challenge 2012 dataset NER task is shown in Figure 22. Several groups of control experiments are carried out under this model as well. The main parameters of the final model are shown in Table 8, and the optimal performance on the test set is obtained after 10 epoch training.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 81)                0
_________________________________________________________________
embedding_1 (Embedding)      (None, 81, 50)            454000
_________________________________________________________________
bidirectional_1 (Bidirection (None, 81, 100)           40400
_________________________________________________________________
time_distributed_1 (TimeDist (None, 81, 100)           10100
_________________________________________________________________
crf_1 (CRF)                  (None, 81, 14)            1638
=================================================================
Total params: 506,138
Trainable params: 506,138
Non-trainable params: 0
_________________________________________________________________
```

Fig. 22. Network of BiLSTM Model on the 2012 Dataset

parameter	value
DENSE_EMBEDDING	50
LSTM_UNITS	50
LSTM_DROPOUT	0.2
DENSE_UNITS	100
BATCH_SIZE	256
MAX_EPOCHS	30
LEARNING_RATE	0.0001

Table 8. Model Parameters of BiLSTM+CRF for the 2012 Dataset

With 70% of training dataset, 15% validation dataset and 15% test dataset, the training process for NER task on the i2b2 2012 dataset is shown below::

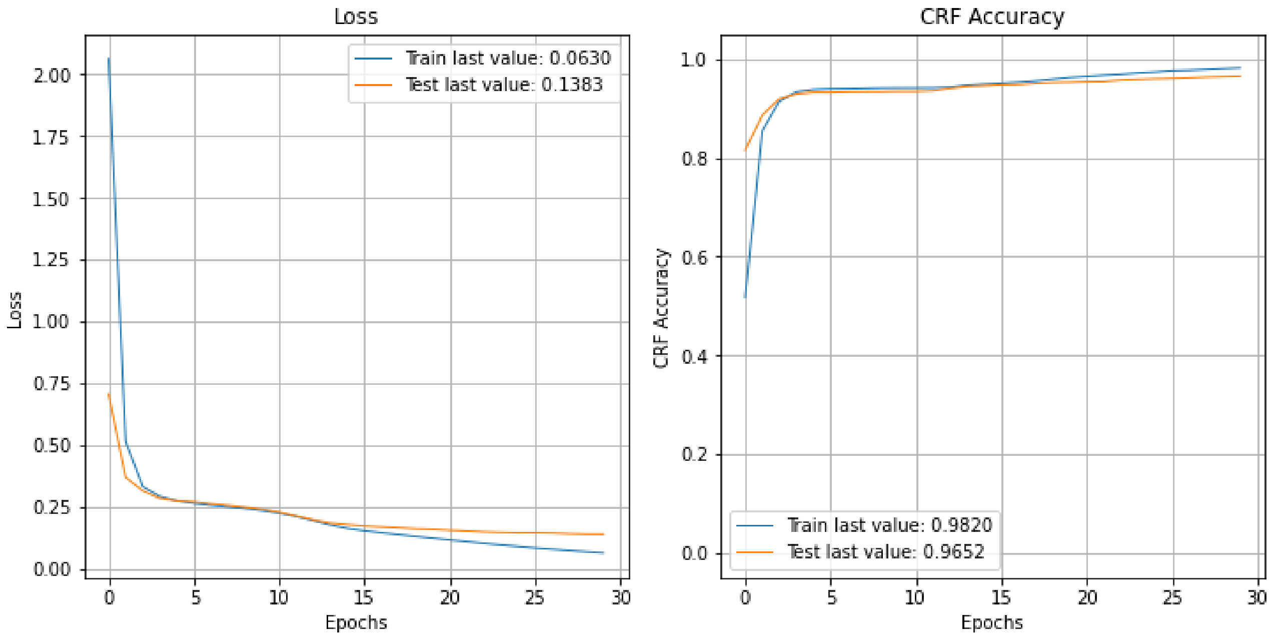


Fig. 23. Training Process on the 2012 Dataset

Table 9 shows the BiLSTM+CRF model performance on the test dataset for the i2b2 challenge 2009; the model is initialised with one-hot word embedding. As mentioned, evaluation metrics were based on exact-matching criteria and with this setting, the BiLSTM+CRF model achieves a 0.98 weighted average accuracy and 0.69 of Macro averaging. Apparently, combined with the evaluation scores of each subcategory, it can be seen that the amount of entities of each category varies significantly due to the uneven data distribution, which leads to the scores between entity categories varying wildly.

As demonstrated in Table 9, this BiLSTM+CRF model has better prediction performance on class ”o” and ”B-m’ in the 2009 dataset.

Table 9. NER Evaluation on the 2009 Dataset with BiLSTM-CRF.

category	Number / %			Number
	precision	recall	f1-score	support
PADDING	99.99	1.00	1.00	111422
B-do	81.88	70.52	75.78	519
I-do	84.09	74.66	79.10	446
B-m	81.35	66.51	73.19	1069
I-m	87.60	70.42	78.07	622
B-f	83.50	77.48	80.37	444
I-f	75.44	61.43	51.81	166
B-du	61.11	15.28	24.44	72
I-du	44.44	17.78	25.40	135
B-r	39.80	20.42	26.99	191
I-r	41.67	10.64	16.95	141
B-mo	83.05	78.61	80.77	374
I-mo	0.00	0.00	0.00	20
o	95.87	98.50	97.17	32429
accuracy	98.63	148050		
macro avg	68.56	53.76	58.55	148050
weighted avg	98.46	98.63	98.50	148050

In general, the BiLSTM+CRF model shows better performance on the 2009 dataset than the 2012 dataset for the NER task. Table 10 shows that it obtains 97.11 of average accuracy, which is slightly lower than the 2009 dataset. Similarly, the effect of the amount of entities of sub-class for training is still not avoidable in the 2012 dataset. For this phenomenon, Precision-Recall is a useful metric of success of prediction evaluation when the classes are imbalanced. In information extraction, precision is a measure of result relevancy, while recall is a measure of how many genuinely relevant results are returned.

4.2.2 Model-I Discussion and Summary

The results of NER are shown in Tables 9 and 10 for the 2009 and 2012 datasets, respectively. For the overall trend, I observed that the prediction performance really depends on the amount of sub-category data for training, which is that the more data, the more information obtained, leading to a more excellent score. For entity categories with a small number of entities, the recognition effect is poorer due to insufficient model learning. For example, sub-classes B-r and I-du have lower precision than B-m and B-do in the 2009 dataset. However, there are some exceptions that with more training data samples, the model shows poorer results for sub-class B-r than sub-class B-du. In addition, the results based on precision, recall and f1-score are pretty different (for most sub-classes, the model has higher precision than recall which means it returns very few results, but most of its predicted labels are correct when compared to the training labels. Namely, the percentage of Negative labels is lower). In conclusion, the BiLSTM+CRF model can not solve the problem

Table 10. NER Evaluation on the 2012 Dataset with BiLSTM-CRF.

category	Number / %			Number
	precision	recall	f1-score	support
PADDING	1.00	99.99	99.99	63867
B-CLINICAL_DEPT	68.85	32.81	0.44	128
I-CLINICAL_DEPT	80.43	54.95	65.29	202
B-EVIDENTIAL	1.00	4.11	7.89	73
I-EVIDENTIAL	0.00	0.00	0.00	6
B-OCCURRENCE	61.69	44.39	51.63	410
I-OCCURRENCE	34.69	7.46	12.27	228
B-PROBLEM	59.10	56.64	57.85	602
I-PROBLEM	75.67	61.92	68.11	864
B-TREATMENT	66.80	68.31	67.55	486
I-TREATMENT	65.87	67.63	66.74	448
B-TEST	58.48	54.18	56.25	299
I-TEST	64.42	63.23	63.82	378
O	87.82	95.83	91.65	75.82
accuracy	97.11	75573		
macro avg	65.99	50.82	53.82	75573
weighted avg	96.89	97.11	96.88	75573

of unbalanced data distribution well. On the dataset with unbalanced data distribution, the performance for the entity category with a small number of samples is poor.

In order to explore the effect of different embedding inputs for the BiLSTM+CRF model, I trained the model with Glove (glove.6B.50d), word2vector and one hot encoding and the average and standard deviations of metrics are recorded. Table 11 demonstrates the performance of each model on the same dataset. Glove embedding achieved 72.48% of precision, 52.58% of recall and 60.86% of f1-score, respectively which increased precision by around 2%. For word2vec embedding, this model achieved exact 73.62% of precision, 60.34% of recall and 61.48% of f1-score achieving an improvement of 1.14% in precision, an increasing of 7.76% in recall and an improvement of 0.62% than Glove embedding.

Table 11. Results of Experiments on the 2009 Test Dataset.

embedding	precision	recall	f1-score
on-hot-BiLSTM+CRF	68.56	53.76	58.55
Glove-BiLSTM+CRF	72.48	52.58	60.86
word2vec-BiLSTM+CRF	73.62	60.34	61.48

4.3 Model-II: BiLSTM-CNN

The Convolutional Neural Network (CNN) used in this approach draws inspiration from Chiu and Nichols (2016) hybrid model, which uses bidirectional LSTMs to learn both character and word-level information. As a result, our model employs word embeddings, in addition to other human-created word features and character-level information retrieved using a convolutional neural network. All of these features are sent into a BiLSTM for each word. Referring to the great work having done and following the tutorial created by Maximilian Hofer (Hofer, 2018) with the code adapted from:Git, I implemented this BiLSTM-CNN model for the NER task on the 2009 dataset. The architecture of this model from that paper is demonstrated in figure 24. The algorithm idea of this model is to make NER prediction by CONCAT of character-level features through CNN and word-level features and then by bidirectional LSTM. This enables the model to make better use of character-level features such as prefix and suffix, which can reduce the work of manual feature construction. Character-level features are obtained by a CNN technology which has been approved as a great success in the NER task and POS tagging task (Chotirat and Meesad, 2021; Labeau et al., 2015). Per-character feature vectors, such as character embeddings and character kinds, are concatenated with a max layer to produce a new feature vector, which is then used to train a model for recognising words.

The BLSTM layer forms the core of the network and has the following three inputs:

1. Character-level patters are identified by a convolutional neural network
2. Word-level input from GloVE embeddings
3. Casing input (whether words are lower case, upper case, etc.)

One of the critical tasks in (Chiu and Nichols, 2016) work is feature fusion, and the core features include Word Embedding, Character Embedding and Additional features. Additional features include Additional word features and Additional character features to the character features and word features. For example, words can be divided into six categories, including all uppercase letters, all lowercase letters, all numbers, some numbers and so on. The model includes 7 layers of network structure:

1. Character embedding layer, which maps the input of characters to 30 dimensional embedding.
2. Dropout layer (0.5), which mitigates the effect of overfitting
3. 1D convolutional layer, which transforms the character dimension size into 1
4. Word embedding layer, which maps the words into 50 dimensional embedding vectors with Glove (glove.6B.50d)

5. Concatenation layer, which combines processed character-level, word-level and casing data into a vector of 80 dimensions

6. BiLSTM layer: using the merged word vector sequence as input, the spatial semantic modelling of the preceding and subsequent text information was carried out, and the bidirectional semantic dependence of word vectors was captured. The high-level feature expression of the context information of medical records was further constructed

7. Dense output layer, which applies softmax function for prediction.

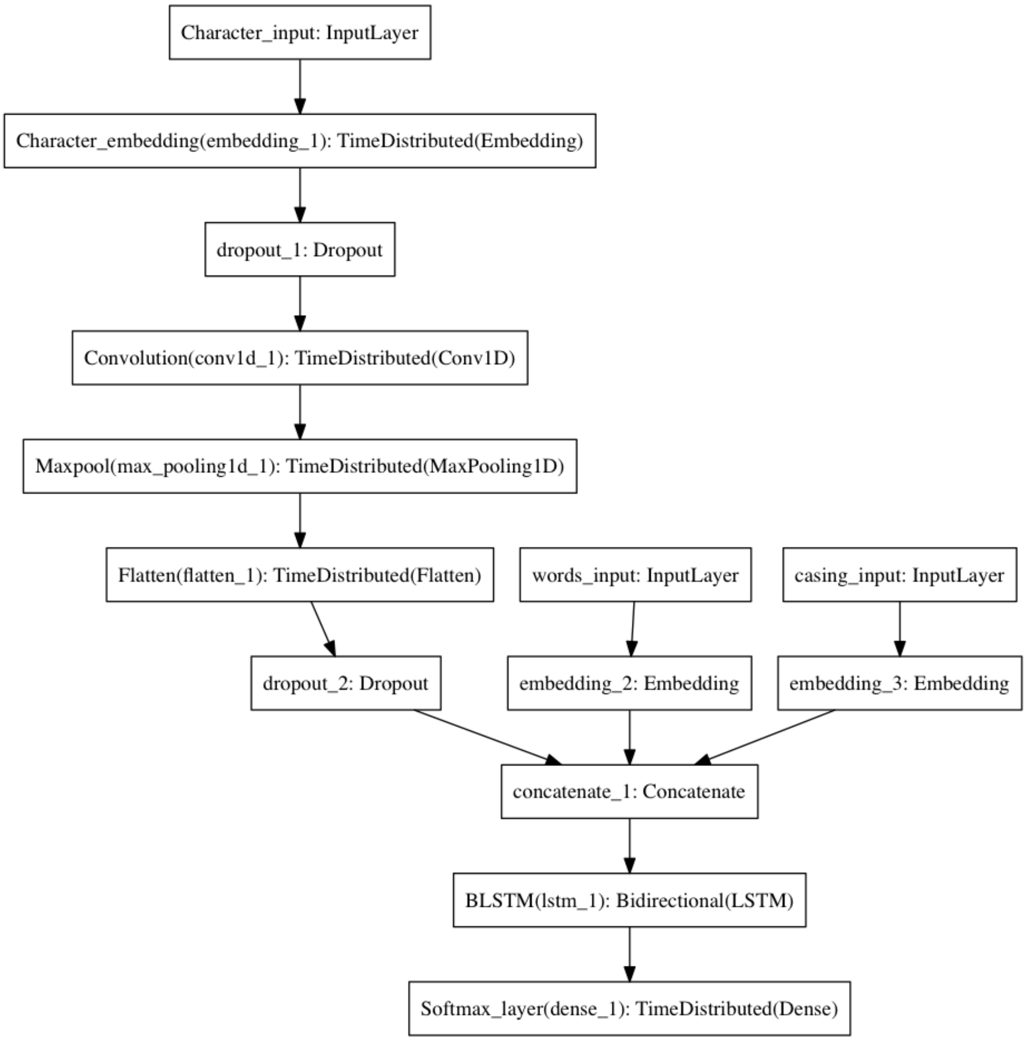

Character_input: InputLayer
Character_embedding(embedding_1): TimeDistributed(Embedding)
dropout_1: Dropout
Convolution(conv1d_1): TimeDistributed(Conv1D)
Maxpool(max_pooling1d_1): TimeDistributed(MaxPooling1D)
Flatten(flatten_1): TimeDistributed(Flatten)
words_input: InputLayer
casing_input: InputLayer
dropout_2: Dropout
embedding_2: Embedding
embedding_3: Embedding
concatenate_1: Concatenate
BLSTM(lstm_1): Bidirectional(LSTM)
Softmax_layer(dense_1): TimeDistributed(Dense)

Fig. 24. Network of BiLSTM-CNN Model on the 2009 Dataset (Chiu and Nichols, 2016)

4.3.1 Model-II Evaluation

This BiLSTM-CNN model was set for training for 30 Epoch but convergence after 14 Epoch. Table 12 shows the parameters used in this model.

parameter	value
DENSE_EMBEDDING	30
DROPOUT	0.5
WORDS_EMBEDDING	50
DROPOUT_RECURRENT	0.25
LSTM_STATE_SIZE	200
CONV_SIZE	3
LEARNING_RATE	0.0105
OPTIMIZER	Nadam()

Table 12. Model Parameters of BiLSTM+CNN for the 2009 Dataset

Table 13 indicates the performance (average precision, recall and f1-score) of the CNN-BiLSTM model, which shows better performance than the BiLSTM+CRF model (exact 85.67% of precision, 87.83% of recall and 88.17% of f1-score).

Table 13. Results of Experiments on the 2009 Test Dataset with BiLSTM-CNN.

model	precision	recall	f1-score
CNN-BiLSTM	75.67	77.83	78.17

4.3.2 Model-II Discussion and Summary

Initially, this BiLSTM-CNN model was created to solve the situation that most NER task models depend on the availability of large amounts of labelled training text dataset (Hofer et al., 2018)which is rare in clinical text because of privacy agreement between patients and doctors. However, for i2b2 2009 dataset, the dataset for training is adequate to retrieve powerful and useful information for the NER task with BiLSTM-CNN models. This model achieves an increase in precision by approximately 10% than the same model on the 2003 CoNNL dataset.

4.4 Model-III: BERT-based CNN

Although the [CLS] tags in BERT output can achieve good classification results, the rich semantic knowledge contained in BERT is not fully utilised. Therefore, in this dissertation, inspired by (Michalopoulos et al., 2020), the convolutional neural network is fused to expand BERT (model structure demonstrated in Fig. 25). After the BERT model receives the processed text, the content of the EMRs text is represented by a vector through the two-layer Transformer mechanism of the

BERT model. The model outputs a vectorised representation of comprehensive semantic information fused by each token vector, sentence vector and feature vector in the EMRs and then inputs the output of the model to the convolutional neural network. In this dissertation, three different convolutional kernels are used to capture different feature information. After the fully connected layer is connected by word vector mapping, the convolutional neural network model further extracts the semantic information of the dialogue text.

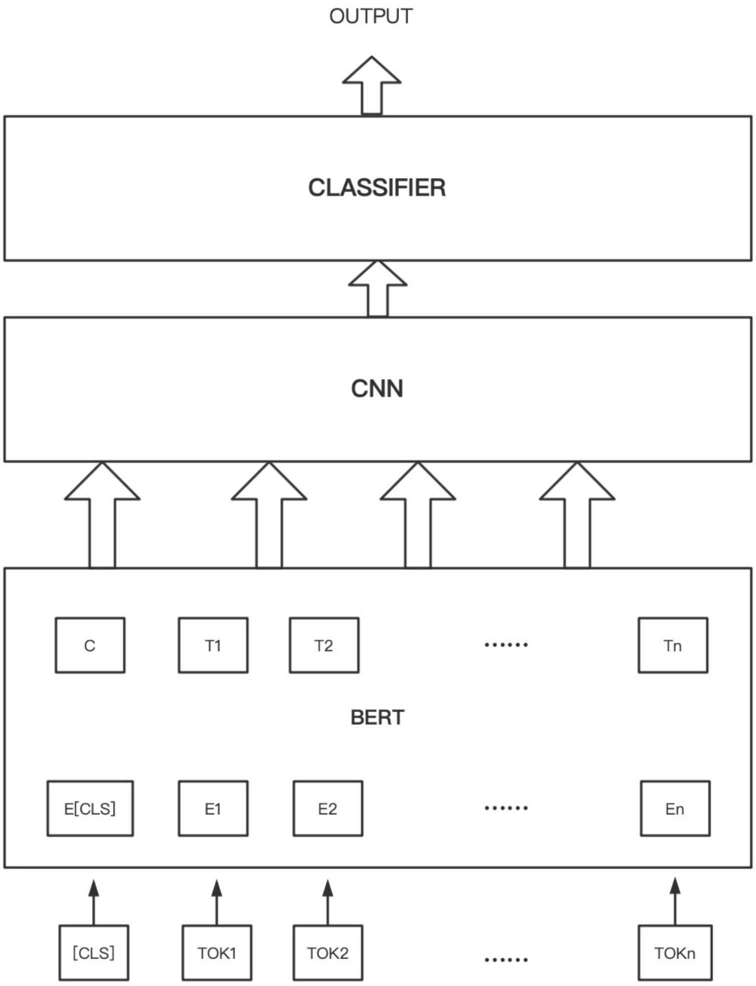


Fig. 25. Architecture of BERT-CNN Model

Code adapted from UmlsBERT and MedicalRelationExtraction I trained this BERT+CNN model based on the parameters on the i2b2 challenge 2012 dataset as follows (Table 14).

parameter	value
BERT_MODEL	$Bio-BERT_{base}$
DROPOUT	0.5
EPOCH	5
DROPOUT	0.1
MAX_SENTENCE_LENGTH	512
HIDDEN_SIZE	768
LEARNING_RATE	1e-05

Table 14. Model Parameters of BERT+CNN

4.4.1 Model-III Evaluation

In order to mitigate the influence of the training dataset imbalance of each sub-class, I down-sampled the dataset to 3000 samples for each class (AFTER, OVERLAP and BEFORE). Table 15 gives the result of the prediction on the training dataset. Table 16 shows the result of $BERT_{base}+CNN$ model performance for the relation extraction task on the test data of the 2012 dataset. This model achieves exact 64.48% of precision, 67.17% of recall and 65.03% of f1-score.

Table 15. Temporal Relation Extraction Evaluation on the 2012 Training Dataset.

category	Number / %			Number
	precision	recall	f1-score	support
AFTER	96.23	97.74	96.86	3000
OVERLAP	93.71	95.85	94.80	3000
BEFORE	97.13	93.24	95.06	3000
accuracy			95.60	9000
macro avg	95.60	95.60	95.60	9000
weighted avg	95.60	95.60	95.60	9000

Table 16. Temporal Relation Extraction Evaluation on the 2012 Test Dataset

category	Number / %			Number
	precision	recall	f1-score	support
AFTER	34.52	49.00	40.52	1122
OVERLAP	67.34	77.23	71.94	4078
BEFORE	91.62	75.28	82.64	6000
accuracy			73.32	11200
macro avg	64.48	67.17	65.03	11200
weighted avg	77.03	73.34	74.47	11200

4.4.2 Model-III Discussion and Summary

As the prediction shows, with BERT as a word vector extraction or directly as a classification model, the BERT+CNN model can achieve a good classification result, indicating that the pre-trained model BERT can extract the semantic information well of dialogue text and also gain a good experimental effect. Although the BERT+CNN model can realise excellent performance for relation extraction, over-fitting on the training dataset is still inevitable, especially for sub-class AFTER and OVERLAP, which show 61.71% and 26.37% lower than the prediction results on the training dataset, respectively. Basically, there is an apparent correlation between the number of testdata samples and prediction precision which is that the more supporting samples, the better the prediction performance.

4.5 Post-processing

In order to apply the prediction results of the NER task for medication recognition and the RE task for temporal relation extraction between medication event and date on medication usage status, I design a set of rules that determine whether a medication is in use. I take the date of Admission and Discharge as a baseline and take the relation between Admission and Discharge Date and medication into consideration since the dataset texts come from de-identified discharge summaries (HOSPITAL COURSE), which give the treatment information of patients during hospital time.

1. If the relation between Admission Date and a medication AFTER and the relation between Discharge Date and the medication is BEFORE or OVERLAP, it means the medication is in use

2. If the relation between Admission Date and a medication OVERLAP and the relation between Discharge Date and the medication is OVERLAP, it means the medication is in use

3. If the relation between a date except for Admission and Discharge Date which is between them and a medication is OVERLAP and the relation between Discharge Date and the medication is BEFORE or OVERLAP, it means the medication is in use

4. Otherwise it is not.

With this set of rules, I manually checked the classification results, showing that nearly three out of ten can be classified correctly. For example, the following text:
Admission Date : [R] 2015-09-14 [R] Discharge Date : [R] 2015-09-19 [R] Service : NEONATOLOGY HISTORY OF PRESENT ILLNESS : The patient is a 3285 gm infant born at 37 5/7 weeks to a 21 year old G3 P1 now 2 mother with prenatal screens as follows : O positive , antibody negative , hepatitis B surface antigen negative , RPR nonreactive , GBS negative . Unremarkable pregnancy except for minor fullness of the left renal pelvis reported during the week prior to delivery . Past OB history remarkable for postpartum depression . Mother was admitted in labor . Baby was delivered by repeat C-section with rupture of membranes at delivery . Apgars were 8 and 9 . CMED CSRU staff was called about 30 minutes of age for grunting , flaring and retractions and the baby was admitted to the CMED CSRU . HOSPITAL COURSE: 1. Respiratory . The Athol Memorial Hospital hospital course was initially consistent with transient tachypnea of the newborn . Chest x-ray revealed mild streakiness of the lung fields . He was initially placed on [L] nasal cannula [L]

The predicted relation between ’nasal cannula’ and Discharge Date is BEFORE, and the predicted relation between ’nasal cannula’ and Admission Date is AFTER, which indicates that the medication treatment is in use according to the rules. Table 17 shows some correct classifications.

Table 17. Medication Status Classification Structured Output

id	medication	status	
		IN USE	OUT OF USE
23	Enfamil	2015-09-14- 2015-09-19	2015-09-19 - UNKNOWN
23	Antibiotics	UNKNOWN- 2015-09-19	2015-09-19 - UNKNOWN
151	Diuretics	1993-03-26 - 1993-04-03	1993-04-03 - 1993-06-13

4.6 Chapter Summary

This chapter gives the results of NER task models (BiLSTM+CRF and BiLSTM+CNN) for the 2009 dataset and 2012 dataset and the RE model (BERT+CNN) for the 2012 dataset, and the analysis of model performance.

5 Conclusions and Future Work

5.1 Conclusions

Up to this section, I have discussed the work I have done. I started with introducing the necessity of temporal information extraction from clinical texts and giving the current research status as well as the direction. Following the commonly used research steps for text information extraction, I presented methods for the NER and RE tasks. With the in-depth research of scholars on artificial intelligence and deep learning and the development of computational power of computers, machine learning and deep learning based techniques (especially RNNs) have been proved better at learning from what has been observed in the text and generating features. I focused on how to recognise the MEDICATION and DATE entities and extract the temporal relationship between them with machine learning-based models. To achieve creating a table presenting MEDICATION use status, I created a set of rules based on relation prediction results to make specific decisions for the status which obtains an acceptable prediction accuracy of approximately 33%.

Named entity recognition, and relation extraction are essential components in the field of natural language processing, as well as critical technologies in application fields such as machine translation, information extraction and knowledge answering. With the development of the information technology, the demand for text processing in various fields is increasing, which also puts forward higher requirements for the tasks of named entity recognition and relation extraction in different environments. Traditional named entity recognition and relation extraction methods rely on manually designed features, which require a lot of human resources and expert knowledge. There are many limitations and shortcomings in processing massive texts with complex multi-fields. With the development of deep learning techniques, many new models have been proposed, and the application of deep learning in named entity recognition and relationship extraction tasks has become increasingly common. Using deep learning for named entity recognition and relationship extraction can ignore the manual feature design step and solve the dimensionality disaster problem caused by sparse data representation. Therefore, inspired by other scholars and researchers, this dissertation proposes a named entity recognition method based on deep neural networks and a BERT-based CNN architecture for relation extraction, combined these two sub-tasks were solved to extract temporal information from clinical texts. The specific research work is as follows:

1. Based on data statistic analysis, make CoNLL format and clean data after pre-processing.
2. Combined with a neural network, build a NER model based on BiLSTM+CRF and BiLSTM+CNN, applying different embedding models as input to explore the effect of sophisticated word embedding (one-hot, Glove and word2vec).
3. With the prediction output of medication entity recognition, build the BERT+CNN model to extract the relation between EVENTS

4. Combining the result of medication and the relation extraction results, create a set of rules to classify whether a medication is in use during a specific time.
5. To generate structured format outputs which make a table telling the information of the use status of medications.

To some extent, these models have achieved the expected performance on the NER task and RE task. However, the result of the ultimate output for medication use status prediction is not well as expected.

5.2 Future Work

This dissertation' main sources of experimental data are the i2b2 challenge 2009 and 2012, datasets which are considerable corpora leading to a lack of computation resources to do more experiments for fine-tuning the best model settings and parameters. Moreover, the proportion of data (those annotated) on the relationship between medications and their corresponding dates is inadequate, which can not provide sufficient information for MEDICATION-DATE relation extraction. Therefore, the performance of information extraction generated by the rules created for patients' medication usage status falls short of ideal standards, which require further improvement. For NER and RE tasks, with the development of NLP and other related fields, more text features could be applied to improve model performance in the future.

Additionally, even though the BiLSTM-BiLSTM and BiLSTM-CRF models have been proved that they could achieve great performance for NER task, more research regarding the structure of the models can be further explored. Besides, my work did not show much of original contribution towards new ideas of modelling improvement, which is supposed to be my next direction of future efforts. Further more, some Bootstrapping relation extraction systems can be taken into account for better accuracy for temporal relation extraction in the future exploring.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2019. A multi-task approach for named entity recognition in social media data. *arXiv preprint arXiv:1906.04135*.

Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen. 2015. Bidirectional recurrent neural network language models for automatic speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5421–5425. IEEE.

Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. 2018. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. SemEval-2016 Task 12: Clinical TempEval. Technical report.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Yung-Chun Chang, Hong-Jie Dai, Johnny Chi-Yang Wu, Jian-Ming Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2013. TEMPTING system: A hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of biomedical informatics*, 46(6):S54–S62.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.

Saranlita Chotirat and Phayung Meesad. 2021. Part-of-speech tagging enhancement to natural language processing for thai wh-question classification with deep learning. *Heliyon*, 7(10):e08216.

William A Chren. 1998. One-hot residue coding for low delay-power product cmos design. *IEEE Transactions on circuits and systems II: Analog and Digital Signal Processing*, 45(3):303–313.

Anne Cocos, Alexander G Fiks, and Aaron J Masino. 2017. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Adrien Coulet, Nigam H Shah, Yael Garten, Mark Musen, and Russ B Altman. 2010. Using text to build semantic networks for pharmacogenomics. *Journal of biomedical informatics*, 43(6):1009–1019.

Bontcheva K Cunningham H, Maynard D. 2002. GATE: an Architecture for Development of Robust HLT Applicas. (2).

Thanh Hai Dang, Hoang-Quynh Le, Trang M Nguyen, and Sinh T Vu. 2018. D3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, 34(20):3539–3546.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Sean R Eddy. 2004. What is a hidden markov model? *Nature biotechnology*, 22(10):1315–1316.

Mats Fredriksen, Alv A. Dahl, Egil W. Martinsen, Ole Klungsøyr, Jan Haavik, and Dawn E. Peleikis. 2014. Effectiveness of one-year pharmacological treatment of adult attention-deficit/hyperactivity disorder (ADHD): An open-label prospective study of time in treatment, dose, side-effects and comorbidity. *European Neuropsychopharmacology*, 24(12):1873–1884.

Chenquan Gan, Qingdong Feng, and Zufan Zhang. 2021. Scalable multi-channel dilated cnn–bilstm model with attention mechanism for chinese textual sentiment analysis. *Future Generation Computer Systems*, 118:297–309.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Tianyong Hao, Xiaoyi Pan, Zhiying Gu, Yingying Qu, and Heng Weng. 2018. A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts. *BMC medical informatics and decision making*, 18(Suppl 1):22–22.

P. Hartley, R. Flamary, N. Jackson, A. S. Tagore, and R. B. Metcalf. 2017. Support vector machine classification of strong gravitational lenses. *Monthly Notices of the Royal Astronomical Society*, 471(3):3378–3397.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Maximilian Hofer. 2018. Deep learning for named entity recognition #2: Implementing the state-of-the-art bidirectional lstm + cnn model for conll 2003. `https://towardsdatascience.com/deep-learning-for-named-entity-recognition-2-implementing-the-state-of-the-art-bidirec`

Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. 2018. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*.

Bin Ji, Rui Liu, Shasha Li, Jie Yu, Qingbo Wu, Yusong Tan, and Jiaju Wu. 2019. A hybrid approach for named entity recognition in chinese electronic medical record. *BMC medical informatics and decision making*, 19(2):149–158.

Min Jiang, Yukun Chen, Mei Liu, S. Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):601–606.

Shaohua Jiang, Shan Zhao, Kai Hou, Yang Liu, Li Zhang, et al. 2019. A bert-bilstm-crf model for chinese electronic medical records named entity recognition. In *2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pages 166–169. IEEE.

Young-Min Kim and Tae-Hoon Lee. 2020. Korean clinical entity recognition from diagnosis text using bert. *BMC Medical Informatics and Decision Making*, 20(7):1–9.

Veysel Kocaman and David Talby. 2021. Spark nlp: natural language understanding at scale. *Software Impacts*, 8:100058.

Matthieu Labeau, Kevin Löser, and Alexandre Allauzen. 2015. Non-lexical neural architecture for fine-grained pos tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 232–237.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yu-Kai Lin, Hsinchun Chen, and Randall A Brown. 2013. MedTime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, 46(6):S20–S28.

Xiao Ling, Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2008. Spectral domain-transfer learning. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 488–496.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Kai Ma, Yongjian Tan, Miao Tian, Xuejing Xie, Qinjun Qiu, Sanfeng Li, and Xin Wang. 2022. Extraction of temporal information from social media messages using the BERT model. *Earth science informatics*, 15(1):573–584.

Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, 273.

Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2020. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *arXiv preprint arXiv:2010.10391*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. 2021. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks.

Niels Peek, Roque Marîn Morales, and Mor Peleg. 2013. Artificial intelligence in medicine : 14th Conference on Artificial Intelligence in Medicine, AIME 2013, Murcia, Spain, May 29-June 1, 2013 : proceedings. Lecture notes in computer science, 7885. Lecture notes in artificial intelligence, Berlin ;. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

A.Heimonen Pyysalo, S.Airola. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9:s3–s2.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Ruth M Reeves, Ferdo R Ong, Michael E Matheny, Joshua C Denny, Dominik Aronsky, Glenn T Gobbel, Diane Montella, Theodore Speroff, and Steven H Brown. 2012. Detecting temporal expressions in medical narratives. *International journal of medical informatics (Shannon, Ireland)*, 82(2):118–127.

Kirk Roberts, Bryan Rink, and Sanda M Harabagiu. 2013. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):867–875.

Sumit Saha. A comprehensive guide to convolutional neural networks. [EB/OL]. `https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53` Accessed Dec 25, 2018.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

K.-H. Schriever. 2014. *G3P : good privacy protection practice in clinical research : principles of pseudonymization and anonymization*. De Gruyter, Berlin ;.

Zhenjie Shi, Zhaowei Dong, Chaoyi Pang, Bailing Zhang, and Lihui Zhang. 2020. Sentiment analysis of e-commerce comments based on bert-cnn. *Intelligent computers and Applications*, 10(2):7–11.

Maria Skeppstedt, Maria Kvist, and Hercules Dalianis. 2012. Rule-based entity recognition and coverage of SNOMED CT in Swedish clinical text. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1250–1257, Istanbul, Turkey. European Language Resources Association (ELRA).

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Keh-Yih Su, ichi Tsujii, Jong-Hyeok Lee, and Oi Yee Kwong. 2004. Lecture Notes in Artificial Intelligence 3248 Subseries of Lecture Notes in Computer Science. Technical report.

Weiyi Sun, Anna Rumshisky, Ozlem Uzuner, Peter Szolovits, and James Pustejovsky. 2012. The 2012 i2b2 temporal relations challenge annotation guidelines. *Manuscript, Available at https://www. i2b2. org/NLP/TemporalRelations/Call. php*.

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):828–835.

Yaodong Tang, Yuchen Huang, Zhiyong Wu, Helen Meng, Mingxing Xu, and Lianhong Cai. 2016. Question detection from acoustic features using recurrent neural network with gated recurrent unit. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6125–6129. IEEE.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Yanhua Wang, Zhihao Yang, Hongfei Lin, and Yanpeng Li. 2012. A syntactic rule-based method for automatic pathway information extraction from biomedical literature. In *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pages 626–633. IEEE.

Lawrence L Weed et al. 1968. Medical records that guide and teach. *N Engl J Med*, 278(11):593–600.

Wikipedia. Long short-term memory. [EB/OL]. `https://en.wikipedia.org/wiki/Long_short-term_memory` Accessed Dec 25, 2018.

Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):1–37.

Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. 2015. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624.

Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. 2017. Clinical named entity recognition using deep learning models. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1812. American Medical Informatics Association.

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I-Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):849–858.

YX Yang, S-K Teo, Eric Van Reeth, CH Tan, IWK Tham, and CL Poh. 2015. A hybrid approach for fusing 4d-mri temporal information with 3d-ct for the study of lung and lung tumor motion. *Medical Physics*, 42(8):4484–4496.

Yi Yao and Gianfranco Doretto. 2010. Boosting for transfer learning with multiple sources. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1855–1862. IEEE.

Wenjing Yuan, Lin Yang, Qing Yang, Yehua Sheng, and Ziyang Wang. 2022. Extracting spatio-temporal information from chinese archaeological site text. *ISPRS International Journal of Geo-Information*, 11(3):175.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 73–78.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. 2018. A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, 81:83–92.

GuoDong Zhou, Dan Shen, Jie Zhang, Jian Su, and SoonHeng Tan. 2005. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC bioinformatics*, 6(1):1–7.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.