A picture containing logo

Description automatically generated

Does parameter sharing between named entity recognition and relations extraction models improve model performance for i2b2 medical concept and relation extraction?

MSc Applied Statistical Modelling and Health Informatics

King’s College London

DATE

01/09/2023

Module Name: Research Project

Module Code: 7PAVREPR

Student Number: K2257531

Supervisor (s): Angus Roberts/ Leo Xinyue Zhang

Word Count: 10180

Project Plagiarism Coversheet

This coversheet must be added to your submission to turnitin, as part of one dissertation submission.

|  |  |
| --- | --- |
| **Title of Project** | Does parameter sharing between named entity recognition and relations extraction models improve model performance for i2b2 medical concept and relation extraction? |
| **Supervisor** | Angus Roberts, Leo Xinyue Zhang |
| **Deadline** | 01/09/2023 15:00 |

**Statement of Academic Integrity**

I have read the College Regulations on Cheating and Plagiarism and state that this piece of work is my own and does not contain any unacknowledged work from any other hand-written, printed or internet source. I further state that I have acknowledged all sources used and quoted in the attached work. I understand the penalties that can be applied in cases of proven cheating and plagiarism.

|  |  |  |
| --- | --- | --- |
| **I have submitted a Word file of the entire project, names ‘STUDENTNUMBER#\_Project”, to Turnitin** | | **YES** |
| **Surname** | Li | |
| **First Name** | Qingzhou | |
| **Student Signature** |  | |
| **Date** | 01/09/2023 | |

Abstract

Accurate information extraction from medical texts has the potential to completely transform healthcare decision-making and research in the field of natural language processing (NLP). The essential tasks of named entity recognition (NER) and relation extraction (RE) make it easier to extract medical ideas and their relationships from clinical narratives. In the difficult setting of i2b2 medical discharge summaries, this work offers a thorough examination into the usefulness of parameter sharing across NER and RE models to improve medical concept and relation extraction.

This work establishes the development of NER, RE, and joint learning strategies in NLP based on a thorough literature assessment. On top of this foundation, the study develops research questions and hypotheses to investigate how parameter sharing would impact NER and RE tasks. The construction of separate NER and RE models as well as its combined counterpart is accomplished through the careful application of a methodology that covers the data source, data preprocessing, model architecture design and training phases. Thorough testing and analysis reveal complex benefits of parameter sharing in RE, and end-to-end relation extraction tasks. Our final parameter sharing model boosted the RE prediction by an absolute difference of 1.0% and for end-to-end relation extraction, the difference made up to 1.3% in strict micro-averaged f1-score.

The discussion delves into the strengths, limitations, and future work of parameter sharing between NER and RE. The identified strengths, showcase the approach's potential to enhance medical text analysis. Notably, the observed enhancements in relation classification validate parameter sharing's capacity to translate shared knowledge and contextual embeddings into tangible improvements, which results the same outcome for end-to-end relation extraction. Acknowledging limitations, such as the absence of statistical significance in certain results, the study suggests avenues for future research. An in-depth error analysis for NER and RE predictions emerges as a promising direction, offering insights into the model's behavior and informing shared feature representations and contextual embeddings. Additionally, the implications of parameter sharing extend beyond technical advancements, encompassing ethical considerations and real-world impact.

By bridging the gap between NER, RE, and joint learning, this paper sheds light on the potential of parameter sharing for medical concept and relation extraction. The results provide a road map for multidisciplinary NLP and healthcare collaboration, encouraging further investigation, validation, and improvement of collaborative learning strategies. Parameter sharing has the potential to change medical information extraction as technology and ethical issues grow, fostering improvements in healthcare delivery, research, and patient-centric decision support.

Table of contents

[Abstract 3](#_Toc144159715)

[Introduction 6](#_Toc144159716)

[Background 6](#_Toc144159717)

[Research question and hypotheses 6](#_Toc144159718)

[Objectives 7](#_Toc144159719)

[Motivation 7](#_Toc144159720)

[Section outline 8](#_Toc144159721)

[Literature review 9](#_Toc144159722)

[NER and RE 9](#_Toc144159723)

[National NLP Clinical Challenges 11](#_Toc144159724)

[Bidirectional LSTM 11](#_Toc144159725)

[Joint learning 12](#_Toc144159726)

[Methodology 14](#_Toc144159727)

[Data Source 14](#_Toc144159728)

[Method 14](#_Toc144159729)

[Evaluation metrics 15](#_Toc144159730)

[Data preprocessing 15](#_Toc144159731)

[Model Architecture Design 16](#_Toc144159732)

[Training 18](#_Toc144159733)

[Results 20](#_Toc144159734)

[Named entity recognition performance 20](#_Toc144159735)

[Relation extraction performance 21](#_Toc144159736)

[End-to-end relation extraction performance 21](#_Toc144159737)

[Discussion 23](#_Toc144159738)

[Strengths 23](#_Toc144159739)

[Limitations 24](#_Toc144159740)

[Future work 26](#_Toc144159741)

[Error Analysis for NER 26](#_Toc144159742)

[Error Analysis for RE 26](#_Toc144159743)

[Conclusion 28](#_Toc144159744)

[References 30](#_Toc144159745)

# Introduction

## Background

In the field of Electronic Health Records (EHRs), there exists massive amount of data of patients, including patients' basic information, disease history, drug history, allergy, adverse drug events, etc. The amount of data cumulates day by day and carries potential information insights. To extract information from these data, Machine Learning (ML) and Natural Language Processing (NLP) techniques are commonly used. Named Entity Recognition (NER) and Relations Extraction (RE) are two fundamental components of this information extraction pipeline, with each task addressing distinct challenges in identifying medical concepts and their interrelated associations.

Traditionally, NER and RE have been treated as separate tasks, with individual models specialized for each. People also used pipeline method for NER and RE, which cannot take full advantage of the benefits in between (Wang, 2021). However, in recent years, the rapid gross of NLP led to significant advancements in various applications, including the joint learning between the two tasks. Existing studies have shown that combining NER and RE tasks into a single joint model have a high performance for many concept types to training the two tasks independently or perform end-to-end tasks which is more close to real-world scenario (Henry et al., 2019). Research has also shown that leveraging inter-task dependencies and sharing knowledge between related tasks can lead to improved model performance in NLP applications (Wang, 2021). The concept of parameter sharing between NER and RE models is one such approach that has gained traction due to its potential for enhancing the overall performance and robustness of information extraction systems.

The i2b2 (Informatics for Integrating Biology and the Bedside) challenge is a well-known benchmark for evaluating NER and RE models on medical text. The National NLP Clinical Challenges (n2c2) organized in 2018, continued the legacy of i2b2, adding 2 new tracks and 2 new sets of data to the shared tasks organized since 2006. Track 2 focused on extraction of adverse drug events (ADEs) from clinical records and evaluated 3 tasks: concept extraction, relation classification, and end-to-end systems (Henry et al., 2019).

## Research question and hypotheses

The aim of this project is to discover the impact of parameter sharing between NER and RE models in terms of performance for i2b2 medical concepts and relation extraction. The research questions are as follows:

1. Does the parameter sharing improve the NER prediction performance in the joint model?
2. Does the parameter sharing improve the RE prediction performance in the joint model?
3. Does the parameter sharing improve the end-to-end relation extraction prediction performance in the pipeline model?

Our hypothesis of these research questions is that the parameter sharing will improve the individual and overall performance of NER and RE predictions, and the end-to-end relation extraction will also be improved by parameter sharing in the joint model.

## Objectives

The project which this paper is going to conduct, follows the similar overall structure as the 2018 n2c2 Track 2 challenge.

Objective 1: data access, we will need to get access to the same i2b2 dataset that was being used in the 2018 n2c2 Track 2 Challenge, training data, test data and gold standard data are all required for further adjustment.

Objective 2: data preprocessing, both the training data and gold standard data contains the raw context and the gold standard entity and relation annotations, the test data only contains the raw context, we need to perform preprocessing for them to make the fit the input of the models and this will be easier for us to do the testing.

Objective 3: pipeline model creation and evaluation, starting with developing two independent named entity recognition and relation extraction models based on the Bi-LSTM model, followed by training with BIO tagging annotations or entity to drug relations, each individual model will then prepare the reported performance of the baseline measure on the gold standard test datasets.

Objective 4: joint model creation and evaluation, build a parameter sharing model. This will be based on the two individual models, which shares some feature embeddings in the Bi-LSTM model, after being trained with BIO tagging annotations and entity to drug relations, we record the model performance on relations as the baseline measure for joint model. To justify our hypothesis, we continue to compare the test results from the pipeline model and joint model for each NER and RE section.

Objective 5: end to end evaluation. We will test the end-to-end model performance by comparing the test results between pipeline model and joint model that generated from the relations formed by the named entities predicted by each model’s NER part.

Objective 6: Analyze the benefits and limitations of parameter sharing in the context of medical text analysis.

## Motivation

The motivation behind this paper is to investigate the impact of parameter sharing between NER and RE models on i2b2 medical concept and relation extraction. We aim to explore whether the integration of task-specific information and shared knowledge can lead to better feature representations and more efficient learning. Additionally, we seek to examine the extent to which parameter sharing can aid in handling complex and ambiguous clinical text, where medical concepts and their relationships are often expressed in various linguistic patterns.

To address these research objectives, we conduct a comprehensive experimental study on a diverse dataset sourced from the i2b2 Challenge, encompassing a wide array of clinical narratives and complex medical relationships. By comparing the performance of dedicated NER and RE models with their parameter-sharing counterparts, we hope to identify the potential benefits and challenges associated with this approach. Furthermore, we assess the impact of varying model architectures, training strategies, and data augmentation techniques on the overall performance of the proposed shared-parameter framework.

The contributions of this paper lie in providing novel insights into the efficacy of parameter sharing between NER and RE models for i2b2 medical concept and relation extraction. The findings from this study may not only advance the state-of-the-art in joint learning information extraction but also shed light on the broader implications of knowledge sharing across related tasks in NLP. Ultimately, our research aims to contribute to the development of more accurate, robust, and efficient information extraction systems for the medical domain, with the potential to significantly impact healthcare decision-making and research.

In conclusion, this paper aims to shed light on the benefits of parameter sharing between NER and RE models for enhancing the performance of medical concept and relation extraction in the challenging domain of i2b2 medical discharge summaries. The findings of this study have significant implications for developing more efficient and accurate NLP systems, ultimately contributing to improved healthcare information management and clinical decision support.

## Section outline

The remainder of this paper is organized as follows:

Section 2 introduces the relevant literature about NER, RE and joint learning, also briefly discuss the choice of neural network algorithms that we use the developing the models.

Section 3 introduces the methodology and details in developing the NER, RE and joint models.

Section 4 presents the results and analysis, addressing the research questions and hypothesis. Section 5 discusses the limitations of the model and potential applications of parameter sharing in medical concept and relation extraction. We also talk about the future work and improvement that could being done to the project.

Section 6 concludes the paper with a summary of findings and future research directions.

# Literature review

## NER and RE

In order to facilitate the collection of data from electronic health records (EHRs), machine learning (ML) and natural language processing (NLP) have been widely deployed in the healthcare system. Two of the most crucial elements of NLP are named entity recognition (NER) and relation extraction (RE), which have been extensively used in tasks like information retrieval, question answering, sentiment analysis, and knowledge graph construction. In this literature review, we delve into the advancements, challenges, and state-of-the-art techniques in the fields of NER and RE.

A named entity is, roughly speaking, anything that can be referred to with a proper name: a person, a location, an organization. The task of named entity recognition (NER) is to find spans of text that constitute proper names and tag the type of named entity recognition (NER) the entity (Daniel and Martin n.d.). Named entity recognition has been widely used in many NLP applications, including text classification and information retrieval. The accurate recognition of these entities is essential for information retrieval and knowledge extraction from unstructured text. Early approaches to NER relied on rule-based systems and hand-crafted features. However, recent years have seen a paradigm shift towards data-driven methods, particularly those based on deep learning and neural networks.

To annotate the NER data, there exists several schemes, with IO being the most popular choice (Alshammari & Alanazi, 2021), which uses ‘I’ to tag all existing entity and ‘O’ to tag the words outside the entity. but it also comes with disadvantages, such as unable to distinguish boundaries. Other annotation schemes, such as IOB or BILOU, has the ability to incorporate boundary information with different techniques (Nasar et al., 2021). The IOB scheme uses three main tags: "B-" for the beginning of an entity, "I-" for tokens inside an entity, and "O" for tokens outside of any entity. For instance, consider the sentence "Apple Inc. is headquartered in Cupertino, California." In IOB tagging, this sentence might be annotated as follows: “B-ORG, I-ORG, O, O, O, B-LOC, O, B-LOC, O”, with ‘ORG’ representing organization and ‘LOC’ represents location, and all punctuations are included in the tagging. The capacity of IOB tagging to discriminate between adjacent items of the same type is one of its key benefits. When numerous entities are close to one another, misunderstanding can be avoided thanks to IOB tagging, which makes it clear where one entity stops and another begins. Moreover, the IOB tagging has the ability to manage named things with multiple words. Multi-word entities can be represented consistently because each token is specifically tagged, ensuring that each token within the entity is tagged correctly.

Relation extraction on the other hand is about finding and classifying semantic relation extraction relations among entities mentioned in a text, like child-of (X is the child-of Y), or part-whole or geospatial relations (Daniel & Martin, 2023). The history of this concept dates back to the 1980s where The Message Understanding Conference (MUC) was organized and developed since (Kartik Detroja et al., 2023). Traditional RE highly relies on the hand-crafted or learned from hand-labelled training examples, such as Riloff (1996) achieved 98% performance of a hand-crafted dictionary. Thus, this method has a big disadvantage of heavily relay on the human made relations and will encounter lots of difficulty when move to a new domain. To minimize the human effect in RE, the recent studies has proposed new RE method such as OpenIE (Etzioni et al., 2008) and lots of deep neural network-based systems.

For NER and RE tasks, there exists several approaches of the NLP methods. The first one is rule-based method. This approach uses a predefined set of rules for information extraction which are pattern-based and context-based. A pattern-based rule relies on the structural form of words, whereas a context-based rule utilizes the surrounding information about the word within the text document. One example to this is that Alfred et al. (2014) used such algorithm in Malay articles and achieved 89.47% on the F-measure value. Wu et al. (2022) proposed rule-based NER and RE approach for mechanical, electrical, and plumbing (MEP) information and reported performance 40% better than deep learning models.

Second approach is machine-learning based. This approach overcomes numerous constraints associated with the preceding rule-based methods. It constitutes a statistical model that aims to create a feature-oriented portrayal of the data under examination. It can identify a present entity name, even when minor spelling differences exist. The machine learning-centric strategy encompasses a dual-stage process for conducting Named Entity Recognition (NER). In the initial stage, the ML model undergoes training using annotated documents. Subsequently, in the following phase, the trained model gets applied to label the unprocessed documents. This procedure mirrors a typical pipeline for machine learning models. Similar to machine learning, this type of approach can be more specified into supervised, semi-supervised and unsupervised approach (Sharnagat, 2014), where supervised methods are subdivided into Hidden Markov Models (HMM), Support Vector Machines (SVM), Maximum Entropy, and Conditional Random Fields (CRF). Semi-supervised techniques leverage a bootstrapping methodology for executing Named Entity Recognition (NER). Unsupervised methods are elucidated through the illustration of a commonly employed unsupervised NER system, known as the KnowItAll system. Additionally, the utilization of unsupervised methods for non-English languages is also discussed within this context (Nasar et al., 2021).

Third approach is deep learning or neural network method. In fact, deep learning is a part of the neural network, they are all capable of detecting the semantic and syntactic relationship between words. Deep Neural Network (DNNs) has founded use in different architectures such as Convolutional Neural Network (CNN) (Liu et al., 2013), Recurrent Neural Network (RNN) (Zhang & Wang, 2015), Long short-term Memory (LSTM) (Gasmi et al., 2019), Gated Recurring Unit (GRU) (M.Abdelgwad et al., 2021), Bidirectional Encoder Representations from Transformers (BERT) (Harnoune et al., 2021). DNNs also has wide range of applications over the past twenty years to tasks such as part of speech tagging (POS) (Gopalakrishnan et al., 2019), sentiment analysis (Zhao et al., 2018), machine translation (Gehring et al., 2016) etc. These applications have given a great leap to the development of NER and RE.

Overall, NER and RE are important tasks in NLP, and recent advances in neural networks and pre-trained word embeddings with different model architectures have led to significant improvements in their performance. In this project, we will explore some approach to the NER and RE tasks and try to reach new state-of-the-arts method.

## National NLP Clinical Challenges

National NLP Clinical Challenges (n2c2) is an outgrowth of the former Informatics for Integrating Biology and the Bedside (i2b2), i2b2 was an NIH-funded National Center for Biomedical Computing (NCBC) based at Partners HealthCare System in Boston from 2004 to 2014 (Kohane, n.d.). Clinical texts include a wealth of information but processing them can be difficult due to their linguistic complexity, ambiguity, and terminology that is peculiar to the field. NLP systems that can help with better healthcare decision-making and outcomes are being developed as a result of n2c2 challenges, which imitate real-world clinical settings. The challenge also facilitate collaboration between NLP researchers, clinicians, and healthcare experts. This collaboration helps bridge the gap between technology and healthcare, ensuring that NLP solutions are not only accurate but also clinically meaningful. The beginning of the n2c2 challenge can trace back to 2006, where Uzuner et al. (2008) proposed method to solve the smoking prediction subtask. It aimed to solve the particular linguistic quirks, domain-specific terminology, and privacy issues related to clinical literature. The challenge's organizers selected datasets from a range of clinical disciplines, including radiological reports, electronic health records, clinical notes, and discharge summaries. From then, total of 11 challenges has taken place, with the most recent 2022 Workshop at AMIA 2022 Annual Symposium at Washington Hilton Hotel in DC in November. The challenge provides high-quality annotated datasets, often considered gold standards, for various clinical NLP tasks. These datasets are crucial for training and evaluating NLP models, enabling researchers to benchmark their methods against state-of-the-art approaches. Significant contributions and developments in clinical NLP have been made as a result of the n2c2 challenges, such as deep learning models (Oleynik et al., 2019), pre-trained language models such as BERT (Guan & Devarakonda, 2020), BioBERT (Alimova & Tutubalina, 2020), and RoBERTa (Kusalavan, 2023), and some feature engineering method (Catelli et al., 2021). Many participants of the challenge release their NLP models, code, and resources as open-source projects. This fosters the growth of a collaborative NLP community, encourages knowledge sharing, and accelerates progress in clinical NLP.

In this project, we use the dataset from the 2018 n2c2 challenge, which has two tracks. Track 1 aims to solve the task “Can NLP systems use narrative medical records to identify which patients meet selection criteria for clinical trials?” (Stubbs et al., 2019). Track 2 focus more on extraction of adverse drug events (ADEs) from clinical records. Our focus is similar to the Track 2, which will be evaluating three tasks: concept extraction, relation classification, and end-to-end systems.

## Bidirectional LSTM

A recurrent neural network (RNN) is any network that contains a cycle within its network connections, meaning that the value of some unit is directly, or indirectly, dependent on its own earlier outputs as an input (Daniel & Martin, 2023).

RNN is a class of neural networks that can process sequential data by maintaining hidden states that capture contextual information from previous inputs. However, traditional RNNs often suffer from the vanishing gradient problem, limiting their ability to capture long-range dependencies. Also, the training of RNNs arises from the need to backpropagate the error signal back through time (Daniel & Martin, 2023).

To address the limitations of traditional RNNs, the Bi-LSTM architecture was proposed. Bi-LSTM incorporates both forward and backward information flow by using two separate LSTM layers that process the input sequence in opposite directions. The forward LSTM captures information from past to future, while the backward LSTM captures information from future to past. The outputs of both directions are concatenated, allowing the model to effectively capture contextual information from both sides of the sequence.

In NLP, Bi-LSTM has been extensively used for various tasks. For instance, Li et al. (2017) applied Bi-LSTM to text classification, achieving state-of-the-art results by effectively capturing contextual information from both directions. Similarly, Ma and Hovy (2016) used Bi-LSTM for named entity recognition, showing improved performance by considering the surrounding context of named entities.

Attention is a mechanism combined in the RNN allowing it to focus on certain parts of the input sequence when predicting a certain part of the output sequence, enabling easier learning and of higher quality. Luo et al. (2017) has used an attention-based BiLSTM-CRF approach to the NER task, and has achieved better performance to other state-of-the-art methods on BioCreative IV chemical compound NER and RE.

In this project, we will use the Bi-LSTM model over the Attention model since Attention models tend to be computationally more expensive compared to Bi-LSTM models due to the need to calculate attention weights for each input element. Moreover, EHR data often requires interpretability for clinical decision-making and analysis. Bi-LSTM models provide hidden state representations at each time step, which can help interpret the internal dynamics of the model.

Overall, Bi-LSTM has proven to be an effective architecture for sequence modeling tasks in NLP. Its ability to capture contextual information from both directions has made it particularly useful for tasks requiring comprehensive understanding of input sequences. With the advent of deep learning, Bi-LSTM and its variants continue to be explored and adapted to further improve performance in a wide range of NLP applications.

## Joint learning

Joint learning involves training a single model to simultaneously perform multiple related tasks. Jointly learning NER and RE has garnered attention due to the inherent connection between the two tasks—accurate NER is often a prerequisite for successful RE. Joint learning exploits shared representations and enables the model to capture subtle interactions between entities and relationships. Research efforts have explored various architectures for joint NER and RE models. One common approach is to perform NER first followed by RE, Graves et al. (2013) proposed a Bi-LSTM model with same approach to address the speech recognition problem. Another approach is to share lower-level representations between the tasks, allowing the model to learn common linguistic features. Deep neural networks, such as BERT (Devlin et al., 2018), BiLSTM (Li et al., 2017) and Transformer-based architectures (Chen et al., 2020), have demonstrated success in joint learning by capturing context and dependencies in the text effectively. According to Liu et al. (2019), the use of Multi-Task Deep Neural Networks (MT-DNN) can achieve new state-of-the-arts results in ten new natural language understanding (NLU) tasks. Li et al. (2018) has combined two approaches and build a BiLSTM-CRF model to recognize entities and a BiLSTM-Attention network to extract relations, he further improves the models by deploy three MTL models by hard parameter sharing.

In Ruder (2017)’s overview of Multi-Task Learning (MTL), he proposed two methods for deep learning: hard parameter sharing and soft parameter sharing. Hard Parameter Sharing involves using the exact same set of parameters (weights and biases) across different tasks or components of a neural network. In this approach, the shared parameters are learned jointly for all tasks, and the same parameters are applied to process the input data for each task. This forces the model to find a common representation that works well for all tasks. Soft parameter sharing, on the other hand, allows for some level of differentiation between tasks or components. Instead of sharing exact parameters, the model encourages the parameters to be similar but not necessarily identical. This is typically achieved by regularizing the differences between parameters using techniques like weight decay or variational methods.

Miwa and Bansal (2016) presented a novel end-to-end neural model for extracting entities and relation between them. Their study used bidirectional sequential and bidirectional tree-structured LSTM-RNNs, which developed a single model to reach even better result than the state-of-the-art, feature-based system on end-to-end relation extraction. Additionally, they discovered that word sequence and dependency tree structure work well. Our study will be depend on these previous studies and explore the impact of parameter sharing on named entity recognition and relation extraction, with a further justification of an end-to-end system.

# Methodology

## Data Source

The data used for this project consisted of 505 discharge summaries drawn from the MIMIC-III (Medical Information Mart for Intensive Care-III) clinical care database (Johnson et al., 2016). The training data consists 303 discharge summaries and the test data consists 202 discharge summaries. These data were also used in 2018 n2c2 Track 2 shared task on adverse drug events.

There are in total of 9 entity types which are: ‘Drug’, ‘Strength’, ‘Form’, ‘Dosage’, ‘Frequency’, ‘Route’, ‘Duration’, ‘Reason’, ‘ADE’ which stands for adverse drug event. As for relation, there are 8 relations in total, which consists of all other entities to ‘Drug’. The number of entities and relations of each type in the test and training sets and the class distributions are given by Henry et al. (2019). Both concepts and relations are very similar for the test, training, and full datasets, and are shown in parentheses in *Table 1* for the full dataset.

表格

描述已自动生成

The training and testing data were provided to us under an NLP data use agreement through an online portal[[1]](#footnote-1), testing data are only being used after the model creation, and only for testing performance of each model.

## Method

The workflow of this project is shown in the picture below, to evaluate our research question, we will build two individual NER and RE models and a parameter sharing model. From there, we can evaluate the impact of parameter sharing on each individual NER and RE prediction and end-to-end relation extraction prediction.

图示

描述已自动生成

Figure 1. Overview of the project and workflow

### Evaluation metrics

There are variants of evaluation methods, such as precision, recall, and F1 with micro- and macro-averaged, with both strict and lenient matching according to Henry et al. (2019). The evaluation metrics we use in this project is strict F1 scores. For strict F1, we need to have all entity position to match as well as its relation.

### Data preprocessing

For training data, we are given with 303 raw medical discharge summaries and its hand annotated entity and relations respectively. Each entities are given with their entity type (e.g. Reason, Drug) and starting and end position in the file, and all the relations are related to the entity ‘Drug’, such as ‘Form-Drug’, or ‘Dosage-Drug’. There are a total of 9 entity types and 8 relation types, as listed in Table 1. We are also being given the annotation guide and some evaluation scripts for our development.

For NER task, we are using the Inside–outside–beginning (IOB) tagging format, where 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, the I-prefix indicates that the tag is inside a chunk, and the O-prefix indicates that the token belongs to no chunk. This scheme was initially proposed by Ramshaw and Marcus in 1995.

The first thing we need to do is to generate entities with corresponding sentence, we will read the files and generate a tag list that contains all the entities in the annotation file, this tag list will then be transformed into a python pandas DataFrame with column names: 'text\_id', 'entity\_id', 'start', 'end', 'concept', 'entity', 'sent\_num'. ‘text\_id’ indicates which number of discharge summaries this entity is in. ‘entity\_id’ follows the order when the entities are being annotated from the raw text, but they are not in the correct sequence as sentence goes on, thus they need an indication in case there are repeated entity. ‘start’ and ‘end’ indicates the start and end position index of this entity in its sentence. ‘concept’ indicates the entity type, and ‘entity’ is the original word itself. Finally, ‘sent\_num’ is the number of sentences in the text that this entity is in.

With the current DataFrame, we are going to create a list that contains the sentences from the discharge summaries and its corresponding IOB tagging. This is quite straight forward, since we have the entity index in the sentence, we can just label the entities with ‘B-‘ and ‘I-‘ tagging and rest of the sentence filled with ‘O’ tags. That is the data preprocessing steps for NER task. A diagram of an example sentence with it’s IOB tagging is formatted in the Figure 2 below.

|  |
| --- |
|  |
| Figure 2. Example of IOB tagging to a sentence. |

Now we move to RE task, since we are given the gold standard relations in the training data, proper formulation is still needed to appropriately arrange these data so that our model can read them and use them for training. We formulate a relation triplet as "start\_position\_1st\_entity, end\_position\_1st\_entity, start\_position\_2nd\_entity, end\_position\_2nd\_entity, relation\_type, 1st\_entity\_type, 2nd\_entity\_type". Each relation is concatenated by all the relation triplets within the sentence, separated by '|'. From the triplet, we can get each relation’s two starting position indexes and end position indexes in its original sentence, and corresponding relation. We can form a list that contains all those information and use it as the RE model’s training data. Figure 3 below shows an example of the relation triplet for a sentence that is properly tagged.

|  |
| --- |
|  |
| Figure 3. Example of relation triplets of a sentence. |

As for the joint model’s training data, we just added a ‘ner’ or ’re’ in front of the entity or relation samples, thus the model can be trained differently with respect to the tasks.

Final step for training data before going into the model is that we separate the data into train and validation with a ratio of 8:2 while maintaining consistent entity and relation distributions across splits. by doing that, our model can update a validation loss and can also be tested on performance before using the actual test data.

### Model Architecture Design

For our models, we decided to use the PyTorch (Paszke et al., 2019) framework with BiLSTM architecture. This is because the Bidirectional LSTM model has ability to handle the information flow both ways, this means that at each time step, the model has access to information not only from previous tokens but also from subsequent tokens. The bidirectional processing of BiLSTM allows them to capture long-range dependencies and relationships within the sequence. Comparing to the traditional LSTM structure, BiLSTM tends to make more informed prediction, thus lots of state-of-the-art results in NER, POS tagging and speech recognition uses this model architecture. Although other models such as BERT has shown its potential in recent years, due to the computational complexity and time efficiency to train that model, we still decided to use the BiLSTM structure.

The creation of model begins with first use BiLSTM to get contextual embeddings for each token. Then we get entity embeddings by averaging its constituting token embeddings. Then we concatenate head and tail embeddings. We make predictions based on the concatenate embedding. We want our model to find combinations that form constituents such that our model depends completely on the input sentence. The main frame of the model architecture design follows the tutorial that Robert Guthrie made with PyTorch of the BiLSTM-CRF model for NER task at PyTorch.[[2]](#footnote-2)

Continuing with our model, we have the embedding layer that capture the contextual information, now we are going to decide some of the choices in model training. And we have one hidden layer that captures and encodes sequential information for input data, since we are using the BiLSTM architecture, we have the hidden layer is structured to operate in forward and backward directions, as shown in the picture below (*Figure 4*). The hidden layer has two sets of LSTM units that computes a forward representation of ht for every word from left to right and a backward ht from right to left. Each LSTM unit maintains a memory cell, input gate, output gate, and forget gate. The memory cell retains information over long sequences, while the gates control the flow of information within the LSTM unit. A figure of a memory cell is shown below (*Figure 5*):

|  |  |
| --- | --- |
| Architecture of BiLSTM model | 图示, 示意图  描述已自动生成 |
| Figure 4. BiLSTM model overview | Figure 5. A Long Short-Term Memory Cell |

The forward function gets the embeddings of an entity or average the head and tail entity of a relation and computes the emission score of that entity or relation, based on that, we can predict the entity or relation. For loss function, we decided to use the Cross-Entropy loss to optimize both NER and RE tasks. Given the input text ***x*** and its gold tag sequence ***y***NER and tag table ***y***RE, we then calculate the following two losses shown in Figure 6, this will be used in the calculating the training loss and update the parameters.

|  |
| --- |
| 文本, 信件  描述已自动生成 |
| Figure 6. Loss function for calculating NER and RE training loss |

The goal is to minimize both losses LNER and LRE. And in the joint model, our goal would be to minimize the sum of both losses.

The parameter sharing strategy used in joint model entails sharing pre-trained word embeddings between the NER and RE models. This approach exploits the contextual information embedded within the embeddings to potentially improve both tasks simultaneously. The word embedding uses a word\_to\_ix dictionary that contains all the words appeared in the training and test corpus, are shared between the input layers of both NER and RE models.

Next, we implemented early stopping function such that our model reaches at a point where its validation loss stops decreasing for five continuous epochs, the model will stop further training.

### Training

During our model training phase, we chose Stochastic Gradient Descent (SGD) as our model optimizer to update our model’s parameters during training.

For the choices of some hyperparameters in the model, we did a hyperparameter tuning in the training phase. We recorded the baseline model performance measure for each run and compare the results for our final choice. For optimizer, we decided to use SGD optimizer with learning rate at 0.1 in all three models. For dimensions of the layers, we set hidden dimension at 64 and embedding dimension at 5 since change this does not make huge difference on the model performance. With this hyperparameter setting, our models have an average of 20 to 30 epochs to achieve best performance.

For training, we are using the sentences with gold standard entities and relations in the training data, NER model will generate each sentence’s entity prediction and RE model will generate each sentence’s relation based on the gold standard entity position, for the end-to-end task, the RE model will generate each sentence’s relation based on the NER model predicted entity position. For our parameter sharing model, it will generate individual performance measure of NER and RE tasks separately, and for end-to-end task, we will use the relations based on NER prediction’s entity position to test the relations.

Evaluation is necessary for each model, after each training, we note down the classification report of the model performance on validation data, this contains the precision, recall and f1-score of each entity and relation, as well as the overall micro-, macro-, and weighted-average value of the three different evaluation metrics. But in the final comparison between joint model and each baseline model, we will determine the model performance with weighted average f1-score. For NER performance measure, we use a special case of classification report designed by Nakayama[[3]](#footnote-3) that can evaluate the performance for named entity recognition or part-of-speech tagging.

End-to-end performance measure will be conducted with two different approaches. Since we have the NER prediction, we will form relations for each sentence with 8 other entities to form a relation with every ‘Drug’ predicted in the sentence, by doing that, some relations might match with the gold standard relations, and some might not. We categories those relations in three types:

1) True Positive (TP), which matches exact same entity positions as well as the relation type with the gold standard test data, we will use them in the end-to-end relation extraction test.

2) False Positive (FP), which are the relations that created from the entities that does not appear to have a match with the gold standard test data, we record them as FP.

3) False Negative (FN), which are the relations that should have been created from the entities as they are in the gold standard test data, but because of the errors and inaccurate prediction of entities, they are not created, we record them as FN.

Now, the model is properly tuned with training, we can do the testing on each model and record the test result for further analysis.

# Results

The results of our study demonstrate the impact of parameter sharing between Named Entity Recognition (NER) and Relation Extraction (RE) models in the context of n2c2 medical concept and relation extraction. We conducted a comprehensive evaluation using the n2c2 dataset and implemented various models to assess the performance improvement achieved through parameter sharing. Since we have three tasks, our results will be discussed in three parts:

1. Individual NER model performance comparing with parameter sharing model’s NER prediction.
2. Individual RE model performance comparing with parameter sharing model’s RE prediction.
3. Pipeline model performance which is using the relations that formed from the individual NER model prediction and tested on the RE model to compare with the parameter sharing model’s end-to-end performance.

## Named entity recognition performance

The result of Individual NER model prediction and parameter sharing model’s NER prediction are shown in the table 1 below, the score of each metrics is taken from the average of 5 runs under the best hyperparameters during training to avoid any accidents or extreme situations.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Individual Model | | | Parameter sharing model | | | | support |
|  | precision | recall | f1-score | precision | recall | | f1-score |
| ADE | *0.23* | *0.12* | *0.16* | *0.39* | *0.20* | *0.26* | | 625 |
| Dosage | 0.85 | 0.83 | 0.84 | 0.84 | 0.82 | 0.83 | | 2681 |
| Drug | 0.77 | 0.85 | 0.81 | 0.80 | 0.84 | 0.82 | | 10575 |
| Duration | *0.53* | *0.48* | *0.50* | *0.56* | *0.46* | *0.51* | | 378 |
| Form | 0.94 | 0.86 | 0.90 | 0.90 | 0.89 | 0.89 | | 4354 |
| Frequency | 0.78 | 0.74 | 0.76 | 0.80 | 0.77 | 0.78 | | 4012 |
| Reason | *0.53* | *0.54* | *0.53* | *0.49* | *0.56* | *0.52* | | 2545 |
| Route | 0.87 | 0.90 | 0.88 | 0.87 | 0.88 | 0.87 | | 3513 |
| Strength | 0.82 | 0.83 | 0.83 | 0.79 | 0.81 | 0.80 | | 4230 |
| micro avg | 0.79 | 0.80 | 0.79 | 0.79 | 0.80 | 0.79 | | 32913 |
| macro avg | 0.70 | 0.68 | 0.69 | 0.72 | 0.69 | 0.70 | | 32913 |
| weighted avg | **0.79** | **0.80** | **0.793** | **0.79** | **0.80** | **0.793** | | 32913 |
| Table 1: Precision, Recall, F1-score of named entity recognition prediction between Individual model and parameter sharing model.  ADE: adverse drug events. | | | | | | | | |

As we can see from the tables, the weighted average precision, recall, and f1-score of NER prediction does not have significant increase using our parameter sharing model, and each entity type has produced similar result which is shown in bold in the table. The shared knowledge and feature representations between NER and RE tasks did not lead to enhanced entity recognition capabilities. Out of 9 entities, there are three entities that perform poorly compare with the other entities: ‘ADE’, ‘Duration’, and ‘Reason’, as shown in Italian in the table. However, ‘ADE’ has gained quite amount of accuracy over the individual model, although the f1-scores for parameter sharing are not particularly high and still not reach the average of the overall score, it has produced a better result than the individual NER model prediction, we will talk more about it in the discussion section.

## Relation extraction performance

The result of Individual RE model prediction and parameter sharing model’s RE prediction are as the Table 2 below, the score of each metrics is taken from the average of 5 runs under the best hyperparameters during training to avoid any accidents or extreme situations.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Individual Model | | | Parameter sharing model | | | | support |
|  | precision | recall | f1-score | precision | recall | | f1-score |
| ADE-Drug | 0.74 | 0.57 | 0.64 | 0.69 | 0.56 | 0.62 | | 607 |
| Dosage-Drug | 0.96 | 0.93 | 0.94 | 0.96 | 0.97 | 0.96 | | 2657 |
| Duration-Drug | 0.95 | 0.82 | 0.88 | 0.95 | 0.88 | 0.91 | | 393 |
| Form-Drug | 0.96 | 0.96 | 0.96 | 0.97 | 0.97 | 0.97 | | 4304 |
| Frequency-Drug | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | | 3961 |
| Reason-Drug | 0.83 | 0.95 | 0.89 | 0.84 | 0.95 | 0.89 | | 2442 |
| Route-Drug | 0.97 | 0.95 | 0.96 | 0.98 | 0.95 | 0.97 | | 3503 |
| Strength-Drug | 0.96 | 0.95 | 0.96 | 0.98 | 0.96 | 0.97 | | 4211 |
| accuracy |  |  | 0.94 |  |  | 0.95 | | 22078 |
| macro avg | 0.92 | 0.89 | 0.90 | 0.92 | 0.90 | 0.91 | | 22078 |
| weighted avg | **0.94** | **0.94** | **0.942** | **0.95** | **0.95** | **0.952** | | 22078 |
| Table 2: Precision, Recall, F1-score of relation extraction prediction between Individual model and parameter sharing model.  ADE: adverse drug events. | | | | | | | | |

Unlike named entity recognition prediction, the results here indicate notable improvements in the relation extraction prediction performance when parameter sharing is employed in the joint model. By leveraging shared contextual information from NER, the joint model achieves a higher accuracy in identifying relationships between medical concepts. The precision and recall for Duration-Drug event relations, and other associations display promising enhancements under smaller test sample size, thus validating our hypothesis. From the table we can see that only ‘ADE-Drug’ relation has a slightly lower f1-score in parameter sharing model, all other relations achieved higher precision, recall, or f1-score, and the overall weighted average precision, recall and f1-score are boosted by 1.0%.

## End-to-end relation extraction performance

The result of pipeline model relation extraction prediction and parameter sharing model’s relation extraction prediction are shown in the Table 3 below, like the two tasks before, the score of each metrics is taken from the average of 5 runs under the best hyperparameters during training to avoid any accidents or extreme situations. Since the training data utilized were the relations were emerged from the predicted entities rather than the gold standard relations, the end-to-end relation classification job should perform worse than the conventional relation extraction task. Due to the fact that we use the anticipated entities as a filter to eliminate any unfavorable cases generated while building relations, our outcome in this case is superior to the regular relation extraction. By doing so, performance measure only keeps true positive results compared to the gold standard relations, increasing resulting f1-score. As we can see from the table 3 below, the performance of the model shows an absolute difference of 1.3% between the parameter sharing model and pipeline model, and specifically, the ‘ADE-Drug’ relation was largely improved from 0.80 to 0.93 in f1-score measure. All other relations have slightly improved as well.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pipeline Model | | | support | Parameter sharing model | | | support |
|  | precision | recall | f1-score | precision | recall | f1-score |
| ADE-Drug | 0.84 | 0.76 | 0.80 | 150 | 0.91 | 0.95 | 0.93 | 95 |
| Dosage-Drug | 0.99 | 0.97 | 0.98 | 1993 | 1.00 | 1.00 | 1.00 | 2040 |
| Duration-Drug | 0.99 | 0.94 | 0.96 | 215 | 0.99 | 0.99 | 0.99 | 153 |
| Form-Drug | 0.99 | 1.00 | 0.99 | 2961 | 1.00 | 1.00 | 1.00 | 3567 |
| Frequency-Drug | 0.99 | 0.99 | 0.99 | 2612 | 1.00 | 1.00 | 1.00 | 2672 |
| Reason-Drug | 0.96 | 0.98 | 0.97 | 1270 | 1.00 | 0.99 | 0.99 | 1240 |
| Route-Drug | 0.99 | 0.99 | 0.99 | 2700 | 1.00 | 1.00 | 1.00 | 2736 |
| Strength-Drug | 0.98 | 0.99 | 0.98 | 2961 | 1.00 | 1.00 | 1.00 | 3090 |
| accuracy |  |  | 0.99 | 15243 |  |  | 1.00 | 15593 |
| macro avg | 0.97 | 0.95 | 0.96 | 15243 | 0.99 | 0.99 | 0.99 | 15593 |
| weighted avg | **0.99** | **0.99** | **0.985** | **15243** | **1.00** | **1.00** | **0.998** | 15593 |
| Table 3: Precision, Recall, F1-score of end-to-end relation extraction prediction between pipeline model and parameter sharing model.  ADE: adverse drug events. | | | | | | | | |

We have mentioned that for the end-to-end task, we categorized the relations into true positives, false positives, and false negatives, and we only evaluate the true positives since our model cannot distinguish the false positives as a type of ‘no-relation’, this is the reason that results in the high values of precision, recall and f1-scores in the table above. The support values in the table shows the number of relation examples in the test relations, and for both models, it is showing a similar distribution, stating that the ability to predict entities have no significant difference. However, when we compare that with the gold standard test relations, which is 22078, both models have much reduced number of test examples, and the full table of distributions for relations formed from predicted entities are shown in the table 4 below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | TP | FP | FN | Total |
| Pipeline model | **15243**(28.0) | 32435(59.5) | 6835(12.5) | 54513 |
| Joint model | **15593**(25.4) | 39218(64.0) | 6485(10.6) | 61296 |
| Table 4. Distribution of true positives, false positives, false negatives in the total relations formed from entity prediction. Values are n(%). | | | | |

As we can see from the table, although the joint model forms more true positive relations and less false negative relations, both models have similar distribution of TPs, FPs, and FNs, which shows the ability of predicting the entities does not vary significantly.

# Discussion

## Strengths

Despite the unsignificant improvement in the named entity recognition result, our parameter sharing model still achieved higher f1-score in the entity ‘ADE’ by 0.1, which partially shows that the parameter sharing does effect some of the entity extraction to be more accurate.

The observed improvements in both RE and end-to-end performance attest to the underlying power of parameter sharing. This approach taps into the intrinsic connections between medical concepts and their relationships, leveraging shared features and contextual information. The enhanced RE performance can be attributed to the holistic view of the text that the joint model acquires, leading to more accurate recognition of relations between drug and other medical entities. This result aligns with findings in recent studies that underscore the importance of contextual and word embeddings in capturing the nuances of medical terminology and their relationships. By comparing our model result with state-of-the-art model, our model was second in recall, fifth in precision and third overall in f1-score. Since the gold standard relations were provided, the measures are conducted in strict matching.

From the table 5 below, it is clear that there is still a gap between our model results and those of state-of-the-art models, we will discuss that in the limitation section afterwards.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1-score |
| Our NER model | 0.79 | 0.80 | 0.793 |
| State-of-the-art NER | 0.8973 | 0.8939 | **0.8956** |
| Our RE model | 0.95 | 0.95 | 0.952 |
| State-of-the-art RE | 0.9715 | 0.9548 | **0.9630** |
| Table 5. Comparison between our NER and RE model result with current state-of-the-art model. | | | |

Furthermore, the end-to-end task benefits from the refined entity relations produced by the parameter sharing model. The joint model's ability to reason about relationships within the context of recognized entities results in more accurate relation predictions. This observation concurs with the notion that information extraction tasks, when interlinked, can leverage synergies to improve overall performance (Li et al., 2018).

The significance of end-to-end relation extraction in the pipeline model and joint model is particularly noteworthy. Although the parameter-shared joint model did not demonstrate superior performance in predicting named entities, but the complex relationships formed from the entity prediction has been improved. This achievement is not only a testament to the effectiveness of parameter sharing but also an affirmation of the interconnectedness of medical information within clinical narratives. The ability to coherently recognize entities and subsequently predict relationships contributes to a more comprehensive understanding of medical texts.

Furthermore, our result of end-to-end task is based on the true positive results after comparing our relations comparing the gold standard relations, thus when we compare it with the result of RE model prediction, it is showing that our relation classification result is improved by the parameter sharing model on the gold standard data. This improvement is another significant validation of the effectiveness of the joint learning approach, the joint learning process exposes the model to a broader range of context, allowing it to better grasp the nuances of medical terms and their interactions. Medical texts often involve intricate and domain-specific associations between entities. The shared feature representations in the parameter sharing model allow it to capture complex linguistic patterns such as adverse drug events, which is pivotal for accurate prediction of diverse relationships. The demonstrated improvement in relation classification underscores the real-world applicability of parameter sharing. In medical practice, accurate identification of relationships between medical concepts is essential for informed decision-making.

## Limitations

Like any research endeavor, this study has its limitations. One prominent limitation of this study pertains to the availability and diversity of datasets. The scope of i2b2 challenge datasets, while valuable, might not fully encapsulate the intricacies of clinical texts encountered in real-world medical scenarios. Real-world medical texts exhibit diverse linguistic patterns, medical specialties, and context-specific nuances that demand careful consideration. The limited size and specific domain focus can potentially hinder the generalizability of findings. Especially for some named entity such as ‘ADE’ and ‘Duration’, only have 959 and 592 training samples respectively, corresponding relations ‘ADE-Drug’ and ‘Duration-Drug’ suffers the same situation of having 1107 and 643 training samples respectively. In addition to having the fewest examples, Duration, Reason, and ADE make up 1%, 8%, and 2% of all concepts, respectively. thus, the model prediction on these three entities is relatively lower than other entities. This was mostly because of unnecessary predictions, but there was also some misunderstanding between ADEs and Reasons. This situation also shows the challenges faced by human annotators, where Durations, ADEs, and Reasons were the most challenging concepts to annotate, while Strength, Route, and Form were the easiest and most straight-forward to find. In fact, when we look into the gold standard annotations, the annotators for this dataset have explained its complexity in the 2018 n2c2 Track 2 paper. In the supplementary appendix provided by Henry et al. (2019), they stated that for these three medical entities, superfluous annotations were the main culprit; for example, low grade fever, lymphedema, viral, zoster px, febrile, cellulitis, bacterial, etc. were marked as ‘Reason’ in contexts when they weren't; difficulty swallowing secretions; emesis; mucositis; and severe nausea were frequently marked as ‘ADE’ in contexts where they weren't.

The model architecture has its own limitations as well. In n2c2 2018 Track 2 challenges, for concept extraction task, among the top 10 performance of the participated teams, 9 of them implemented Conditional random fields (CRFs) (Henry et al., 2019). Our BiLSTM model could perform better with CRF implementation, since CRF enhances sequence labeling performance and computes the joint probability of label sequences considering both the local features and the transition scores to achieve a better performance at in the global context. By incorporating CRF into the BiLSTM model, the combined architecture becomes more capable of capturing both local and long-range dependencies within sequences. Due to the limited time frame to develop the model and training, we were not able to implement that with our current model.

The relation classification task in 2018 n2c2 Track 2 on the other hand, various methods were proposed by different participants, including Random Forest (Chapman et al., 2018), Support Vector Machine (SVM) (Lafferty et al., 2001), a deep learning ensemble (Christopoulou et al., 2019), attention-based BiLSTM (Zeng et al., 2015) and CRF implemented BiLSTM (Xu et al., n.d.) or LSTM model (Sorokin & Iryna Gurevych, 2017). However, the BiLSTM model we used in relation extraction task was performing relatively outstanding compare with the best groups’ model. As we mentioned in strengths section, the result of relation extraction of our parameter sharing model can reach up to third in overall strict f1-score measure. Consider that we did not use the CRF layer, this result is quite special, and would perform even better with the implementation of CRF layer.

The approach we used to evaluate the end-to-end relation extraction is a compromission of the actual methodology to the end-to-end system evaluation. In end-to-end systems, we use the NER part of the pipeline or joint model to process raw narrative text and discover entities then find relations of those entities to their medications, such that we will form some relations that are not exactly the same as the ones in gold standard relations, and we classified them into false positives and false negatives, which we did not use them in the testing. The only relations we tested were the true positives when comparing our entity formed relation with the gold standard relations. Thus, the model produces such high accuracy. The optimal way to evaluate the end-to-end result is to test on all the relations that are formed, and for the relations that are not in the gold standard statistics, the model has the ability to classify them as no relation such that our end-to-end performance measure can be evaluated on behalf of the whole relation list generated and get a more reliable result.

The reported gains in relation classification by parameter sharing are encouraging, but there is a serious drawback caused by the test findings' lack of statistical significance that must be addressed. The lack of statistical significance raises concerns regarding the robustness and generalizability of the observed gains despite the model's performance appearing to have improved. An objective evaluation of the observed improvements might be possible by using statistical significance tests, such as t-tests or ANOVA. These tests would assist in establishing whether the performance differences between the parameter sharing model and the baseline model are statistically significant or merely coincidental. The most suitable mitigation we found, is the approximate randomization proposed by Noreen (1989), and were used in the n2c2 2018 Track 2 challenge to test the statistical significance between the systems, The null hypothesis argues that two systems will generate identical scores when compared. Predictions from both systems are gathered, jumbled, and redistributed in order to test this. An evaluation measure (such as precision, recall, or F score) is produced using the reassigned predictions, and the significance of the change in score is discovered. For every potential shuffle, this is exhaustively performed; for bigger datasets, it is approximated by a specified number of random shuffles. An assessment of the significance of the difference between the two methods is made using the count of shuffles that resulted in significant changes and the total number of shuffles compared. By using this method, we can tell the significance between our joint model with individual models or pipeline model, which can justify our result further together with the absolute difference value we used in the results.

## Future work

The promising results and insights gained from the parameter sharing approach pave the way for several compelling avenues of future research. Beyond the scope of the current study, exploring these directions can provide deeper insights into the potential of joint learning for medical concept and relation extraction. One particularly valuable area of investigation is conducting an error analysis to unravel the intricacies of the model's performance in predicting Named Entity Recognition and Relation Extraction.

### Error Analysis for NER

An in-depth error analysis for NER predictions could shed light on the model's strengths and limitations in identifying medical entities. By meticulously examining instances where the model succeeds and fails, researchers can gain a better understanding of the challenges posed by different types of medical terms, linguistic variations, and ambiguous contexts. This analysis could lead to valuable insights into the specific types of entities that benefit most from parameter sharing.

We could investigate cases where the model's NER predictions align with the gold standard annotations and those where discrepancies arise. By categorizing errors based on entity types, complexity, and linguistic context, patterns of misclassification could be identified. For example, if we can find for each entity, how many instances have our model mislabel and unable to predict the correct gold standard entity. Such patterns can guide the refinement of the model, suggesting potential areas for fine-tuning shared embeddings and contextual features. This level of granularity in error analysis could help bridge the gap between current performance and optimal accuracy. Ratinov (2009) has analyzed issues in misconceptions that underlie the development of an NER system, our model could use some of the methods he proposed to find reasons behind the missingness patterns.

### Error Analysis for RE

Similarly, an error analysis for relation extraction can illuminate the nuances of the model's predictions in capturing complex relationships between medical entities. Insights about the model's understanding of various sorts of relationships may be gained by examining situations in which the model correctly recognizes and categorizes relations as well as those in which it incorrectly does so. The situations covered by this research may involve drug-drug interactions, drug-disease correlations, and other clinically important linkages.

It is possible to further categorize error patterns in relation categorization based on things like entity proximity, context ambiguity, and the function of various entities inside relationships. Researchers can gain a more sophisticated knowledge of the sorts of textual signals that impact correct connection prediction by analyzing the causes for the model's successes and failures. This understanding could guide the refinement of shared feature representations and contextual embeddings to better capture the intricacies of medical relationships.

Overall, the error analysis for both NER and RE task holds great potential in guiding the development of parameter sharing models for medical concept and relation extraction, the analyzed result may also apply to end-to-end tasks with the same approach. By uncovering the specific challenges and intricacies in NER and RE tasks, researchers can refine the shared parameter approach to target the areas where improvements are most needed. This iterative process of analysis, refinement, and validation could ultimately lead to more accurate and robust models that excel in understanding complex medical texts.

# Conclusion

The pursuit of accurate and efficient information extraction from medical texts has garnered significant attention in the realm of natural language processing. Important elements of this project are named entity recognition (NER) and relation extraction (RE), which make it easier to extract medical ideas and their intricate interactions from clinical narratives. This paper embarked on a comprehensive exploration of the potential of parameter sharing between NER and RE models for enhancing medical concept and relation extraction, particularly within the challenging landscape of i2b2 medical discharge summaries.

A rigorous examination of the advantages, drawbacks, and potential applications of the parameter sharing technique was the hallmark of the trip taken in this work. A detailed literature review was used to lay the groundwork, which emphasized the development of n2c2, NER, RE, BiLSTM and joint learning approaches in NLP. This backdrop emphasized the need of examining the interaction between NER and RE, which led to the creation of the research questions and hypotheses that determined the course of the study.

The methodology section provided an overview of the data source, preprocessing, and model building phases that led to the construction of separate NER and RE models as well as their combined counterpart. Thorough training and testing and analysis that followed provided deep understandings of the performance gains made possible by parameter sharing. The outcomes demonstrated the approach's potential by demonstrating enhancements in individual RE and end-to-end relation extraction tasks.

For named entity recognition, our parameter sharing model achieved a strict micro-averaged f1-score of 0.793 using the BiLSTM structure which is similar to the state-of-the-art systems approach (BiLSTM-CRFs) that incorporate additional features. The result of named entity recognition task did not show the improvement for our parameter sharing model, we discussed several limitations that may cause the unsignificant result, including model architecture, feature embeddings, and dataset diversity.

For relation extraction task, our parameter sharing model has reached a strict micro-averaged f1-score of 0.952 using the BiLSTM structure, this result has 1.0% absolute difference to the individual relation extraction model. As for the end-to-end relation classification task, our parameter sharing model got 0.997 in predicting the true positive relations which is 1.3% higher than the pipeline model. Both results have contributed to justify our research hypothesis. The improvements observed in relation classification and end-to-end task validated the approach's capacity to translate shared knowledge and contextual embeddings into tangible enhancements in medical text analysis.

The discussion section delved into the strengths of parameter sharing, revealing its potential to revolutionize medical concept and relation extraction. The parameter sharing approach's ability to capture contextual cohesion and enhance data efficiency presented compelling evidence of its efficacy. We also acknowledged the limitations of the current study, particularly the absence of statistical significance in certain test results. Other limitations such as diversity of dataset, model architecture are also taken into account. These limitations highlighted the complexity of evaluating model improvements and prompted discussions on potential mitigations and future avenues for exploration.

Parameter sharing's potential was investigated, and this revealed some exciting new paths for future study and development. A thorough error analysis to analyze the intricacies of NER and RE predictions is one major step forward. This line of inquiry has the potential to reveal the model's thought processes, points of strength, and points of weakness in terms of accurately representing medical entities and connections. The knowledge gathered through error analysis may help to improve contextual embeddings, shared feature representations, and model training techniques, ultimately leading to more precise and contextually aware information extraction systems.

The route followed in this study, in sum, highlights the value of multidisciplinary cooperation between NLP and healthcare. The potential to use contextual word embeddings and shared information to transform how people read medical texts increases as technology develops. The advantages of parameter sharing between NER and RE model upgrades fan the flame of creativity, encouraging more investigation, validation, and development of collaborative learning strategies. By advancing healthcare delivery, research, and patient-centered decision support, parameter sharing has the potential to transform the landscape of medical information extraction with ongoing research and ethical issues at the forefront.

# References

1. [Wang, 2021] Wang, S. (2021). The survey of joint entity and relation extraction. In *Computing and Data Science: Third International Conference, CONF-CDS 2021, Virtual Event, August 12-17, 2021, Proceedings 3* (pp. 363-381). Springer Singapore.
2. [Henry et al., 2019] Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, *27*(1), 3-12.
3. [Alshammari & Alanazi, 2021] Alshammari, N., & Alanazi, S. (2021). The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, *22*(3), 295-302.
4. [Nasar et al., 2021] Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, *54*(1), 1-39.
5. [Daniel & Martin, 2023] Daniel, J., & Martin, J. (2023). *Speech and Language Processing Relation and Event Extraction*. https://web.stanford.edu/~jurafsky/slp3/21.pdf
6. [Kartik Detroja et al., 2023] Detroja, K., Bhensdadia, C. K., & Bhatt, B. S. (2023). A Survey on Relation Extraction. *Intelligent Systems with Applications*, 200244.
7. [Riloff, 1996] Riloff, E. (1993, July). Automatically constructing a dictionary for information extraction tasks. In *AAAI* (Vol. 1, No. 1, pp. 2-1).
8. [Etzioni et al., 2008] Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, *51*(12), 68-74.
9. [Alfred et al., 2014] Alfred, R., Leong, L. C., On, C. K., & Anthony, P. (2014). Malay Named Entity Recognition Based on Rule-Based Approach. *International Journal of Machine Learning and Computing*, *4*(3), 300–306.
10. [Wu et al., 2022] Wu, L. T., Lin, J. R., Leng, S., Li, J. L., & Hu, Z. Z. (2022). Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. *Automation in Construction*, *135*, 104108.
11. [Sharnagat, 2014] Sharnagat, R. (2014). Named entity recognition: A literature survey. *Center For Indian Language Technology*, 1-27.
12. [Liu et al., 2013] Liu, C., Sun, W., Chao, W., & Che, W. (2013, December). Convolution neural network for relation extraction. In *International conference on advanced data mining and applications* (pp. 231-242). Berlin, Heidelberg: Springer Berlin Heidelberg.
13. [Zhang & Wang, 2015] Zhang, D., & Wang, D. (2015). Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.
14. [Gasmi et al., 2019] Gasmi, H., Laval, J., & Bouras, A. (2019). Information extraction of cybersecurity concepts: An LSTM approach. *Applied Sciences*, *9*(19), 3945.
15. [M.Abdelgwad et al., 2021] Abdelgwad, M. M., Soliman, T. H. A., Taloba, A. I., & Farghaly, M. F. (2022). Arabic aspect based sentiment analysis using bidirectional GRU based models. *Journal of King Saud University-Computer and Information Sciences*, *34*(9), 6652-6662.
16. [Harnoune et al., 2021] Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z., & El Asri, B. (2021). BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, *1*, 100042.
17. [Gopalakrishnan et al., 2019] Gopalakrishnan, A., Soman, K. P., & Premjith, B. (2019, July). Part-of-speech tagger for biomedical domain using deep neural network architecture. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.
18. [Zhao et al., 2018] Zhao, W., Guan, Z., Chen, L., He, X., Cai, D., Wang, B., & Wang, Q. (2017). Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *30*(1), 185-197.
19. [Gehring et al., 2016] Gehring, J., Auli, M., Grangier, D., & Dauphin, Y. N. (2016). A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.
20. [Kohane, n.d.] Kohane, I. (n.d.). *About n2c2*. N2c2.Dbmi.hms.harvard.edu. <https://n2c2.dbmi.hms.harvard.edu/about-n2c2>.
21. [Uzuner et al., 2008] Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, *15*(1), 14-24.
22. [Oleynik et al., 2019] Oleynik, M., Kugic, A., Kasáč, Z., & Kreuzthaler, M. (2019). Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association*, *26*(11), 1247-1254.
23. [Guan & Devarakonda, 2020] Guan, H., & Devarakonda, M. (2019). Leveraging contextual information in extracting long distance relations from clinical notes. In *AMIA Annual Symposium Proceedings* (Vol. 2019, p. 1051). American Medical Informatics Association.
24. [Alimova & Tutubalina, 2020] Alimova, I., & Tutubalina, E. (2020). Multiple features for clinical relation extraction: A machine learning approach. *Journal of biomedical informatics*, *103*, 103382.
25. [Kusalavan, 2023) Kusalavan, K. (2023). Named entity recognition in the medical domain.
26. [Catelli et al., 2021] Catelli, R., Casola, V., De Pietro, G., Fujita, H., & Esposito, M. (2021). Combining contextualized word representation and sub-document level analysis through Bi-LSTM+ CRF architecture for clinical de-identification. *Knowledge-Based Systems*, *213*, 106649.
27. [Stubbs et al., 2019] Stubbs, A., Filannino, M., Soysal, E., Henry, S., & Uzuner, Ö. (2019). Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, *26*(11), 1163-1171.
28. [Daniel & Martin, 2023] Daniel, J., & Martin, J. (2023). *Speech and Language Processing Sequence Processing with Recurrent Networks.* <https://web.stanford.edu/~jurafsky/slp3/9.pdf>
29. [Li et al., 2017] Li, F., Zhang, M., Fu, G., & Ji, D. (2017). A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, *18*(1), 1-11.
30. [Ma & Hovy, 2016] Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.
31. [Luo et al., 2017] Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, *34*(8), 1381-1388.
32. [Graves et al., 2013] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). Ieee.
33. [Devlin et al., 2018] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
34. [Chen et al., 2020] Chen, M., Lan, G., Du, F., & Lobanov, V. (2020, November). Joint learning with pre-trained transformer on named entity recognition and relation extraction tasks for clinical analytics. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop* (pp. 234-242).
35. [Liu et al., 2019] Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.
36. [Li et al., 2018] Li, F., Liu, W., & Yu, H. (2018). Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. *JMIR medical informatics*, *6*(4), e12159.
37. [Ruder, 2017] Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
38. [Miwa & Bansal, 2016] Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.
39. [Johnson et al., 2016] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*(1), 1-9.
40. [Ramshaw & Marcus, 1999] Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157-176). Dordrecht: Springer Netherlands.
41. [Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.
42. [Xu et al., n.d.] Xu, J., Lee, H. J., Ji, Z., Wang, J., Wei, Q., & Xu, H. (2017, November). UTH\_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017. In *TAC*.
43. [Chapman et al., 2018) Chapman, A. B., Peterson, K. S., Alba, P. R., DuVall, S. L., & Patterson, O. V. (2018, May). Hybrid system for adverse drug event detection. In *International Workshop on Medication and Adverse Drug Event Detection* (pp. 16-24). PMLR.
44. [Lafferty et al., 2001] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
45. [Christopoulou et al., 2019] Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., & Ananiadou, S. (2020). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, *27*(1), 39-46.
46. [Zeng et al., 2015) Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015, September). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1753-1762).
47. [Sorokin & Iryna Gurevych, 2017] Sorokin, D., & Gurevych, I. (2017, September). Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1784-1789).
48. [Noreen, 1989] Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses* (pp. 20-41). New York: Wiley.
49. [Ratinov & Roth, 2009] Ratinov, L., & Roth, D. (2009, June). Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)* (pp. 147-155).

1. <https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/> [↑](#footnote-ref-1)
2. <https://pytorch.org/tutorials/beginner/nlp/advanced_tutorial.html> [↑](#footnote-ref-2)
3. Nakayama, H. GitHub. https://github.com/chakki-works/seqeval [↑](#footnote-ref-3)