Does parameter sharing between NER and RE models improve model performance for i2b2 medical concept and relation extraction?

Qingzhou Li

February 16, 2023

1 Introduction

Medical text information extraction is a crucial task in the field of medical informatics. Named entity recognition (NER) and relation extraction (RE) are two key tasks in information extraction, where NER identifies named entities such as diseases, symptoms, and treatments, while RE identifies the relationships between these entities. The i2b2 (Informatics for Integrating Biology and the Bedside) challenge is a well-known benchmark for evaluating NER and RE models on medical text.

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 [DMa]. 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 [DMb]. In this project, we will be using these two approaches to evaluate performance between models build based on them.

In general, people used the pipeline method for NER and RE, which cannot take full advantage of the benefits in between [Wan21]. By using sharing representations between related tasks, we can improve our model based on the original one, this is called multi-task learning (MTL) [Rud]. 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 [HBF+19]. This may be improved by sharing parameters between the NER and RE models. For example, Fei Li believes that parameter sharing between the subtasks of a joint model is effective since they are influenced by correlated subtasks [LZFJ17]. However, it is still not well established how parameter sharing between NER and RE models impacts performance for the i2b2 medical concept and relation extraction task.

The purpose of this study is to explore the impact of parameter sharing between NER and RE models on performance for the i2b2 medical concept and relation extraction task. This study is important because it will provide insights into how parameter sharing can improve the performance of NER and RE models on medical text. Additionally, this study will also determine if parameter sharing can improve the computational efficiency of NER and RE models, which is critical in real-world applications.

Study Aims:

Evaluate the performance of NER and RE models trained with and without parameter sharing for the i2b2 medical concept and relation extraction task. Compare the computational efficiency of NER and RE models trained with and without parameter sharing for the i2b2 medical concept and relation extraction task.

Specific Procedures and Materials:

To achieve our study aims, we will use the i2b2 2018 Track 2 challenge corpus, which contains the annotated medical text for NER and RE tasks. We will train NER and RE models with and without parameter sharing using deep learning techniques such as long short-term memory (LSTM). We will evaluate the performance of the models using metrics such as F1 score and computational efficiency using a runtime and memory usage.

Hypothesis/es:
We hypothesize that the joint NER and RE model with shared parameters will achieve a higher F1 score compared to models trained on NER and RE tasks independently. We also hypothesize that the joint NER and RE model with shared parameters will be more computationally efficient compared to models trained on NER and RE tasks independently.

2 Methodological Considerations

The data for this study will be sourced from the i2b2 challenge, a widely used benchmark for evaluating NER and RE models on medical text. The challenge includes annotated medical records from a variety of sources, mostly are electronic health records (EHRs). The data was originally produced by healthcare providers and has been annotated for NER and RE tasks.

The criteria for selecting material for this study will be based on the availability of annotated data for both NER and RE tasks and the relevance of the data to the medical domain. We will consider a date range of the last 5 years to ensure that the data is recent and relevant. No secondary data is required for this study, and a research passport is not needed as the data is publicly available through the i2b2 2018 Track 2.

One potential issue regarding the reliability and validity of large data sets is the possibility of missing or inconsistent data. To address this, we will perform data cleaning and pre-processing to ensure that the data is consistent and reliable. This will involve removing any irrelevant or duplicate data, correcting any inaccuracies, and transforming the data into a format suitable for analysis.

In terms of research design, we will use a controlled experiment to evaluate the impact of parameter sharing between NER and RE models on performance and efficiency for the i2b2 medical concept and relation extraction task. We will use a pre-existing NER and RE model as a baseline and compare the performance of this model with a modified version that includes parameter sharing.

The methodology for this study will be based on machine learning and deep learning techniques. Recently, many neural networks have been used in NER, for example CNN and RNN. The most common and useful model now is LSTM-CRF [HXY15]. We will use a long short-term memory (LSTMs) method [MB] to develop the NER and RE models. The models will be trained on annotated data from the i2b2 challenge and evaluated using common metrics such as precision, recall, and F1 score.

In terms of performance checking, we will use a cross-validation technique to evaluate the performance of the models and ensure that they generalize well to new data. We will also compare the performance of the models on a test set of annotated data from the i2b2 challenge to provide an objective evaluation of the models.

The data analysis for this study will be performed using Python programming language and relevant libraries such as Pandas and nltk. We will use these tools firstly to perform data pre-processing and cleaning and then to train and evaluate the NER and RE models.

In conclusion, this study will provide valuable insights into the use of parameter sharing for NER and RE models on medical text and advance our understanding of the field. The methodology and models described in this proposal will ensure that the results of the study are reliable and valid, and the use of Python and relevant libraries will allow us to perform the data analysis in an efficient and effective manner.

3 Timetable/Schedule

References

[DMa] Jurafsky Daniel and James Martin. *Speech and Language Processing Relation and Event Extraction.*

[DMb] Jurafsky Daniel and James Martin. *Speech and Language Processing Relation and Event Extraction.*

DATE	TASK
November - January	Students should meet with project supervisors
November - January	Data Requisition Requests: Requisitioning access to data and software.
January	Literature exploration and self-learning
February	Research proposal write-up
17 February 2023	**Research Proposal Deadline**
February - March	Literature exploration and self-learning
March - April	Data exploration and data pre-processing
May - June	Independent NER and RE model building
June - July	Parameter sharing model building
July - August	Performance comparison and writing up
August	Poster Creation and Printing AO
22 August 2023	**Poster Submission**
24 August 2023	**Live Poster Presentation**
Late August	**Final Write up and proof-reading**
1 September 2023	**Deadline for Submission**

[HBF+19] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 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, 10 2019.

[HXY15] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv:1508.01991 [cs]*, Aug 2015.

[LZFJ17] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1), Mar 2017.

[MB] Makoto Miwa and Mohit Bansal. *End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures.*

[Rud] Sebastian Ruder. *An Overview of Multi-Task Learning in Deep Neural Networks *.*

[Wan21] Sai Wang. The survey of joint entity and relation extraction. In Weijia Cao, Aydogan Ozcan, Haidong Xie, and Bei Guan, editors, *Computing and Data Science*, pages 363–381, Singapore, 2021. Springer Nature Singapore.