END-TO-END RELATION EXTRACTION USING SEMI-SUPERVISED PRE-TRAINING

A Thesis

by

PRANOY KOVURI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Bobak Mortazavi
Committee Members,	Ruihong Huang
	Xiaoning Qian
Head of Department,	Dilma Da Silva

August 2019

Major Subject: Computer Science

ABSTRACT

Information extraction (IE) extracts meaningful knowledge from data. Two important tasks in IE are named entity recognition and relation extraction. Existing approaches in relation extraction treat entity and relation extraction as two separate tasks. They model them in a pipeline approach and rely on external linguistic resources to improve the performance. On contrary, we design a generalized system for end-to-end relation extraction without utilizing any external resources. Our approach identifies entities and relations jointly using a single model, and concurrently identifying all relations between all predicted entities. Through this work, we introduce multi-task fine-tuning on pre-trained models as an approach for related tasks and show that it gives significant performance improvements for each of the individual tasks. Our model performs comparably to the state of the art on Biocreative V Chemical Disease Relation corpus in detecting chemical and diseases and chemically induced disease relation F1-score. We outperform the existing state of the art results on nominal relation classification for SemEval-2010 Task 8 by Test F1 86.9 (2.2 point absolute improvement), without incorporating any external resources or tools.

Better information extraction techniques can help identify patient risks more efficiently and thus will be helpful in patient care. Clinical notes are crucial for predicting events during a patient stay in hospital since they contain valuable information which correlates with the event occurrence. Hence, we study identifying Intensive care unit (ICU) readmission risks using clinical notes for heart disease patients, considering different subsets of these notes but focusing on Echocardiography notes. This work builds a representation of the clinical notes and accounts for additional modality including time series based vital data and different patient descriptors. We outperform previous work on predicting ICU readmission clinical event measured by AUROC (0.634) and F1-score (0.73) without textual modality (Baseline - 0.62 AUROC and 0.72 F1). Additionally, we give the clinician a way of visual interpretation of the important text for the model prediction using attention scores.

DEDICATION

To my Parents and my niece Siri.

ACKNOWLEDGMENTS

I would like to express my gratitude to Professor Bobak Mortazavi my research supervisor, for his guidance and useful discussions during this research work. He always believed in me and taught me the essentials of research work. I would also like to thank Dr. Ruihong Huang, for her advice and assistance in innovative ideas and novelty discussions. I would also like to thank Dr. Xiaoning Qian for serving on my committee and for his constant support. My deepest grateful thanks are extended to Mr. Prafulla Kumar Choubey for his help during the research, to Mr. Satya Kesav for interesting discussions, which helped me understand complex learning algorithms. I wish to thank my parents for their constant support and encouragement throughout my study. Specifically, I would like to thank my mother for always believing in me. Finally, I would also like to extend my thanks to all members of the laboratory and friends for their helping me in tough times.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Bobak Mortazavi [advisor] and Professor Ruihong Huang of the Department of Computer Science and Engineering and Professor Xiaoning Qian of the Department of Department of Electrical and Computer Engineering.

Funding Sources

Graduate study was supported by teaching and research assistantship from the Department of Computer Science and Engineering at Texas A&M University.

NOMENCLATURE

IE	Information Extraction
NER	Named Entity Recognition
NLP	Natural Language Processing
RE	Relation Extraction
EERE	End-to-end relation extraction
EHR	Electronic Health Record
CID	Chemically induced Disease
CDR	Biocreative V Chemical Disease Relation dataset
SemEval	SemEval-2010 Task 8
MIMIC-III	Medical Information Mart for Intensive Care
ICU	Intensive care unit
BPE	Byte pair encoding
Echo	Echocardiogram
GPT	Generative Pre-trained Transformer
BERT	Bidirectional Encoder Representations from Transformers

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Information extraction is the task of extracting meaningful information or knowledge in machine-readable form [1]. The sources of extraction can vary from text in news to medical notes. The type of text can be unstructured (e.g. news journal), semi-structured (e.g. medical notes) or semi-structured machine-readable documents (e.g. Electronic Health Record). Information extraction plays a crucial role in language understanding [2]. The structured information could consist of Entities, relationships between them and the events they are involved in. For instance "Golden Gate Bridge is located in San Francisco" can be represented using the tuple $Located\_In(GoldenGateBridge, SanFrancisco)$. Here the $Located\_In$ is a relation which relates the Entities $GoldenGateBridge$ and $SanFrancisco$. Tuples of length greater than two can also be used to represent a relation present between multiple entities. The data extracted from the text is then stored in a knowledge base. This knowledge base can be further processed for understanding and gaining insights and making challenging decisions.

There are many applications of Information Extraction (IE) in Natural Language Processing (NLP) like question answering, event extraction, etc. This is especially needed in the Medical field since we can identify potential causes of various diseases. For example, IE on MRI reports can reveal various insights to the doctors about prevalence of diseases in body anatomy locations, which otherwise would have been hard to discover [3].

Named entity recognition (NER) and relation extraction (RE) are used ubiquitously. Some well-known applications of NER and RE are Ontology creation [4], document summarization [5], biological or biomedical information extraction [6], question answering [7], and knowledge base population [8].

In this work, we focus on the task of end-to-end relation extraction (EERE). Rephrasing Pawar *et al.* [9], end-to-end relation extraction means identifying boundaries of entities, identifying their types and appropriate semantic relation for each pair of these identified entities. This task has been extensively studied in the past. Early works divided this task as a pipeline of two tasks, entity

recognition and identifying relations among them separately using pipelined models [10, 11]. Both of these tasks are complementary to each other. Entities help decide which relations are important and vice versa. One example can be seen in Figure 1.1, if IBM and New York are given as entities being related, their information constrains the problem for the relations being classified. In this regard, recent works have explored end-to-end joint modeling [9, 12]. This was shown to be vital for better performance - F1 scores of both NER and RE [13, 14]. One common way of joint modeling is by parameter sharing. Hence we use a single model for extracting both named entities and relations training in an end-to-end manner. We perform extensive ablation studies for this multi-task training and summarize our observations.

Figure 1.1: Three entities *IBM*, *New York* and *June 16, 1911* are related by two relations *Founding-loc* and *Founding-year* respectively

Early feature based works and recent works using end-to-end neural models use external linguistic resources to obtain better performance. State of the art works in relation extraction also use linguistic structure-based features [15, 16] and external knowledge based information [17]. These external linguistic resources are costly to achieve in low resource domains like biomedical and less studied languages like Hindi. Furthermore obtaining these features at the document level for relation classification needs manual crafting of effective features [17], often requiring additional tools. Obtaining these features can be costly while decoding and would add as a preprocessing step every time we want to extract relation triples. This preprocessing acts as a barrier entailing more computational costs than systems that process the text directly. Hence we use raw sentences as inputs to our model without using any external resources.

Better feature extraction techniques help good classification of relations, and as stated above it might be difficult to use external linguistic resources for feature extraction. Hence we use archi-

tectures pre-trained on language modeling objective. This is a comparatively simpler technique to obtain in low resource domains since free text is a ubiquitous resource. These pre-trained architectures help learn low-level features efficiently and help in training target tasks by providing better initialization points. Accordingly, this circumvents linguistic resources usage. In this work, we use only raw words as inputs for a single joint model that modifies a pre-trained architecture. More specifically, we amend Generative Pre-trained Transformer (GPT) [18] and Bidirectional Encoder Representations from Transformers (BERT) [19] architectures in two different ways for relation extraction. Section 3 gives the essentials of the design.

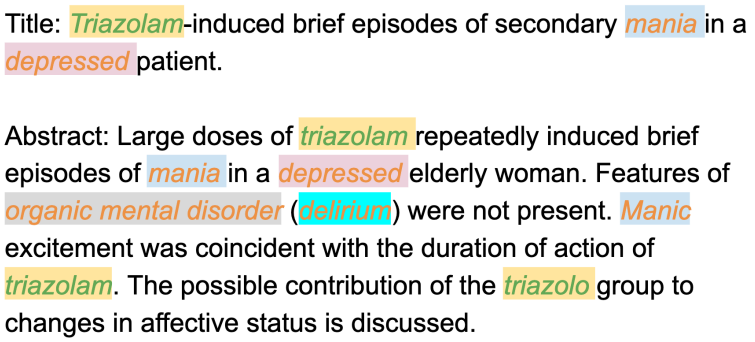

Figure 1.2: Example abstract from the CDR dataset. The green font is chemicals and the orange font is diseases. Each entity mention is highlighted in the same color. The entities being related can be in different sentences and can have multiple mentions with different mention form. Here the entity Manic is of type disease which is induced from chemical entity Triazolam.

In this work, we primarily address chemical-induced disease (CID) relation extraction. Figure 1.2 shows a typical example of a PubMed abstract annotated with entity mentions of chemicals and diseases in the Biocreative V CDR (CDR) dataset, here entity *Manic* is positively related to the entity *Triazolam* [20, 21]. These relations are important in drug discovery, biocuration, drug safety etc [21]. Manual annotating these abstracts is limited by human capabilities and hence automated extraction techniques are needed. CDR PubMed abstracts are annotated at the document level and the relations are expressed across entities in different sentences. The entities can have multiple mentions and this requires global information and understanding across the whole docu-

ment. However, since all the mentions are given as inputs, entity linking is not necessary in this case. With few exceptions most existing research in relation extraction addresses entity pair with a single mention and focus on intrasentence relation extraction, relying on local surface features of entity mentions. Hence, we address these shortcomings in our approach using self-attention for relating long distant entities with mentions across multiple sentences. We use max pooling to aggregate information from all entity mentions in the document. Moreover, we extract all relations in a document concurrently between any pair of extracted entities, unlike any existing work. Furthermore, we work on nominal relation classification for SemEval-2010 Task 8 (SemEval) [22], we somewhat alter our methodology to account for given entity information, we elaborate the subtleties in Section 3.4.2.

This work addresses all the aforementioned difficulties of earlier techniques for the task of end-to-end relation extraction. We model entities and relations using a joint model. We augment pre-trained architectures based on self-attention with a relation extraction head without using any external resources and only relying on the raw sentences as inputs. For RE head, we modify the pre-trained architecture using a biaffine function to model the interactions between head and tail representations. We use max pooling over masked tokens with entities participating in the relation triple being extracted for each relation candidate, this is based on a linguistic observation to pick the most dominant interaction. We experiment model pre-training based on GPT and BERT [18, 19] and select the empirically best-performing architecture. We justify the better performance of GPT in Section 4. In this work, we follow encode once extract multiple strategy, thus extracting all relations in a document at the same time, we train and decode our model in the same way. We select self-attention based pre-trained architectures for their better empirical performance in other tasks [18, 19] and also for learning long distant relations.

We show that our approach performs comparative to the baseline on CDR for chemically induced diseases task using a much simpler model. We outperform state of the art on nominal relation classification task on SemEval. We note that we do not use any ensemble models, any external resources, or additional weakly labeled data for training, unlike existing works. Additionally, we

give an elaborative ablation summary with observations for fine-tuning pre-trained architectures over cross domains tasks and will open source the codebase. Through this work, we introduce multi-task fine-tuning as an effective approach.

As previously mentioned, large scale labeled datasets are difficult to obtain in many domains like biomedical, clinical, health care, etc. The two most obvious reasons are the cost of experts advice needed for labeling and lack of resources for a language. For example, in medicine, it is hard to label a large set of radiology reports with the diagnosis of the disease since it is hard to find expert radiologists [23]. To demonstrate the effectiveness of our work, we apply our technique in a realistic scenario of prediction of a clinical event based on information from Electronic Health Records (EHR). Because of the lack of labeled data, typical models with huge number of parameters cannot be trained. Hence, this work primarily investigates model pre-training based on the language model objective. For this task, we focus on heart patients with an echocardiogram (Echo) or ultra-sound. We represent an Echo note associated with a patient using the above architectures and predict the probability of a patient being readmitted to the Intensive care unit (ICU). This research is performed upon the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database [24]. We use CNN [25] and LSTM [26] based models as baselines for our model's performance.

The article is structured as follows. Chapter 2 reviews related work about relation extraction, document level relation extraction, pre-trained architectures and attention in relation extraction. Chapter 3 first briefly explains joint model and then discusses our approach for all three tasks. In Chapter 4 we discuss about the data and results of our approach on all three tasks. This article then summarizes and concludes. Future research direction is mentioned at the end.

2. LITERATURE SURVEY

There is a great deal of work in the existing literature in relation extraction, and in this chapter we briefly go over the broadly used strategies.

2.1 End-to-end relation extraction works

Early works in EERE focused on feature extraction techniques and treated EERE as two pipelined tasks. Recent works have explored neural methods [12, 27, 28], nevertheless do not forgo feature extraction based methods. A lot of works have explored neural methods for features extraction from dependency parse trees and use these features to classify relations. For this Xu *et al.* [27] first used shortest paths between two entities head words in the dependency parse tree as a sequence. They modelled this sequence using a LSTM [26] to classify relations. On similar lines Santos *et al.* use CNN based approach [29]. The tree-based information has been encoded in a number of different ways. Miwa and Bansal [12] had success modelling the dependency tree using a bi-directional Tree LSTM. They also explored other possible information including subtree of the lowest common ancestor.

State of the art works have explored graph based approaches on dependency parse trees. Zhang *et al.* propose an extension of graph convolution networks for relation extraction [16]. They use a strategy based on shortest paths, in dependency parse trees, for pruning. Christopoulou *et al.* employ a walk-based approach for simultaneously treating multiple pairs in a sentence, modelling interactions among the pairs [15]. Most of these works use external linguistic resources by means of feature extraction tools, linguistic knowledge or additional weakly labelled data. As this leads to additional overhead, we only use document abstract for CDR and sentence for SemEval. Li *et al.* use self-attention architecture for relation extraction [30], we differ from them in subsequent ways. The authors model named entities and relations among them using different models, i.e. they use a cascaded LSTM and attention-based model. Our model is a joint architecture in which we use hard parameter sharing based on a pre-trained model.

2.2 Document Level works and Bio works

Most works in current literature are focused on short sentences relation extraction, only few works are about extracting relations from entities for their mentions across sentences or in an abstract [17, 31, 32]. These works are based on syntactic parse features classifiers and Peng *et al.* [32] use a graph LSTM over dependency tree. This work uses a simple model without using any LSTM, CNN and any features based on syntactic information to extract CID from a medical abstract. Peng *et al.* [17] use rich features based on linguistic and domain knowledge for designing a support vector machine based classifier for CID. Our work motivates from the work of [33], which uses a self-attention and convolution based model to encode a biological abstract. We adopt their model of jointly predicting named entities and relations between them, but our work differs in the following key ways. Firstly, we use architectures solely based on self-attention architecture and also model pre-training schemes on language model objective. Additionally, our model is simple by using max-pool for all mentions and predicts all relations in the document concurrently.

2.3 Model Pre-training

There are many works in existing literature about unsupervised and semi-supervised learning and it has been a long-studied topic in NLP. The famous methods are the word embedding models including Glove [34] and Word2Vec [35]. The most recent methods are based on language models [18, 19, 36] and learning contextual embeddings of words [37]. The state of the art methods learn a surrogate objective function based on language modelling and use transfer learning schemes to fine tune the model for predicting the specific task. Dai and Le first introduced a language modelling objective as a semi-supervised approach for model fine-tuning [38]. The authors design a model based on sequential learning using recurrent neural networks. Only their model requires millions of documents for fine-tuning. Howard and Ruder develop universal language model fine-tuning for text classification in which they use various implementation tricks for proper fine tuning [36]. They design pre-training techniques for neural networks using a language modelling objective and then fine tuning to a classification task with supervision. The authors show

that their model can achieve considerable accuracies from very few examples. Their main contribution comes from the training and optimization techniques including discriminative fine tuning, slanted triangular learning rate schemas and gradual unfreezing of layers. Radford *et al.* [18] implement a language model based on the transformer architecture by Vaswani *et al.* [39] and design task specific input transformations for fine-tuning to various tasks. The authors discuss the various latest advances in the language modelling research and how they can be used to improve specific tasks such as Question-Answering or sentiment classification. The self-attention model roots from the transformer architecture [39], a sequence transduction architecture originally designed in the context of language translation tasks. State of the art work from Devlin *et al.* [19] redefines the bidirectional masked language model objective in a Cloze task scenario. Their model is also grounded on transformer architecture from Vaswani et al. [39], but is a larger model and is trained on very large corpora. We use the weights of pre-trained model by Radford *et al.* [18] and Devlin *et al.* [19] to initialize our architecture, and select the best performing.

2.4 Attention in Relation Extraction

Only few existing works use attention mechanism for extracting relationships [8, 40]. Zhang *et al.* add position data for the subject and object through attention mechanism for the task of relation extraction, more specifically slot filling [8]. Motivated from their work, we adopt masking over entity mentions in our model which give the position information. Nguyen *et al.* apply deep biaffine attention over LSTM [40], we modify our encoder and decoder in a similar fashion, but greatly simplify their model (the details are explained in Section 3). Alt *et al.* apply self-attention based pre-trained architecture to relation extraction [41]. In contrast to our work, their work addresses relation classification for a given pair of named entities in a single sentence and do not train their model based on multi-task objective.

2.5 Applying NLP to EHR

Prior work has applied NLP to EHR but in limited ways. Tran *et al.* [42] extract unigrams and model them using logistic regression for preterm birth prediction. Wagholikar *et al.* [43] used reg-

ular expressions to extract ejection fraction from Echo notes. However these works are rule-based and thus limited in recall. Regular expressions only capture limited patterns and hard to generalize. Marafino *et al.* [44] explored text modelling using logistic regression and augment these features to clinical trajectory based features for predicting mortality. This work explores regular expressions and Lucene based Apache Solr [45] for constructing feature based representation of the Echo note. We explored this avenue to compare our deep learning based representations. Hassanpour *et al.* [46] developed a corpus for Named Entity Extraction and modelled the data using Conditional Markov Model and Conditional Random Field. Cornegruta *et al.* [23] used Deep Learning based Bi-LSTM technique to model the 4 entities which the authors annotate for a list of 2000 documents. Unlike their work, we adopt an abstractive representation instead of extractive. Rajkomar *et al.* [47] worked on predicting various clinical outcomes including In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis using deep learning. But the authors did not include free text in their model. On similar lines, Pabkin *et al.* [48] work on predicting ICU readmission based on data apart from free text. Our work is most similar to the work of Liu *et al.* [49], which uses CNN and LSTM based architectures to model text in the EHR. Our work differs from them in the following ways. Firstly, we modify pre-trained models to account for label in-provision. Secondly, we work on attention based models which inherently account for interpretability to predict ICU readmission.

3. PROPOSED APPROACH

3.1 Summary

We introduce an approach for EERE and relation classification modifying pre-trained architectures trained on language modeling objective. The goal is to build a generic framework for relation extraction at the sentence level, and across a document, but also to apply our model in a realistic setting. Henceforth, we study EERE on CDR dataset and relation classification on SemEval dataset. Finally, we use these architectures for extracting features from a medical note for predicting a clinical event, more specifically predict ICU readmission using Echo notes. For these tasks, we modify transformer [39] encoder pre-trained on variants of language model objective. For EERE on CDR corpus and encoding clinical note, we use GPT [18] and for SemEval we use BERT [19]. In this work, we use the terminology introduced by Devlin *et al.* [19], calling GPT as decoder and BERT as encoder. We use these architectures to modify the tokenized sub-word representations. These contextual representations are then used to predict entities and classify relations. We illustrate two variants of our approach for CDR (EERE) and SemEval (relation classification) respectively. For CDR variant we first predict entities and then relations. For each of the entity pair, we use max pooling over all its mention pairs for both entities to design a relation candidate. For SemEval variant we experiment with multiple variants of encoding a given sentence and at the end layer, we apply a linear layer over the classification token representation. We follow a similar approach to encode the clinical note. We call the prior approach CDR variant and later SemEval variant.

3.2 Input

Let S be the input to our model. S for CDR represents a PubMed abstract, for SemEval a sentence and medical note for Echo reports. Here we describe the transformation to the input text before giving the text to the Language representation model. First, the text is tokenized to sub-word representations and then followed by the embedding layer. After embedding the tokenized

entities, input transformation is applied for aiding in classification. Then, position information is added and the representations are sent to the Language representation model.

3.2.1 Tokenization and Embedding

The model tokenizes the input to N tokens and embeds each word to d dimensions i.e. embeddings in $\mathbb{R}^d$, where N is 512 and d is 768. We choose these parameter values to be consistent with the pre-trained architectures. For tokenization, we use the byte pair encoding (BPE) algorithm following standard implementation with 40000 merges [50]. The embedding layer takes in byte pair encoded sub-word representations and gives out d dimensional vectors as outputs, this embedding layer is accompanied by the pre-trained architecture. After tokenization of the input sentence S we obtain the tokenized version of the input represented as $U = \{u_1, ..., u_n\}$.

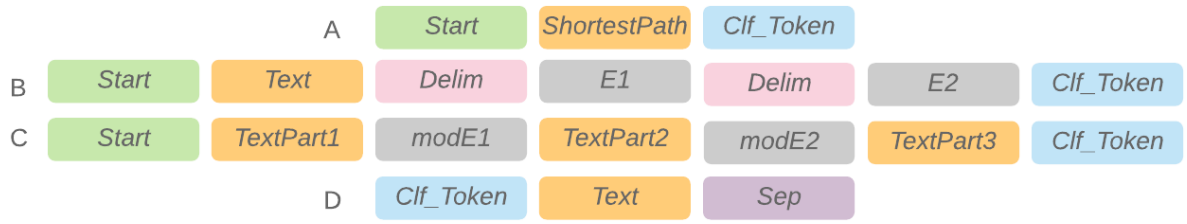

Figure 3.1: Input Transformations for fine-tuning. A describes the transformation for the shortest dependency path sequence. B is a transformation for the surface sequence where the entity information is presented at the end, C illustrates a more natural transformation for surface form sequence, here mod modification symbolizes that each entity is prepended and appended with special delimiters. D describes the transformation for BERT

3.2.2 Input Transformation

For each task, we transform the embedded input U for fine-tuning. Due to the training of the pre-trained model on plain text, input transformations are needed to apply it to our tasks. These transformations help the model discriminate the input and also act as sentence representation corresponding to the classification token [18, 19]. Figure 3.1 summarizes the input transformations

we adopted and we further elaborate the implementation details below. We denote X to the transformed version of U.

3.2.2.1 CDR

We do not transform CDR dataset since we needed no sentence embedding and we needed a consistent representation for entity recognition. This strategy was empirically chosen. While classifying relationships, we attempted to give sentence-based representation but found that of little help.

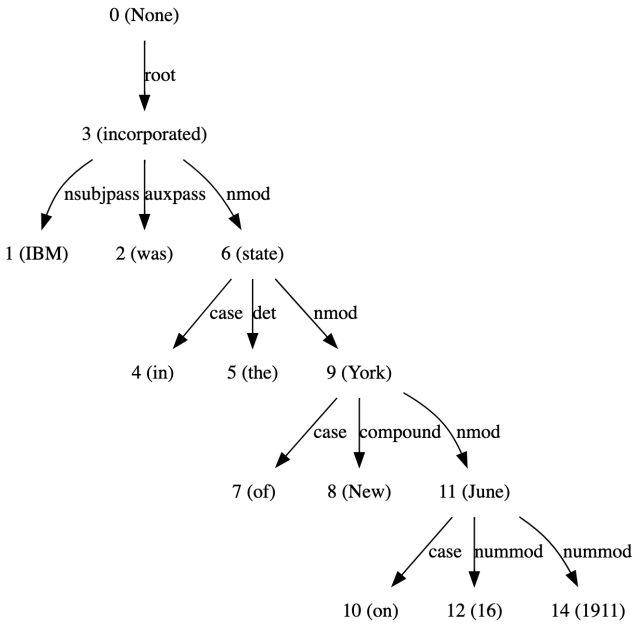

Figure 3.2: Dependency Parse with labelled edges. For entities IBM and $NewYork$ the shortest path consists of path through the Lowest Common Ancestor i.e. $incorporated\ state$

3.2.2.2 SemEval

For SemEval dataset we experimented multiple input transformations. Following Miwa and Bansal [12], we model the shortest path in the dependency path tree as a sequence, since this was shown to be highly correlative of the relation being extracted. Figure 3.2 illustrates an example of shortest path between the two entities. We use Figure 3.1a transformation for the shortest path,

treating it as a sequence. Model pre-training for BERT and GPT [18, 19] was done on the surface form, hence this variant did not perform well. We then experimented two variants of the surface form as shown in Figure 3.1b and 3.1c. We augmented entity location information using delimiters. In first variant we append the sentence with two entities separated by delimiters, here the entities repeat at the end of the input. And in the second variant, we give the model a more natural variant of the surface form. Here we highlighted the entities with delimiters. Since this is a more natural way to provide information to the model, we see that that the performance is improved. We then experiment SemEval with Bert for which we only use SegmentA embeddings, since there is only one sentence. The transformation is simple and can be seen in Figure 3.1 D

3.2.2.3 Echo notes

For Echo notes, we followed a strategy similar to Figure 3.1a, but shortest path replaced with Echo note. As noted earlier we observed that GPT based decoder worked best with predicting ICU readmission and hence we report the same in Section 4.

3.2.3 Position based information

We use the learned position embedding and hence we limit the test sentence length to N, i.e. 512. The dimension of position embedding is consistent with the token dimension d, i.e 768. Consistency is maintained since this information is added to the token representation following the Equation 3.1. This position related information is important for the model to understand the sequential nature of the input.

$$x_i = e_i + p_i \tag{3.1}$$

Here e_i corresponds to the i^{th} token embedding representation of the u_i token obtained by selecting i^{th} row from W_e. Similarly, p_i corresponds to the i^{th} position embedding representation obtained by selecting i^{th} row from W_p

$$s^{(0)} = XW_e + W_p \tag{3.2}$$

Equation 3.2 summarizes the input layer operations including embedding and adding position information. Here X is the sequence of the tokens $X = (x_1, .x_n)$ is the output from Input transformation block, W_e is the learned embedding weight matrix and W_p is the position encoding. X is then projected into the embedding space and added with a learned position embedding with the same dimension. The embedding matrices W_e and W_p are part of the pre-trained architecture. $s^{(0)}$ is the representation of the 1^{st} layer input.

3.3 Language representation model

We first encode all tokens through sequential encoder i.e. pre-trained self-attention based transformer. Transformer model by Vaswani *et al.* [39] is an architecture dispensing recurrence entirely. The authors tackle the machine translation problem using a sequence transduction model. The encoder and the decoder are made of attention only and position wise feedforward neural networks without any convolution or recurrent elements. The rationale behind the model is supported by two main reasons. Firstly, the model supported better utilization of computation since it could be paralyzed over the sequence length. Secondly, the model could directly attend from any position to any other position, making the number of computations to learn dependencies between any two positions in constant time. This architecture helps the model tackle the problem of vanishing gradient for long inputs explicitly. Noting the potential and empirical efficiency of self-attention architecture towards language tasks, we inculcate the same in our model.

Semi-supervised pre-training helps in better initialization of the model by pre-training on language model objective on large corpora. Hence, we modify GPT and BERT architectures for our approach. Models for CDR and Echo note datasets are based on GPT architecture, they were pre-trained on BooksCorpus dataset (800M words) [51]. The SemEval variant based on BERT is additionally trained on English Wikipedia (2,500M words). Since the use of Transformers is prevalent, and our implementation of Language representation model strictly follows standard architecture, we omit a comprehensive background of this part of the model. We recommend readers to Vaswani *et al.* [39].

The attention consists of h heads. The model is a $L = 12$ layer transformer. GPT decoder uses

masking and the BERT encoder doesn't. We refer the reader more comprehensive information from Radford *et al.* [18] and Devlin *et al.* [19] work. The following equations summarize the language representation model, which could be encoder or decoder.

$$s^{(g)} = TransformerBlock^{(g)}(s^{(g-1)}) \forall g \in [1, L] \tag{3.3}$$

The model consists of L residual multi-head self-attention layers with self-attention and position-wise feed forward neural networks as sublayers. Denoting the g^{th} self-attention block as $TransformerBlock^{(g)}(.)$, the output of layer $s^{(g)}$, and $LayerNorm(.)$ layer normalization, the above recurrence in 3.3 is applied to the input and it gives the final representation $s^{(p)}$. Each $TransformerBlock^{(g)}(.)$, consists of Multi-head self-attention and feed forward neural network.

3.4 Model Variants

3.4.1 CDR

For EERE, we design an approach to jointly model named entities and relations between those predicted entities. The language representation model parameters act as shared parameters between these two tasks. In some architectures, no shared parameters are used for detecting entities and relations. Previous research [12] has shown that having shared parameters in a single model helps model the interactions needed for both tasks and thus improves performance. Training entities and relations separately in a pipeline fashion can cascade errors, also entity prediction cannot be benefited from relation classification and vice versa. Hence, we train both tasks jointly.

Figure 3.3 gives an overview of the model architecture for CDR variant, here we intentionally hide some details for clarity.

Figure provides an overview of the model architecture for CDR variant, here we deliberately conceal some details for visual clarity. Internal working of the self-attention architecture is abstracted and the entity prediction is not shown. We extract p entities $\{e_1, e_2, ...e_p\}$ having q relations $\{r_1, r_2, r_3..r_q\}$. Each relation is a triple of $\{e_i, e_j, r_k\}$, here e_i and e_j are the entities in the relation r_k.

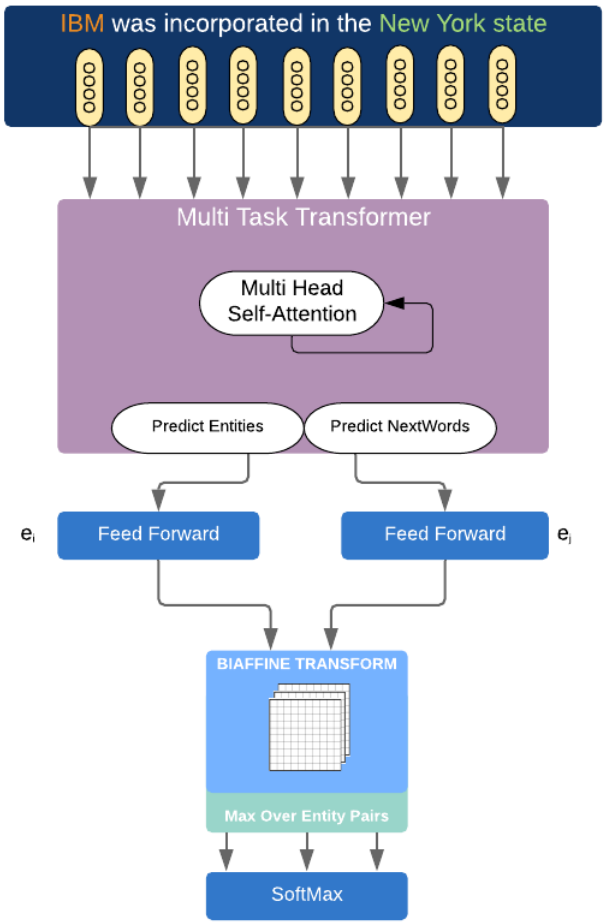

Figure 3.3: End-to-end Relation Extraction Architecture. Inputs are sub word byte pair embeddings. Inputs are then passed through the self-attention encoder and entity prediction layer. Relation extraction is made possible from a Biaffine head with masked max-pooling based on entity information

3.4.1.1 Named Entity Recognition

Rephrasing Pawar *et al.* [9], given a sentence as our input, our goal is to assign words with Begin Inside Out (BIO) tags and its entity type. Since we operate at sub-word level, we treat each sub-word as I tag to denote the entity span. Entity predictions are done using the output of the final layer which is then passed through a linear layer and then softmax layer. Equation 3.4 summarizes the NER operation for classifying entities. We note the loss associated with NER as L_{ner}. Equation 3.4 gives the loss for NER for a single abstract.

$$C = W^{entity} s^{(g)} \tag{3.4}$$

$$L_{ner} = 1/N \sum_{t=1}^{N} \log P(y_t | s^{(g)}) \tag{3.5}$$

3.4.1.2 Relation Extraction

For classifying relations between each of those identified entities, we define a relation extraction head on top of self-attention architecture. The representations $s^{(L)}$ are sent to two representational spaces to representing head and tail in relation candidate. Equation 3.6 summarizes the linear layers for head and tail representations

$$\begin{aligned} e^{head} &= W^{(1)}_{head}(ReLU(W^{(0)}_{head} s^{(L)})) \\ e^{tail} &= W^{(1)}_{tail}(ReLU(W^{(0)}_{tail} s^{(L)})) \end{aligned} \tag{3.6}$$

$$scores(head, tail) = masked(max(A_{ij})) \tag{3.7}$$

For calculating relation candidates, we pass the above head and tail representations through a biaffine layer to capture the dependencies between the words in two spaces. We calculate the following tensor based on the biaffine weight matrix $W_{relations}$, whose dimensions are dX We then mask the biaffine output with dimensions (N, N, R), where R is the number of classes in relations.

The outputs are then max-pooled for obtaining relation candidates and trained using Logistic loss. We note the loss associated with Relation Extraction for Q relation as L_{rel}.

3.4.1.3 Training

The multi-task loss consists of the NER, Rel Ex loss and Language modeling loss. During ablation studies we noted that this is necessary for training both the relation extraction and NER task. The loss equation for training objective is summarized as follows in the Equation 3.8

$$L_{total} = L_{rel} + ner\_coef * L_{ner} + lm\_coef * L_{lm} \tag{3.8}$$

In Equation 3.8, L_{lm} refers to the language model loss. Here ner_coef refers to the coefficient corresponding to the NER loss and lm_coef corresponds to the Language model loss

3.4.2 SemEval and Echo Note dataset

For fine tuning on SemEval dataset, we pass <*clf*> token representation through a linear+softmax layer acting as the final layer of our model. For the best performance we use BERT for fine-tuning on SemEval dataset and follow the input transformations mentioned in Section 3.2.2. Analogously, for Echo notes we use a similar model but fine-tune using GPT. The below figure summarizes the model we used for these tasks.

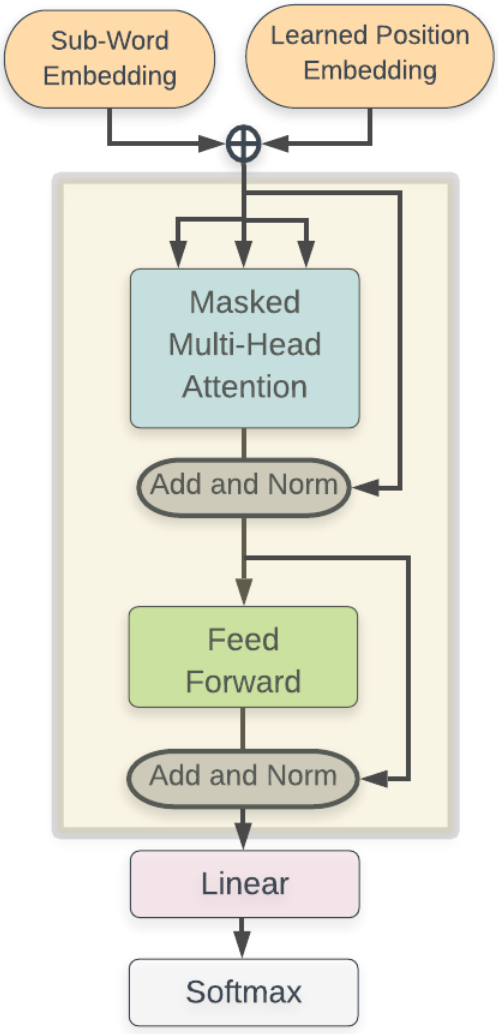

Figure 3.4: Transformer architecture used in representing the SemEval relation candidate and Echo Note.

4. DATA AND EXPERIMENTS

We evaluate our approach on three datasets. First, on Biocreative V Chemical Disease Relation corpus for the task of end-to-end relation extraction [20, 21]. Second, on SemEval-2010 Task 8 dataset on nominal relation classification [22]. Finally, on predicting ICU readmission on MIMIC-III cohort of patients with Echo notes [24]. We use the first two datasets to showcase the prowess of our approach to relation extraction and the last dataset to understand the approach in a medical setting.

4.1 End-to-end Relation Extraction Results

Title: Electrocardiographic evidence of myocardial injury in psychiatrically hospitalized cocaine abusers. **Abstract**: The electrocardiograms (ECG) of 99 cocaine-abusing patients were compared with the ECGs of 50 schizophrenic controls. Eleven of the cocaine abusers and none of the controls had ECG evidence of significant myocardial injury defined as myocardial infarction, ischemia, and bundle branch block.

The ORANGE COLOR are the diseases and the GREEN COLOR are the chemicals

2 Relations Extracted

- cocaine *CAUSES* myocardial infarction
- cocaine *CAUSES* bundle branch block

Figure 4.1: Example abstract from the CDR corpus

A unique gold standard dataset was created in BioCreative V challenge, including manual annotations of chemicals, diseases entities and their Chemically induced disease (CID) relationships for 1500 PubMed papers. The dataset was derived from the Comparative Toxicogenomics Database (CTD), which curates interactions between genes, chemicals, and disease. The dataset statistics are presented in Table 4.1, the dataset is equally split into training, validation, and test. An example abstract is shown in Figure 4.1.

Table 4.1: Data statistics of CDR chemical and disease entities and CID relations

Data split	Docs	Chemical mention	Disease Mention	Pos relations	Neg relations
Train	500	5203	4182	1038	4280
Development	500	5347	4244	1012	4136
Test	500	5385	4424	1066	4270

Table 4.2: Summary of entity recognition for chemical and diseases entities, different averages are measured for clarity

Average Metric	**P**	**R**	**F1**
micro avg	0.93	0.93	0.93
macro avg	0.82	0.79	0.81
weighted avg	0.93	0.93	0.93

Table 4.3 compares our results of relation extraction with other state of the art works. We note that the model apart from Verga *et al.* [33] use external linguistic resources and are based on ensemble models. Hence our model performs equivalently with no ensembles and no external resources. Table 4.2 summarizes the entity extraction results performance.

As noted earlier, we use GPT for CDR modified with relation extraction head. We saw that this configuration empirically performed best since the model has regularization loss from the language model. BERT did not perform relatively well and hence we did not note its performance in the ablation study in Table 4.4. The best performance for both entity recognition and relation extraction was recorded only in the multi-task scenario, i.e. both tasks help each other. We choose to pick the most dominant interaction i.e. Max-Pool among interactions, this further increased performance.

4.2 Nominal Relation Classification Results

We work on a dataset on nominal relation classification (SemEval-2010 Task 8). SemEval-2010 Task 8 has 9 relation types between entities and a tenth type $Other$ when the entities are not related. It is a Multi-Way Classification of Semantic Relations between pairs of Nominals

Table 4.3: Comparison of performance for various SOTA model on CID relation extraction

Models and Settings	**P**	**R**	**F1**
MaxEnt (Gu et al., 2016)	62	55.1	58.3
Pattern rule-based (Lowe et al., 2016)	59.3	62.3	60.8
LSTM-based (Zhou et al., 2016)	**64.9**	49.3	56
LSTM-based & PP (Zhou et al., 2016)	55.6	68.4	61.3
CNN-based (Gu et al., 2017)	60.9	59.5	60.2
CNN-based & PP (Gu et al., 2017)	55.7	68.1	61.3
BRAN (Verga et al., 2017)	55.6	**70.8**	**62.1**
SVM+APG (Panyam et al., 2018)	53.2	69.7	60.3
Our Model (Predict all relations at Once with Entities Prediction)	51.4	70.35	*59.95*

Table 4.4: Ablation study for CDR dataset for model based on GPT

Model Ablations	**F1**
Train only Head, Batch size 8	48.90
Train Entire Model, Batch size 8	54.00
Include part of validation data for training, Batch size 8	55.11
Decrease batch size 4	57.21
Change Optimizer to Adam lr=1e-4, betas=[0.9, 0.999], eps=1e-8	59.4
Change Optimizer and introduce Max-Pool	**59.95**

and relations are asymmetric. We neglect this other relation type and consider it a 10th type. The dataset consists of 8,000 training and 2,717 test sentences, and each sentence is annotated with a relation between two given entities. One example of the relation candidate is *"People in Hawaii might be feeling <e1>aftershocks</e1> from that <e2>powerful earthquake</e2> for weeks."* Here Cause-Effect(e1, e2) does not hold, but Cause-Effect(e2, e1) does. Most of the input candidates are short sentences similar to this sentence, thus the relationships are between entities having single mention and are present in the same sentence. We use this setting since we wanted to completely analyze our model on a different setting contrasting to CDR before. The dataset types are summarized in Table 4.5 below. Official score is macro-averaged F1-score (9+1)-way classification, with directionality.

Table 4.5: Data statistics of SemEval relation types including an example of each class

Relation Type	Example	Freq
Cause-Effect	Smoking causes cancer.	1331 (12.4%)
Instrument-Agency	The murderer used an axe.	660 (6.2%)
Product-Producer	Bees make honey.	948 (8.8%)
Content-Container	The cat is in the hat.	732 (6.8%)
Entity-Origin	Vinegar is made from wine.	974 (9.1%)
Entity-Destination	The car arrived at the station.	1137 (10.6%)
Component-Whole	The laptop has a fast processor.	1253 (11.7%)
Member-Collection	There are ten cows in the herd.	923 (8.6%)
Message-Topic	You interrupted a lecture on maths.	895 (8.4%)
Other	N/A	1864 (17.4%)
Total		10717(100%)

We describe the results in Table 4.6. For baselines, we use all the previous works mentioned in the table. SemEval is an intensively studied dataset, this gives our model a fair comparison. The results in this table correspond to our best performing model with BERT and maximum epochs of 10. Hence, we do a thorough ablation study with various input transformations, different self-attention encoder based on GPT and BERT and various hyper-parameters involved. The summary is presented in Table 4.7

4.3 ICU readmission results

4.3.1 Cohort

We select Echo notes from Echocardiography reports from Medical Information Mart for Intensive Care (MIMIC-III) database [24]. The MIMIC-III dataset is comprised of deidentified information detailing over $60,000$ ICU stays from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. This information was collected as part of routine clinical care and, as such, is representative of the information that would be available to clinicians in real-time. The database contains a wide range of numerical data such as lab results and patients' vitals which has been extracted to predict readmission in previous work. MIMIC-III also contains a large amount of unstructured information in the form of free text notes. These notes are taken down by medical

Table 4.6: Relation Classification Results on SemEval data and comparison with state of the art works

MODELS AND SETTINGS	**Macro-F1**
Feature Based	
Jin et al. (2019) SVM-RBF	0.781
Zeng et al. (CNN and feature based)	0.827
Settings - No External Knowledge Resources	
dos Santos et al. (2015)(CNN based)	0.841
Xu et al. (2015) (lstm-crf)	0.84
Miwa and Bansal et al. (2017) (Tree LSTM based)	0.844
Cai et al. 2017 (RCNN) **no direction prediction	0.854
Lee et al. (2019) (bilstm-attention)	0.847
$OurModel(Attention - only)$	0.869

Table 4.7: Model Ablations for Nominal Relation Classification on SemEval dataset

Model Ablations and Experiments	**F1**
RelShortestPath (<start >+ <shortest_path >+ <clf_token >)	0.772
Bilinear Fashion	0.8
RelSeq - <start>+ <sentence>+ <delim>+ <e1>+ <delim>+ <e2>+ <clf_token>	0.839
RelSeq - <start>+ <text1><e1>+ <text2>+ <e2>+ <text3>+ <clf_token>	0.859
Change Model Initialization masked LM(Devlin et al.)	0.841
Add additional Output Layer	0.842
Predict Dependencies	0.85
Predict Dependencies and Increase Epochs to 10	0.869

professionals during patients' stays and contain a large amount of both qualitative and quantitative information regarding the patients' hospital stay. The total number of Echo Notes is 45794 and the number of unique patients with Echo notes is 29173, hence we use these patients as the cohort for predicting ICU readmission. Patients are dropped from this cohort if they are missing either time for ICU admission or discharge as this makes it impossible to calculate readmission times, resulting the cohort size to $25,320$. This cohort is split into training, validation, and testing groups using an $80/5/15$ split for our work.

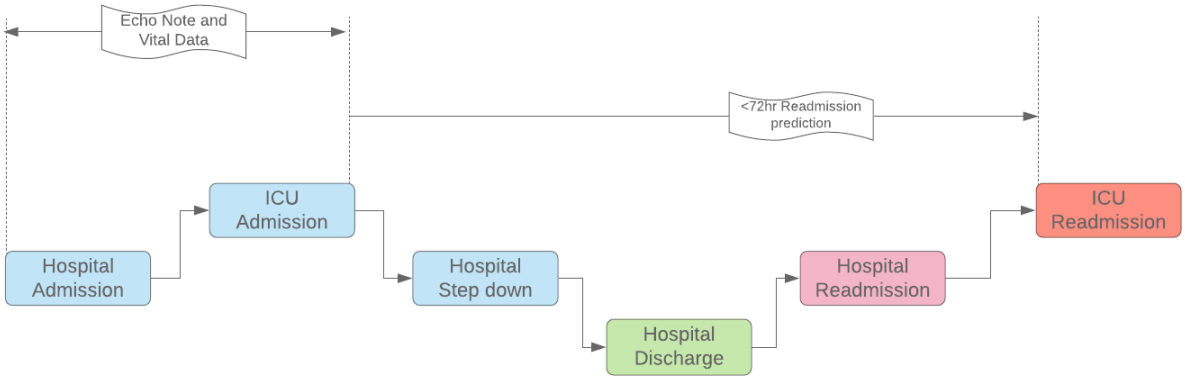

Figure 4.2: Patient flow in an ICU admission, the red color implies the patient gets readmitted to the hospital, here we categorize the patient as positive sample if he gets readmitted within 72 hours of ICU discharge

4.3.2 Task

The task can be better understood by Figure 4.2. The patient passes through various stages in a typical hospital stay. The patient's stay in the hospital starts with hospital admission followed by ICU admission. It is important to note that our patient cohort only consists of patients who have been to the ICU at least once. The patient flow after the ICU can vary and marks the event we aim to predict in this work. We also visualize the words which the <*clf*> token attends to since this is used for classification. This provides partial interpretability to understand the model's focus for prediction. Figure 4.3 showcases a typical example of this visualization. We note that since the model is a multi-headed model to learn different features. This visualization is for only a single head, which learns some features. Other heads learn other features. This example is cherry picked for better understanding. For this note, the model actually picks on important words for predicting readmission for patient who was actually readmitted.

4.3.3 Results

We summarize our results on ICU readmission in Table 4.8. In the cohort ICU patients being readmitted are 1327 (positive class) and patients with readmission are 27941 (negative class). Since

Table 4.8: Results on ICU readmission task

Experiment Description	**F1**	**AUC**
Full Data - complete zero prediction	0	0.5
Train Language Model first and then classification Experiment	0	0.49
Language Model Loss + Classification Loss	0.73	0.565
Model with subsampled data -sigmoid - n iterations = 3	0.72	0.599
Baseline without Text data	0.72	0.62
Model with subsampled data -sigmoid - Weighted Loss Function	0.73	0.634

the dataset is unbalanced, models with a large number of parameters tend to overfit. Hence we use subsampling for training and use GPT instead of BERT since while fine-tuning GPT we can regularize the main loss with language model loss. We treat Pabkin *et al.* [48] model as our baseline, we train their model on Echo cohort to obtain the metrics noted in row 5 of Table 4.8.

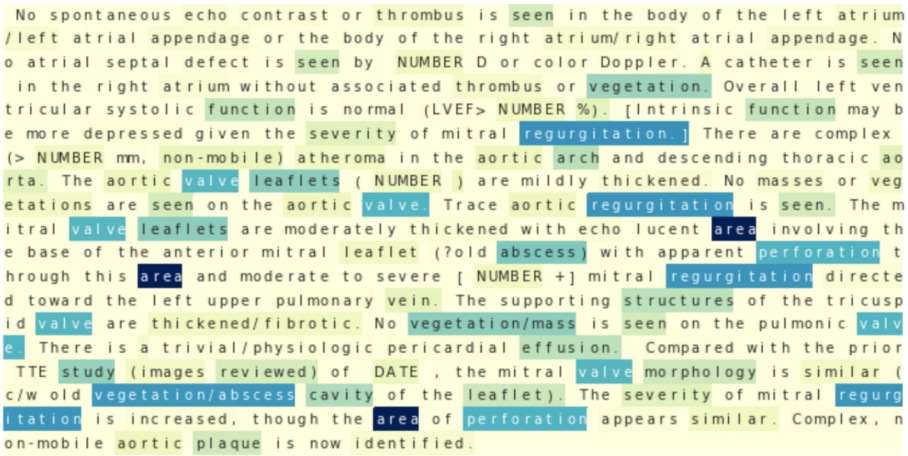

Figure 4.3: Attention scores visualized on a example Echo note. Here we can see that the patient has severe health conditions, Intrinsic function is more depressed, Trace Aortic regurgitation, Our model picks up on these terms and gives the correct prediction for the patient who is readmitted in the future

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

We designed a novel framework for end-to-end relation extraction using multi-task training on self-attention based pre-trained architectures. Our model performed comparative to the baseline on CDR (CID relations F1-score), and outperforms previous state of the art on SemEval relation classification dataset by 2.2 absolute F1 score for relations. We demonstrated an approach to extract relations at sentence and document level without using any external linguistic resource. We show that multi-task fine-tuning is helpful for related tasks. This approach is generic and can be applied across tasks such as syntactic parsing, semantic parsing and Semantic role labelling. We further extract features from medical notes for heart patient to predict ICU readmission, we outperform the baseline by 14% absolute AUROC. Through this we showcase an example of applying pre-trained models on clinical notes and predicting important events.

5.2 Future Work

We have observed in practice that this approach is more efficient than alternatives for practical tasks and part of future work will be measuring this much more precisely and conclude empirically. Naturally, this work applies to model semantic parsing and we would, therefore, like to apply this model to this task. One key avenue is to explore the different ways of inculcating structural information into the model. In this work, we have observed multiple cases of overfitting because of the large number of parameters for massive models, hence we encourage research to understand techniques to evade this.

REFERENCES

[1] J. H. Martin and D. Jurafsky, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

[2] M. Rodrigues and A. Teixeira, *Advanced applications of natural language processing for performing information extraction*. Springer, 2015.

[3] I. Spasić, B. Zhao, C. B. Jones, and K. Button, "Kneetex: an ontology–driven system for information extraction from mri reports," *Journal of biomedical semantics*, vol. 6, no. 1, p. 34, 2015.

[4] N. Kaushik and N. Chatterjee, "Automatic relationship extraction from agricultural text for ontology construction," *Information processing in agriculture*, vol. 5, no. 1, pp. 60–73, 2018.

[5] W. Li, "Abstractive multi-document summarization with semantic information extraction," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1908–1913, 2015.

[6] M. Abulaish and L. Dey, "Biological relation extraction and query answering from medline abstracts using ontology-based text mining," *Data & Knowledge Engineering*, vol. 61, no. 2, pp. 228–262, 2007.

[7] M. Yu, W. Yin, K. S. Hasan, C. d. Santos, B. Xiang, and B. Zhou, "Improved neural relation detection for knowledge base question answering," *arXiv preprint arXiv:1704.06194*, 2017.

[8] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, "Position-aware attention and supervised data improve slot filling," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 35–45, 2017.

[9] S. Pawar, P. Bhattacharyya, and G. Palshikar, "End-to-end relation extraction using neural networks and markov logic networks," in *Proceedings of the 15th Conference of the European*

Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 818–827, 2017.

[10] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1083–1106, 2003.

[11] Z. GuoDong, S. Jian, Z. Jie, and Z. Min, "Exploring various knowledge in relation extraction," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 427–434, Association for Computational Linguistics, 2005.

[12] M. Miwa and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures," *arXiv preprint arXiv:1601.00770*, 2016.

[13] M. Miwa and Y. Sasaki, "Modeling joint entity and relation extraction with table representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1858–1869, 2014.

[14] Q. Li and H. Ji, "Incremental joint extraction of entity mentions and relations," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 402–412, 2014.

[15] F. Christopoulou, M. Miwa, and S. Ananiadou, "A walk-based model on entity graphs for relation extraction," *arXiv preprint arXiv:1902.07023*, 2019.

[16] Y. Zhang, P. Qi, and C. D. Manning, "Graph convolution over pruned dependency trees improves relation extraction," *arXiv preprint arXiv:1809.10185*, 2018.

[17] Y. Peng, C.-H. Wei, and Z. Lu, "Improving chemical disease relation extraction with rich features and weakly labeled data," *Journal of cheminformatics*, vol. 8, no. 1, p. 53, 2016.

[18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*, 2018.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu, "Overview of the biocreative v chemical disease relation (cdr) task," in *Proceedings of the fifth BioCreative challenge evaluation workshop*, pp. 154–166, 2015.

[21] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "Biocreative v cdr task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, 2016.

[22] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pp. 94–99, Association for Computational Linguistics, 2009.

[23] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, "Modelling radiological language with bidirectional long short-term memory networks," *arXiv preprint arXiv:1609.08409*, 2016.

[24] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, pp. 160035 EP –, May 2016. Data Descriptor.

[25] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] K. Xu, Y. Feng, S. Huang, and D. Zhao, "Semantic relation classification via convolutional neural networks with simple negative sampling," *arXiv preprint arXiv:1506.07650*, 2015.

[28] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, "Classifying relations via long short term memory networks along shortest dependency paths," in *proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1785–1794, 2015.

[29] C. N. d. Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," *arXiv preprint arXiv:1504.06580*, 2015.

[30] L. Li, Y. Guo, S. Qian, and A. Zhou, "An end-to-end entity and relation extraction network with multi-head attention," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp. 136–146, Springer, 2018.

[31] H. Poon, K. Toutanova, and C. Quirk, "Distant supervision for cancer pathway extraction from text," in *Pacific Symposium on Biocomputing Co-Chairs*, pp. 120–131, World Scientific, 2014.

[32] N. Peng, H. Poon, C. Quirk, K. Toutanova, and W.-t. Yih, "Cross-sentence n-ary relation extraction with graph lstms," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 101–115, 2017.

[33] P. Verga, E. Strubell, and A. McCallum, "Simultaneously self-attending to all mentions for full-abstract biological relation extraction," *arXiv preprint arXiv:1802.10569*, 2018.

[34] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[36] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[37] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[38] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in neural information processing systems*, pp. 3079–3087, 2015.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[40] D. Q. Nguyen and K. Verspoor, "End-to-end neural relation extraction using deep biaffine attention," in *European Conference on Information Retrieval*, pp. 729–738, Springer, 2019.

[41] C. Alt, M. Hübner, and L. Hennig, "Improving relation extraction by pre-trained language representations," 2018.

[42] T. Tran, W. Luo, D. Phung, J. Morris, K. Rickard, and S. Venkatesh, "Preterm birth prediction: Deriving stable and interpretable rules from high dimensional data," in *Conference on machine learning in healthcare, LA, USA*, 2016.

[43] K. B. Wagholikar, C. M. Fischer, A. P. Goodson, C. Herrick, M. Rees, E. Toscano, C. A. MacRae, B. M. Scirica, A. S. Desai, and S. N. Murphy, "Extraction of ejection fraction from echocardiography notes for constructing a cohort of patients having heart failure with reduced ejection fraction (hfref)," in *Journal of Medical Systems*, 2018.

[44] B. J. Marafino, M. Park, J. M. Davies, R. Thombley, H. S. Luft, D. C. Sing, D. S. Kazi, C. DeJong, W. J. Boscardin, M. L. Dean, *et al.*, "Validation of prediction models for critical care outcomes using natural language processing of electronic health record data," *JAMA network open*, vol. 1, no. 8, pp. e185097–e185097, 2018.

[45] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in action: covers Apache Lucene 3.0*. Manning Publications Co., 2010.

[46] S. Hassanpour and C. P. Langlotz, "Information extraction from multi-institutional radiology reports," *Artificial intelligence in medicine*, vol. 66, pp. 29–39, 2016.

[47] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[48] A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. H. Krumholz, and J. B. Mortazavi, "Prediction of icu readmissions using data at patient discharge," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4932–4935, IEEE, 2018.

[49] J. Liu, Z. Zhang, and N. Razavian, "Deep ehr: Chronic disease prediction using medical notes," in *Proceedings of the 3rd Machine Learning for Healthcare Conference* (F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, eds.), vol. 85 of *Proceedings of Machine Learning Research*, (Palo Alto, California), pp. 440–464, PMLR, 17–18 Aug 2018.

[50] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[51] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.