Incorporating medical knowledge in BERT for clinical relation extraction

Arpita Roy and **Shimei Pan**
Department of Information Systems
University of Maryland, Baltimore County
Maryland, USA
{arpita2, shimei}@umbc.edu

Abstract

In recent years pre-trained language models (PLM) such as BERT have proven to be very effective in diverse NLP tasks such as Information Extraction, Sentiment Analysis and Question/Answering. Trained with massive general-domain text, these pre-trained language models capture rich syntactic, semantic and discourse information in the text. However, due to the differences between general and specific domain text (e.g., Wikipedia text versus clinic notes), these models may not be ideal for domain-specific tasks (e.g., extracting clinical relations). Furthermore, it may require additional medical knowledge to understand clinical text properly. To solve these issues, in this research, we conduct a comprehensive examination of different techniques to add medical knowledge into a pre-trained BERT model for clinical relation extraction. Our best model outperformed the state-of-the-art systems on the benchmark i2b2/VA 2010 clinical relation extraction dataset.

1 Introduction

In recent years pre-trained language models (PLMs) such as ELMo (Peters et al., 2017), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and GPT (Radford et al., 2018) have become very popular as they can effectively boost the performance of diverse NLP tasks such as Information Extraction (Shi and Lin, 2019; Jia et al., 2020), Sentiment Analysis(Gao et al., 2019), Question/Answering(Lv et al., 2020) and language entailment (Devlin et al., 2019). These models are trained on large text corpora using self-supervised tasks such as masked language modeling (MLM) and next sentence prediction. As these models can learn meaningful context-sensitive text embeddings, they are frequently used to encode input text in many downstream text analysis tasks. However, PLMs trained on general-domain text (e.g., books, Wikipedia and webdata) may not be ideal for domain-specific NLP applications (e.g., biomedical NLP). In this research, we explore how medical knowledge can be added to PLMs to facilitate clinical relation extraction.

Previously, significant effort has been made on adding domain knowledge into PLMs. Based on the types of knowledge added, we can group the work into two categories: integrating domain text in PLMs (Lee et al., 2020; Peng et al., 2019a; Gu et al., 2020) and integrating domain-specific knowledge graphs into PLMs (Wang et al., 2021; Zhang et al., 2019; Peters et al., 2019). In this study, we conduct a comprehensive investigation of these methods. We test their effectiveness in integrating knowledge from Unified Medical Language System (UMLS) into BERT for clinical relation extraction. Here we focus on UMLS because it is one of the most widely used bio-medical knowledge sources for clinical NLP. Among all the PLMs, we focus on BERT since it often achieves the state of the art performance on diverse NLP tasks. The main contributions of our work include:

- We conducted a comprehensive empirical analysis of the effectiveness of applying diverse knowledge integration techniques to combine medical knowledge encoded in UMLS with embeddings from pre-trained BERT models for clinical relation extraction.
- We proposed several knowledge fusion methods such as ClinicalBERT-EE-RI-CT/ST/SG, ClinicalBERT-EE-ED-CT and ClinicalBERT-EE-KB-MLM for clinical relation extraction.
- Our proposed method ClinicalBERT-EE-RI-ST achieved the state of the art performance on a benchmark clinical relation extraction dataset.

2 Related Work

In this section, we survey the representative work on topics that are most relevant to this research: (a) incorporating domain text in BERT during model training and (b) combining knowledge graph infor-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5357–5366
November 7–11, 2021.

mation with BERT. We also summarize the state-of-the-art techniques for extracting clinical relations from text.

Incorporating domain text in BERT: There are quite a few BERT models which have been trained (or fine-tuned) with bio-medical text: BioBERT (Lee et al., 2020) is pre-trained on PubMed abstracts and PMC full-text articles; ClinicalBERT (Alsentzer et al., 2019) is pre-trained on the clinic notes in the MIMIC-III database (Johnson et al., 2016); BlueBERT (Peng et al., 2019a) is pre-trained on the PubMed abstracts and the clinical notes in MIMIC-III; PubMedBERT (Gu et al., 2020) is pre-trained using abstracts from PubMed and full-text articles from PMC. Among them, BioBERT and BlueBERT are initialized with weights from a general-domain BERT model, ClinicalBERT is initialized from BioBERT, and PubMedBERT is trained from scratch.

Combining knowledge graph information with BERT: The simplest method to combine knowledge graph information with BERT is concatenation. For example, (Jeong et al., 2020) combines knowledge graph embedding trained using Graph Convolution Networks (GCN) with BERT embeddings for citation recommendation. Efforts to directly inject knowledge graph information into BERT can be further categorized into the following categories. (a) Joint optimization with knowledge graph objectives: Clinical KB-BERT (Hao et al., 2020) pre-trained BERT with a knowledge graph objective. In addition to predicting masked words, a triplet classification objective is added, where given a triplet of two concepts and a relation in UMLS, the model aims to correctly predict if the relationship exists between the two concepts. (b) Fusing entity embeddings from knowledge graphs with BERT: (Peters et al., 2019) first retrieves pre-trained entity embeddings from a knowledge graph, then uses them to update BERT word embeddings via word-to-entity attention. (Weinzierl et al., 2020) incorporates entity embeddings learned from a UMLS knowledge graph into BERT using adversarial learning. (c) Augmenting BERT input with knowledge graph information: (Liu et al., 2020) presents K-BERT in which triples from knowledge graphs are added into the input sentences before sent to BERT. In (Mitra et al., 2019), relevant knowledge statements are assigned to each training instance and BERT is fine-tuned on the modified training data to facilitate question answering tasks.

Clinical relation extraction: Early work on clinical relation extraction employed supervised machine learning with a wide range of hand-crafted features such as lexical features, syntactic features as well as semantic features extracted from external knowledge resources such as UMLS, cTAKES, and Medline. Among them, (de Bruijn et al., 2010; Minard et al., 2011) derived concept mapping and concept types based on UMLS. Later, pre-trained word embeddings (e.g. Word2Vec) became the most popular input features for relation extraction. A neural network-based classifier (e.g., CNN, LSTM, GCN) is often used to predict clinical relations based on word embeddings (Sahu et al., 2016; Luo et al., 2018; Li et al., 2019; Ningthoujam et al., 2019). Word embedding features were frequently combined with additional features such as word types, POS tags, IOB encoding of semantic concepts, relative distance, and dependency relations to further improve performance (Hasan et al., 2020). Recently, BERT-based text embedding has gained dominance due to it superior performance (Peng et al., 2019b). (Wei et al., 2019) is the first to combine BERT embeddings with traditional IOB tags. (Hasan et al., 2020) combines part-of-speech, IOB encoding, relative distance and dependency tree information with BERT. (He et al., 2020; Weinzierl et al., 2020) utilized pre-trained UMLS knowledge graph embeddings to enhance BERT.

Despite a substantial body of research on clinical relation extraction, it is still an open question in terms of what is the best method to integrate bio-medical knowledge graphs (e.g., UMLS) into BERT for clinical relation extraction.

3 Methodology

The main steps in this research include: (a) generating text embeddings using BERT, (b) aligning the entities in text with the concepts in UMLS, (c) generating UMLS knowledge graph embeddings and (d) integrating UMLS knowledge with BERT.

3.1 Generating Text Embeddings Using BERT

(Wu and He, 2019) shows that incorporating information about the target entities along with a BERT sentence representation greatly benefits relation classification. To implement this, given a sentence S, we insert four markers e11, e12, e21 and e22 at the beginning and end of the two target entities (e1,e2) in a relation. In the i2b2 relation

extraction dataset, the ground truth entity locations were provided as the input to the relation extraction model. After inserting these special tokens, for a sentence "The patient was given ibuprofen for high fever." with target entities "ibuprofen" and "fever" becomes: "[CLS] The patient was given e11 ibuprofen e12 for high e21 fever 22 . [SEP]". Based on the positions of the two target entities in the BERT embedding, entity embeddings (EE) can be calculated. Then sentence embedding derived from the [CLS] token embedding and the entity embeddings are concatenated and passed through a fully connected layer to generate a representation that contains both sentence and entity embeddings (BERT+EE).

3.2 Text and UMLS Concept Alignment

To incorporate external knowledge from UMLS into language models, we need to identify UMLS concepts in clinical notes. We use Apache cTAKES (Savova et al., 2010) to extract named entity mentions in clinical notes and align them with the concepts in UMLS. cTAKES is a clinical Text Analysis and Knowledge Extraction System that extracts clinical information from unstructured text. It processes clinical notes, identifies types of clinical named entities such as drugs, diseases/disorders, signs/symptoms, anatomical sites and procedures and maps them to UMLS concepts. Using cTAKES we can map 46,305 out of 58,688 entities in our dataset to UMLS concepts.

3.3 Generating UMLS Knowledge Graph Embeddings

The Unified Medical Language System (UMLS) (Bodenreider, 2004) is a repository of biomedical vocabularies developed by the National Library of Medicine of US. It has three knowledge sources: (a) **The Metathesaurus** integrates millions of concepts from over 200 vocabularies. The Metathesaurus is organized by concepts, each concept is characterized by a unique concept identifier (CUI), definition, attributes and relationships with other concepts. For example, the concept "Headache" (CUI C0018681) has a definition "The symptom of pain in the craninal region" and it is related to the concept "Acetanilide" (CUI C0000973) with the relation "may treat". (b) **Semantic Network** provides consistent categorization of all concepts represented in the UMLS Metathesaurus. Each concept in the Metathesaurus is assigned one or more semantic types, which are linked with each other through semantic relationships. Each semantic type has an identifier, a definition, a few examples, and a few relationships. Semantic groups are smaller and coarser-grained semantic type groupings. For example, the semantic type of "Headache" is "Sign or Symptom" and its semantic group is "Disorder". (c) **SPECIALIST Lexicon and Lexical Tools** include a large syntactic lexicon and tools for normalizing strings, generating lexical variants, and creating indexes.

Both the Metathesaurus and the Semantic Network can be considered as multi-relational knowledge graphs with nodes representing concepts or semantic types and edges representing relations. Each relationship is represented as a triplet (h, r, t), indicating a relationship (r) between two nodes (h and t). We use a subset of the Metathesauras and the complete semantic network to create our knowledge graph (KG). Specifically, we select all the CUIs extracted from our dataset using cTAKES. Then we select a subset of the Metathesauras by collecting all CUIs and relations that are one hop away from the initial set of CUIs. We connect this graph with the semantic network by including CUIs and their semantic type relationships. This created knowledge graph contains 312,474 nodes and 1,613,019 relations. An example of the created knowledge graph can be seen in figure 1.

Once the knowledge graph is created, we tested the effectiveness of several popular Knowledge Graph Embedding (KGE) models such as a translation-based model TransE (Bordes et al., 2013), two semantic matching models DistMult (Yang et al., 2014) and ComplEx (Trouillon et al., 2016) and a convolution network based models (Dettmers et al., 2018) and ConvKB (Nguyen et al., 2017) to create UMLS knowledge graph embeddings. We evaluate the effectiveness of these methods on a link prediction task, which predicts an entity that has a specific relation with a given entity, i.e., predicting h given (r, t) or t given (h, r). Among these KGE methods, ComplEx performed the best on the link prediction task. As a result, we only use knowledge graph embeddings from ComplEx in our experiments. From KGE, we can extract concept embedding, semantic type embedding, semantic group embedding and relation embedding.

3.4 Integrating UMLS knowledge with BERT

In our experiments, we primarily use ClinicalBERT trained on clinical text corpora. We systematically

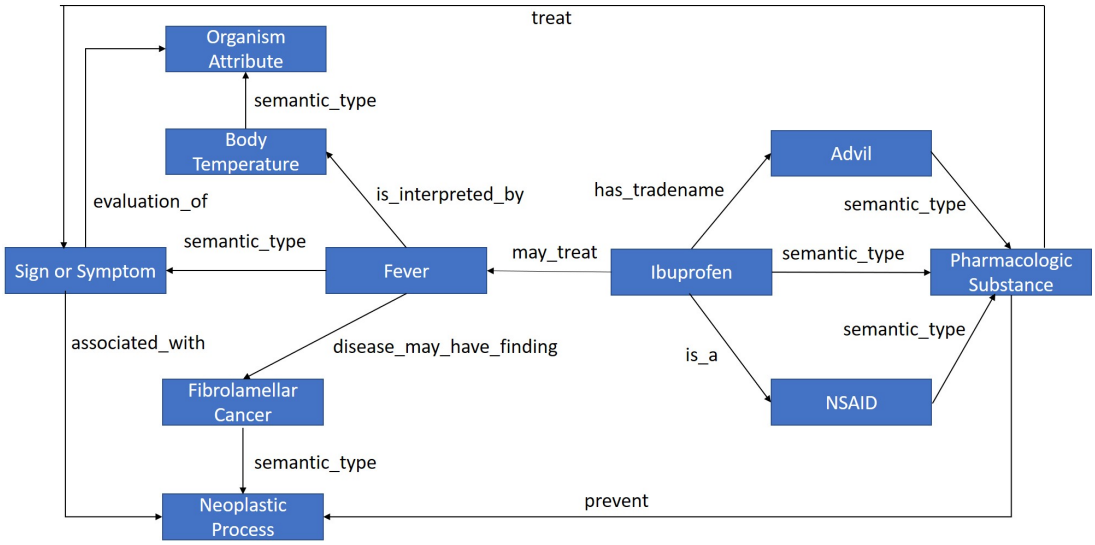

Figure 1: A snippet of a knowledge graph created from UMLS

investigate different techniques to infuse knowledge from UMLS with pre-trained ClinicalBERT. We have examined the following methods.

ClinicalBERT-EE-KGE: The first technique we tried is to combine knowledge graph embedding with the text embeddings from ClinicalBERT and feed them to the relation classifier. For two entities in an input sentence, we retrieve their respective concept embeddings (CT), semantic type embeddings (ST) and semantic group embeddings (SG) from KGE. In addition, for a pair of concepts mapped from two entities in a sentence, we use KGE to predict the UMLS relation between them. Then we retrieve the UMLS relation embedding from KGE. Finally, we concatenate all the KGE embeddings with the sentence and entity embeddings from ClinicalBERT for relation classification. Please note that in this approach, text embeddings and knowledge graph embeddings are in two separate embedding spaces.

ClinicalBERT-EE-MLP: Effectively merging knowledge graph embeddings with BERT can be tricky. Because pre-trained language models, such as BERT, are often trained for 2 to 5 epochs with smaller learning rate during fine-tuning, whereas graph embedding features extracted from KGE need to be trained for much longer with a higher learning rate. If we directly concatenate the BERT output with the KGE features, the relation classifier might not benefit much from the KGE features. To solve this issue, we first train a multi-layer perceptron (MLP) with knowledge graph embeddings for relation classification. The output of the MLP hidden layer is combined with BERT text embeddings in relation classification. The use of a trained MLP ensures that the model does not underfit when trained in an ensemble with pre-trained BERT models for a small number of epochs.

ClinicalBERT with Relation Indicator: In each input sentence, relevant knowledge from a knowledge graph is injected into an input sentence to BERT, which transforms the original sentence into a knowledge-enriched text input. We add knowledge from UMLS as the second sentence in the BERT input. Then, we feed both the original input sentence and the synthesized second sentence to pre-trained ClinicalBERT and it will use these knowledge enriched sentences to predict relation labels. With this method, we inject UMLS knowledge directly into the BERT embedding space. To construct this second sentence, first, we find corresponding CUIs for the two entities in a sentence using cTAKES. Then we use pre-trained KGE to predict the UMLS relation between them. Then we construct the second input sentence in the form of "concept1 relation concept2". For the input sentence "The patient was given ibuprofen for high fever", we first map "ibuprofen" and "fever" to their UMLS CUIs. Pre-trained KGE predicts the UMLS relation between them is "may_treat". Then we construct the second sentence as "ibuprofen may treat fever". This KGE-predicted UMLS relation can potentially act as a relation indicator that may help to differentiate relation class labels. To pass this relation indicator information to BERT, we use special tokens before and after the relation indica-

tor phrase. These tokens are used to extract the relation indicator embedding from BERT. Finally, the combined sentence embedding, entity embedding and relation indicator embedding are used for relation classification. For example, the final input would be "[CLS] Patient was given e11 ibuprofen e12 for high e21 fever e22 . [SEP] ibuprofen r31 may treat r32 fever . [SEP]". We also tried a variety of templates to generate the second sentence where each entity is replaced by its semantic type or semantic group. In the same example, the second sentence would be "pharmacologic substance r31 may treat r31 sign or symptom" or "drug and chemical r31 may treat r31 disease and disorder", where "pharmacologic substance" and "sign or symptom" are the semantic types of "ibuprofen" and "fever", and "drug and chemical" and "disease and disorder" are the semantic groups of "ibuprofen" and "fever". We call these models ClinicalBERT-EE-RI-CT, ClinicalBERT-EE-RI-ST and ClinicalBERT-EE-RI-SG where "RI" stands for "relation indicator", "CT" for "concept", "ST" for "semantic type" and "SG" for "semantic group".

ClinicalBERT with Entity Definition: In this method, we fine-tune BERT not only with input sentences but also with the text descriptions of the two entities. For entities in an input sentence, we extract their corresponding concept definitions from UMLS. They are used as the input to BERT to get concept embeddings (ClinicalBERT-EE-ED-CT). We can also generate semantic type embeddings using its definitions (ClinicalBERT-EE-ED-ST). These definitions are fed to a separate BERT model as input. Text representations are extracted based on the special entity markers we inserted. We use the [CLS] token embeddings related to the concept and semantic type definitions as the concept and semantic type embeddings. These embeddings are concatenated with the text embeddings of the input sentence in relation classification. There are no definitions for semantic groups in UMLS.

ClinicalBERT-EE-KB: UMLS knowledge is infused into BERT by jointly optimizing both a knowledge graph objective and a masked language model objective. Jointly optimizing the two objectives can implicitly integrate knowledge from external knowledge graphs into language models. Here we adopt the pre-trained Clinical KB-BERT (Hao et al., 2020) in our analysis.

ClinicalBERT-EE-KB-MLM: In this method, we pre-train BERT with UMLS information with only the masked language model (MLM) objective. We use the abbreviations provided by UMLS to map a triple into a natural language sentence (e.g., generating a sentence like "fever may be treated by ibuprofen" based on the triple (fever, may_be_treated_by, ibuprofen). In this way, we can get a set of sentences based on the triples in UMLS. We have created a total of 1,613,019 UMLS sentences. We then fine-tune ClinicalBERT with these UMLS sentences using only the MLM objective. By transferring the knowledge graph into natural language texts, we fuse UMLS knowledge with BERT in the same representation space. We call this model ClinicalBERT-EE-KB-MLM.

Summary of Methods: Table 1 summarizes the methods we proposed to add domain knowledge to BERT. We characterize them along multiple dimensions: infusion stage, type of domain knowledge added, form of domain knowledge added and fusion methods.

Fusion stage: Domain knowledge fusion can happen (a) during BERT model training, which results in a BERT model that is aware of the domain information encoded in clinical notes or UMLS (BERT-train) and (b) during BERT prediction, where domain knowledge is combined with the input or output of BERT in relation classification (we call them BERT-PredIn or BERT-PredOut).

Knowledge type: Additional domain knowledge can be characterized into (a) domain text corpora such as clinical notes or PubMed publications (Text-Corpora), (b) UMLS concept, semantic type, semantic group, relation as well as UMLS triples with two entities and one relation (UMLS-CT, UMLS-ST, UMLS-SG, UMLS-RE, UMLS-triple) and (c) UMLS concept and semantic type definitions (UMLS-CTD and UMLS-STD).

Knowledge form: Before fusion, domain knowledge is transformed into (a) embedding features extracted from KGE or KGE with MLP fine turning (embedding) or (b) text which are either synthetic sentences generated from UMLS or entity descriptions extracted from UMLS (text) or (c) training objective where knowledge graph training objective is combined with the BERT training objective to fuse knowledge in BERT (Training-Obj).

Fusion method: Finally, in terms of fusion methods, we characterize them into (a) concatenation where BERT features are concatenated with knowledge graph features in relation classification, (b)

Method	Stage	Knowledge Type	Knowledge Form	Fusion Methods
ClinicalBERT-EE-KGE	BERT-PredOut	UMLS-(CT,ST,SG,RE)	Embedding	Concatenate
ClinicalBERT-EE-MLP	BERT-PredOut	UMLS-(CT,ST,SG,RE)	Embedding	Concatenate
ClinicalBERT-EE-RI-CT	BERT-PredIn	UMLS-(CT,RE)	Text	BERT-Fuse
ClinicalBERT-EE-RI-ST	BERT-PredIn	UMLS-(CT,RE)	Text	BERT-Fuse
ClinicalBERT-EE-RI-SG	BERT-PredIn	UMLS-(SG,RE)	Text	BERT-Fuse
ClinicalBERT-EE-ED-CT	BERT-PredIn	UMLS-CTD	Text	BERT-Fuse
ClinicalBERT-EE-ED-ST	BERT-PredIn	UMLS-STD	Text	BERT-Fuse
ClinicalBERT-EE-KB	BERT-train	UMLS-Triple	Training-Obj	Joint-Opt
ClinicalBERT-EE-KB-MLM	BERT-train	UMLS-Triple	Text	BERT-Tune

Table 1: Summary of the knowledge fusion methods

joint optimization with both BERT and knowledge graph objectives (Joint-Opt), (c) BERT fine-tuning on sentences synthesized from UMLS using only the BERT training objective (BERT-Tune) and (d) BERT-fusion where additional domain knowledge is provided as the second sentence to BERT so that BERT itself becomes the fusion mechanism to combine domain knowledge with each input sentence (BERT-Fuse).

4 Experiments and Results

4.1 Dataset Description

For this study, we use the clinical relation extraction dataset from the 2010 i2b2/VA on Natural Language Processing Challenges for Clinical Records. This dataset contains discharge summaries and progress reports from different healthcare providers. The relation extraction task is to identify nine target relations between three types of medical concepts: treatments, problems and tests. The dataset used during the 2010 i2b2/VA challenge includes a total of 394 training reports, 477 test reports, and 877 un-annotated reports. After the challenge, however, only a part of the data was publicly released. The dataset we downloaded from the i2b2 website[1] only includes 170 training and 256 testing documents. Descriptions and statistics of the target relations can be found in table 2.

4.2 Experiment Details

In all the BERT-based classifiers, we use both the BERT sentence embedding and entity embedding (EE). We train all classifiers for 5 epochs with a learning rate of 0.00002 and a batch size of 8 with a softmax classification layer.

In addition, for ClinicalBERT-EE-KGE we concatenate 700 dimensional KGE with 768 dimensional BERT text representations. This 1468 dimensional vector is used as the input to the softmax layer. To implement ClinicalBERT-EE-MLP, we first train a MLP with KGE as the input. We use a single hidden layer consisting of 128 hidden units, a tanh activation function and a final linear layer with a softmax function to make predictions. We train this model for 50 epochs with a learning rate of 0.001 and a batch size of 32. After the MLP model is trained, we combine the output of the hidden layer (a 128 dimensional vector) with the BERT text embeddings (a 768 dimensional vector). This combined vector is connected to a linear layer and a softmax function to make predictions. In ClinicalBERT-EE-RI-CT/ST/SG, we employ different variations of the second input sentence. Text embedding is created by combining sentence embedding, entity embedding and relation indicator embedding. We connect the combined embeddings to a fully connected layer. In addition, we combine sentence embedding (768 dimensional vector) with two concept embedding (each a 768 dimensional vector) to create a 2304 dimensional input vector in ClinicalBERT-EE-ED-CT/ST. ClinicalBERT-EE-KB uses only the 768 dimensional text embedding. We pre-train ClinicalBERT-EE-KB-MLM with 4.4 million token text created from the UMLS knowledge graph with 425K steps and a learning rate of 0.00002. We initialized this model with weights from ClinicalBERT.

4.3 Baseline Methods

To compare with the current state-of-the-art, we consider systems that employ the same number of training instances and define the classification task with the same granularity level as ours. So far, we have found only two existing systems (Li et al., 2019; Hasan et al., 2020) meeting these criteria. In addition, (He et al., 2020) and (Weinzierl et al., 2020) also integrate UMLS knowledge into BERT

[1] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Relation Type	Total Sentences
Treatment improve or cure medical problem (**TrIP**)	203
Treatment worsen medical problem (**TrWP**)	133
Treatment caused medical problems (**TrCP**)	526
Treatment administered medical problem (**TrAP**)	2617
Treatment was not administered because of medical problem (**TrNAP**)	174
Test reveal medical problem (**TeRP**)	3053
Test conducted to investigate medical problem (**TeCP**)	504
Medical problem indicates medical problems (**PIP**)	2203
No Relation (**None**)	19932

Table 2: Statistics of the relation extraction dataset from the 2010 i2b2/VA challenge

for this task. Their reported weighted F1 scores are 0.747 and 0.782 respectively. But we did not include them in table 3 because either the total number of relations considered or the classification granularity (number of relation classes) are different from ours. To systematically investigate the effectiveness of different knowledge fusion methods, we consider multiple baselines. To evaluate the advantages of text embeddings generated from BERT over static embedding models (e.g. Word2Vec, Doc2Vec), we include baselines with static embeddings as well. We train a Word2Vec (Mikolov et al., 2013) and Doc2Vec (Le and Mikolov, 2014) model using the MIMIC-III clinical corpus (Johnson et al., 2016) plus the sentences in the i2b2 dataset. We implemented the following baselines with static text embeddings.

- (Word2Vec+biLSTM) : To generate sentence representations, we use a bidirectional LSTM to aggregate word embeddings in a sentences. This sentence representation is fed into a fully connected layer with ReLu activation and a linear layer with softmax activation.
- KGE+Word2Vec+biLSTM: Here we concatenate pre-trained knowledge graph embeddings with sentence embeddings generated by Word2Vec+biLSTM. These combined embeddings are used for classification.
- Doc2Vec : We use Doc2Vec generated sentence representations for classification.
- KGE+Doc2Vec : We combine Doc2Vec sentence presentations with pre-trained knowledge graph embeddings for classification.

To show the impact of domain text, we consider a baseline model where we use text embedding from BERT trained on general domain text (BERT-EE). We also consider a baseline where we only use sentence representations from BERT (BERT). This is to show the advantage of incorporating entity embeddings.

4.4 Performance Evaluation

In this section, we evaluate the effectiveness of different methods in incorporating additional knowledge into BERT. We use a 80%-20% train and test split and report the average result over multiple runs for each model. We calculate per class (9 class) and weighted F1 scores. The results of all the models can be found in table 3.

Quality of Text embedding: To investigate the quality of different types of sentence embedding techniques, we first compare the results from Word2Vec+biLSTM/Doc2Vec and BERT-base. Here sentence embedding from BERT-base (BERT) achieved significant improvement (3.96% and 29.91%) over sentence representations learned from Word2Vec+biLSTM and Doc2Vec. Moreover, entity informed text representation from BERT-base (BERT-EE) achieved an impressive 24.57% performance boost over a model that used only the BERT sentence embedding (BERT).

Impact of domain text: To see the effect of domain text, we use ClinicalBERT. Since entity-informed text representation achieves high performance, we continue to use that. Our result shows that ClinicalBERT-EE improves the performance by 0.97% over a general domain BERT-EE model.

Impact of UMLS knowledge: To demonstrate the impact of additional UMLS knowledge, we combine knowledge graph embeddings with both static text embeddings and ClinicalBERT text embeddings. While combined with Word2Vec and Doc2Vec embeddings, UMLS information results in a 10.23% and a 15.81% performance gain over

Model Name	No Relation	PIP	TeCP	TeRP	TrAP	TrCP	TrIP	TrNAP	TrWP	9 Class F1 Score
(Li et al., 2019)	N/A	0.6333	0.6117	0.8444	0.7974	0.6213	0.6159	0.4227	0.4457	0.7434
(Hasan et al., 2020)	0.9275	0.7896	0.6437	0.8685	0.8057	0.6320	0.5000	0.4025	0.2262	0.8808
Word2Vec-biLSTM	0.8398	0.2924	0.2047	0.4536	0.5334	0.3298	0.1453	0.0621	0	0.6979
KGE-Word2Vec-biLSTM	0.8610	0.4554	0.4374	0.6939	0.64040	0.4604	0.4141	0.3537	0.0719	0.7693
Doc2Vec	0.8057	0.0362	0	0.0165	0.0299	0	0	0	0	0.5585
KGE-Doc2Vec	0.8277	0.0091	0	0.5434	0.2604	0.0689	0	0	0	0.6468
BERT	0.8529	0.3496	0.4508	0.4059	0.5487	0.5229	0.4827	0.6153	0.4878	0.7256
BERT-EE	0.94252	0.8005	0.6870	0.8940	0.8438	0.7109	0.7059	0.7288	0.5053	0.9039
ClinicalBERT-EE	0.9473	0.8085	0.7102	0.9045	0.8654	0.7559	0.6935	0.7469	0.5304	0.9127
ClinicalBERT-EE-KGE	0.9486	0.8149	0.7373	0.9058	0.8693	0.7671	0.7027	0.7945	0.51851	0.9162
ClinicalBERT-EE-MLP	0.9459	0.8209	0.7019	0.9008	0.8616	0.7823	0.7449	0.7177	0.5087	0.9124
ClinicalBERT-EE-RI-CT	0.9456	0.8277	0.7115	0.9003	0.8684	0.7745	0.7489	0.7490	0.4685	0.9133
ClinicalBERT-EE-RI-ST	**0.9490**	0.8270	0.7358	**0.9146**	0.8691	**0.7980**	0.7218	0.7955	0.4835	**0.9181**
ClinicalBERT-EE-RI-SG	0.9469	0.8290	0.7195	0.9018	0.8605	0.7740	0.7331	0.7404	0.5460	0.9140
ClinicalBERT-EE-ED-CT	0.9449	0.8205	0.7357	0.9074	**0.8721**	0.7081	**0.7567**	0.7415	**0.6037**	0.9137
ClinicalBERT-EE-ED-ST	0.9439	0.8128	0.7113	0.8994	0.8563	0.7692	0.7123	0.7536	0.5806	0.9107
ClinicalBERT-EE-KB	0.9473	**0.8301**	**0.7429**	0.9102	0.8792	0.7821	0.7435	**0.8195**	0.5094	0.9177
ClinicalBERT-EE-KB-MLM	0.9425	0.8211	0.6967	0.8947	0.8597	0.7112	0.7446	0.7491	0.5150	0.9078
Support	19932	2203	504	3053	2617	526	203	174	133	29345

Table 3: Overall System Performance

Word2Vec-biLSTM and Doc2Vec, ClinicalBERT-EE-KGE (F1=0.9162) provides a 0.38% increase in performance over ClinicalBERT-EE (F1=0.9127). However, ClinicalBERT-EE-MLP (F1=0.9124) did not perform as well. We hypothesize that important information is lost when the 128 dimensional hidden layer vectors are used (versus the 700 dimensional knowledge graph embedding vectors). Next, we try to inject knowledge graph information directly into BERT input. Out of the three variations, ClinicalBERT-EE-RI-ST (with semantic type information and relation indicator in the second input) performed the best. This is our overall best performing model, achieving an F1-score of 0.9181. The relation indicator predicted by KGE may play an important role to boost the performance. Adding domain knowledge provides 0.59% improvement over ClinicalBERT-EE. From the results of ClinicalBERT-EE-ED-CT and ClinicalBERT-EE-ED-ST we can see that concept definitions and semantic type definitions did not help much. We hypothesize that the entity embeddings learned from BERT or the concept embeddings from KGE may be more precise than the embeddings learned from their text definitions. Next, we move on to pretrain BERT with knowledge graph information. Here we can see that ClinicalBERT-EE-KB is our second-best performing model with an F1-score of 0.9177. Finally, when we pre-train BERT with knowledge graphs using a Masked Language Model objective, we see that result (F1=0.9101) slightly went down compared to ClinicalBERT-EE (F=0.9127). This indicates that incorporating knowledge into BERT using knowledge graph objective may be more efficient than injecting UMLS sentences with a language model objective.

5 Conclusions

In this research, we have explored a wide range of techniques to incorporate the bio-medical knowledge base UMLS into BERT for clinical relation extraction. Based on our results, we found that (a) locating, extracting and adding entity embeddings from BERT is highly effective for relation extraction (24.57% improvement); (b) general-domain BERT with entity embedding achieved very high performance for clinical relation extraction (0.9039 F1-score); (c) adding domain-specific information such as domain text (in ClinicalBERT) or UMLSdomain knowledge (in ClinicalBERT-EE-RI-ST) only results in moderate performance gain (0.97% increase for adding domain text, an additional 0.59% increase for adding UMLS and total 1.56% for adding both); (d) the most effective method to fuse UMLS knowledge into BERT is BERT itself. The best performing model ClinicalBERT-EE-RI-ST, transforms a corresponding triplet inferred from UMLS into a natural language sentence, which is added as the second sentence to BERT.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*, pages 1–9.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2010. Nrc at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the NAACL-HLT, Volume 1*, pages 4171–4186.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Boran Hao, Henghui Zhu, and Ioannis Paschalidis. 2020. Enhancing clinical bert embedding using a biomedical knowledge base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661.

Fatema Hasan, Arpita Roy, and Shimei Pan. 2020. Integrating text embedding with traditional nlp features for clinical relation extraction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 418–425. IEEE.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. BERT-MK: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290, Online. Association for Computational Linguistics.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124(3):1907–1922.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced bert pre-training for chinese ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196.

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

Zhiheng Li, Zhihao Yang, Chen Shen, Jun Xu, Yaoyun Zhang, and Hua Xu. 2019. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. *BMC medical informatics and decision making*, 19(1):22.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901–2908.

Yuan Luo, Yu Cheng, Özlem Uzuner, Peter Szolovits, and Justin Starren. 2018. Segment convolutional neural networks (seg-cnns) for classifying relations in clinical notes. *JAMIA*, 25(1):93–98.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, USA*, pages 8449–8456. AAAI Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Anne-Lyse Minard, Anne-Laure Ligozat, and Brigitte Grau. 2011. Multi-class SVM for relation extraction from clinical reports. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, Hissar, Bulgaria. Association for Computational Linguistics.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. How additional knowledge can improve natural language commonsense question answering? *arXiv preprint arXiv:1909.08855*.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2017. A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv preprint arXiv:1712.02121*.

Dhanachandra Ningthoujam, Shweta Yadav, Pushpak Bhattacharyya, and Asif Ekbal. 2019. Relation extraction between the clinical entities based on the shortest dependency path based lstm. *arXiv preprint arXiv:1903.09941*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019a. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019b. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sunil Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, Berlin, Germany.

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

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080. PMLR.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. 2019. Relation extraction from clinical narratives using pre-trained language models. In *AMIA Annual Symposium Proceedings*, volume 2019, page 1236. American Medical Informatics Association.

Maxwell A Weinzierl, Ramon Maldonado, and Sanda M Harabagiu. 2020. The impact of learning unified medical language system knowledge embeddings in relation extraction from biomedical texts. *Journal of the American Medical Informatics Association*, 27(10):1556–1567.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2361–2364.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.