UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus

George Michalopoulos[1], Yuanxin Wang[1], Hussam Kaka[1], Helen Chen[1], Alex Wong[1]
[1]University of Waterloo
{gmichalo, yuanxin.wang, hussam.kaka, helen.chen, alexander.wong}@uwaterloo.ca

Abstract

Contextual word embedding models, such as BioBERT and Bio_ClinicalBERT, have achieved state-of-the-art results in biomedical natural language processing tasks by focusing their pre-training process on domain-specific corpora. However, such models do not take into consideration expert domain knowledge.

In this work, we introduced UmlsBERT, a contextual embedding model that integrates domain knowledge during the pre-training process via a novel knowledge augmentation strategy. More specifically, the augmentation on UmlsBERT with the Unified Medical Language System (UMLS) Metathesaurus was performed in two ways: i) connecting words that have the same underlying 'concept' in UMLS, and ii) leveraging semantic group knowledge in UMLS to create clinically meaningful input embeddings. By applying these two strategies, UmlsBERT can encode clinical domain knowledge into word embeddings and outperform existing domain-specific models on common named-entity recognition (NER) and clinical natural language inference clinical NLP tasks.

1 Introduction

In recent years, the volume of data being collected in healthcare has become enormous. Consequently, in order to use this vast amount of data, advanced Natural Language Processing (NLP) models are needed. This has led to the creation of highly-performing optimized NLP models focused on the biomedical domain.

Contextual word embedding models, such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019) have achieved state-of-art results in many NLP tasks. Initially tested in a general domain, these models have also been successfully applied in the biomedical domain by pre-training them on biomedical corpora, leading to state of the art performance in a variety of biomedical NLP tasks (Lee et al. 2019), (Alsentzer et al. 2019). However, current biomedical applications of transformer-based Natural Language Understanding (NLU) models have yet to incorporate expert domain knowledge into their embedding pre-training process.

The Unified Medical Language System (UMLS) (Bodenreider 2004) metathesaurus is a compendium of many biomedical vocabularies with associated information such as synonyms and hierarchical groupings. It allows for the connection of words that represent the same underlying 'concept'. For example, the words 'lungs' and 'pulmonary' share a similar meaning and can thus be mapped to the same concept unique identifier (CUI) *CUI: C0024109*. More specifically, UMLS uses the CUI in order to connect all the terms from all the source vocabularies that have the same clinical meaning. Additionally, UMLS also allows grouping of concepts according to their semantic types (McCray, Burgun, and Bodenreider 2001). For example, 'heart' and 'liver' are clustered to the 'ANATOMY' group and 'mass' and 'bleeding' are clustered to the 'DISORDER' group.

In this paper, we presented and publicly released[1] a novel architecture for augmenting contextual embeddings with clinical domain knowledge. Specifically:

1. We are the first, to the best of our knowledge, to propose the usage of domain (clinical) knowledge from a clinical Metathesaurus in the pre-training phase of a BERT-based model (UmlsBERT) in order to build 'semantically enriched' contextual representations that will benefit from both the contextual learning (BERT architecture) and domain knowledge (UMLS metathesaurus) .
2. We proposed a new multi-label loss function for the pre-training of the Masked Language Modelling (Masked LM) task of UmlsBERT that considers the connections between clinical words using the CUI attribute of UMLS.
3. We introduced a semantic group embedding that enriches the input embeddings process of UmlsBERT by forcing the model to take into consideration the association of the words that are part of the same semantic group.
4. Finally, we demonstrated that UmlsBERT outperforms two popular clinical-based BERT models (BioBERT and Bio_ClinicalBERT) and a general domain BERT model on different clinical named-entity recognition (NER) tasks and on one clinical natural language inference task.

The rest of paper is organized as follows. Related work is presented in section 2. The data that were used to pre-train and test the new UmlsBERT are described in section 3. The characteristics of the proposed UmlsBERT architecture for augmenting contextual embeddings with clinical knowledge are detailed in section 4. Finally, the results of the downstream tasks and the qualitative analysis are reported in section 5 and a conclusion and a plan for future work are presented in section 6.

[1] https://github.com/gmichalo/UmlsBERT

2 Related Work

2.1 Contextual Word Embeddings

Traditional word embedding methods such as word2vec (Mikolov et al. 2013) and FastText (Bojanowski et al. 2016) produce a constant context-independent vector representation for each word. Therefore, they cannot distinguish between the different meanings that a word might have in a given context.

In (Peters et al. 2018), contextualized word embeddings were introduced in a bidirectional language model (ELMo). This allowed the model to change the embedding of a word based on its imputed meaning, which was derived from the surrounding context. It enabled the model to achieve state of the art results in different NLP tasks. Subsequently, (Devlin et al. 2019) proposed the Bidirectional Encoder Representations from Transformers (BERT) which used bidirectional transformers (Vaswani et al. 2017) to create context-dependent representations that led to further performance improvements. For both models, pre-training was done on massive corpora and the produced context-sensitive embeddings can be used for downstream tasks. Furthermore, the BERT model can be further fine-tuned for a downstream task by being integrated to a task-specific architecture.

Other approaches tried to enhance the BERT architecture by injecting external knowledge from a knowledge base. Sense-BERT (Levine et al. 2020) is pre-trained to predict the supersenses (semantic class) of each word by incorporating lexical semantics (from the lexical database WordNet (Miller 1995)) into the model's pre-training objective and by adding supersense information to the input embedding. GlossBERT (Huang et al. 2019) focused on improving word sense disambiguation by using context-gloss pairs on the standard sentence-pair classification task of a BERT model. Finally, LiBERT (Lauscher et al. 2019) uses the synonyms and direct hyponym-hypernym pairs knowledge to improve its performance by creating an additional task in the pre-training phase of the model for recognizing a semantic relation between an encoded word pair.

2.2 Contextual Clinical Embeddings

There have been multiple attempts to explore the performance of contextual models in the biomedical domain. BioBERT is a BERT-based model which was pretrained on both general (BooksCorpus and English Wikipedia) and biomedical corpora (PubMed abstracts and PubMed Central full-text articles) (Lee et al. 2019). The authors demonstrated that incorporating biomedical corpora in the pre-training process does indeed improve the performance of the model in down-stream biomedical tasks. This is likely due to the fact that medical corpora can contain expressions and terms that are not usually found in a general domain corpus (Habibi et al. 2017).

In addition, (Zhu, Paschalidis, and Tahmasebi 2018) and (Si et al. 2019) trained a ELMo and BERT language models, respectively, on clinical text in order to improve their performance on the i2b2 2010 (Uzuner et al. 2011) and 2012 (Sun, Rumshisky, and Uzuner 2013) tasks. Finally in (Alsentzer et al. 2019), the authors built and released Bio_ClinicalBERT by further pre-training BioBERT on clinical text of the MIMIC-III v1.4 database (Johnson et al. 2016). They showed that the usage of clinical specific contextual embeddings can be beneficial for the performance of a model on different clinical NLP down-stream tasks (e.g entity recognition tasks).

3 Data

The Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC-III) dataset (Johnson et al. 2016) was used for pre-training the UmlsBERT model. It consists of anonymized electronic medical records of over forty-thousand patients who were admitted to the intensive care units of the Beth Israel Deaconess Medical Center (Boston, MA, USA) between 2001 and 2012. In particular, UmlsBERT was trained on the **NOTEEVENTS** table, which contains 2,083,180 rows of physician and nursing patient notes and diagnostic test reports.

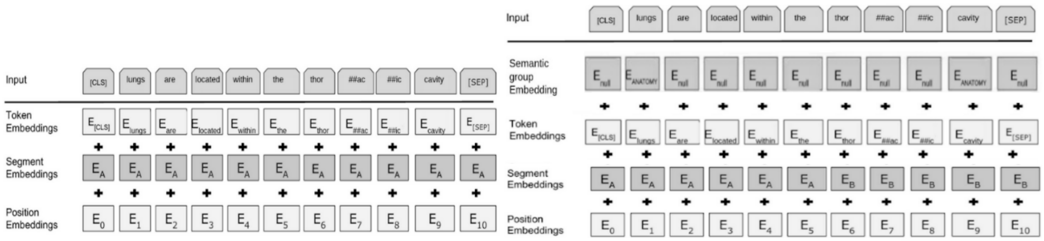
We evaluated the effect of the novel features of the UmlsBERT model on the MedNLI natural language inference task (Romanov and Shivade 2018) and on four i2b2 NER tasks (all in IOB task which was defined for the CoNLL-2003 shared task on the named-entity recognition (NER) task). In particular, we experimented on the following i2b2 tasks: i2b2 2006 de-identification challenge (Uzuner, Luo, and Szolovits 2007), i2b2 2010 concept extraction challenge (Uzuner et al. 2011), i2b2 2012 entity extraction challenge (Uzuner et al. 2011) and i2b2 2014 de-identification challenge (Stubbs, Kotfila, and Uzuner 2015). These datasets were chosen given their use in benchmarking prior biomedical BERT models, thereby allowing for result comparison. In addition, they are publicly available which enables reproducibility of our results, and meaningful comparison with future studies. Table 1 lists the statistics of all the datasets.

		# Sentences		
Dataset	Train	Dev	Test	classes
MedNLi	11232	1395	14 22	3
i2b2 2006	44392	5547	18095	17
i2b2 2010	14504	1809	27624	7
i2b2 2012	6624	820	5664	13
i2b2 2014	45232	5648	32586	43

Table 1: Number of sentences for the train/dev/test set of each dataset. We adopt the same splits that were used in (Alsentzer et al. 2019). We also included the number of classes for each dataset.

4 Methods

In this section, the proposed architecture for integrating UMLS-based features in the UmlsBERT's pre-training process is presented. In subsection 4.1, the Masked LM task of the BERT architecture is described. In subsection 4.2, this paper's contribution for incorporating clinical domain knowledge into the pre-trained process of a BERT model is explored. In particular, the methodology for enriching input embeddings with semantic group information and the new


Input
[CLS]
lungs
are
located
within
the
thor
##ac
##ic
cavity
[SEP]
Semantic group Embedding
E_{null} $E_{ANATOMY}$ E_{null} E_{null} E_{null} E_{null} E_{null} E_{null} E_{null} $E_{ANATOMY}$ E_{null}
Token Embeddings
$E_{[CLS]}$ E_{lungs} E_{are} $E_{located}$ E_{within} E_{the} E_{thor} $E_{\#\#ac}$ $E_{\#\#ic}$ E_{cavity} $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_A E_A E_A E_A E_A
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

(a) BERT

(b) UmlsBERT

Figure 1: Examples of: **(a)** Original input vector of BERT model (Devlin et al. 2019). **(b)** Augmented input vector of the UmlsBERT where the semantic group embeddings were available. For the words ‘chest’ and ‘cavity’, their word embeddings were enhanced with the embedding of the semantic group ‘ANATOMY’ $E_{anatomy}$. The rest of the words are not related to a medical term, so a zero-filled tensor E_{null} was used instead.

loss-function which is used for learning the connection of words through their corresponding CUI’s is analyzed.

4.1 BERT Model

The original BERT model (Devlin et al. 2019), based on multi-layer bidirectional transformers (Vaswani et al. 2017), can generate contextualized word representations. Incorporating information from bidirectional representations allows the BERT model to capture more accurately the meaning of a word based on its surrounding context (sentence).

The pre-training phase of the BERT model consists of two self-supervised tasks: Masked Language Modelling (LM), in which a percentage of the input is masked at random and the model is forced to predict the masked tokens, and Next Sentence Prediction, in which the model has to determine the positional relation between two sentences. Since UmlsBERT is focused on augmenting the Masked LM task with clinical information from the UMLS metathesaurus, we omitted the description of the Next Sentence Prediction task and focused on the details of the Masked LM task instead.

In Masked LM, 15% of the tokens of each sentence are replaced by a [MASK] token. For the j^{th} input token in the sentence, where $w_j \in \mathbb{R}^D$ is a 1-hot vector corresponding to it, an input embedding vector $u^{(j)}_{input}$ is created by the following equation:

$$u^{(j)}_{input} = p^{(j)} + SEGseg^{(j)}_{id} + Ew_j \tag{1}$$

where $p^{(j)} \in \mathbb{R}^d$ is the position embedding of the j token in the sentence and d is the transformer’s hidden dimension. Additionally, $SEG \in \mathbb{R}^{d\times 2}$ is called the segment embedding and $seg_{id} \in \mathbb{R}^2$ a 1-hot vector is the segment id and indicates the sentence to which the token belongs (in Masked LM, the model is using only one sentence hence the segment id indicates that all the tokens belong to the first sentence). Finally, $E \in \mathbb{R}^{d\times D}$ is the token embedding where D is the length of the model’s vocabulary.

The input embedding vectors are passed through multiple attention-based transformer layers where each layer produced a contextualized embedding of each token. Finally, for each masked token w, the model outputs a score vector $y \in \mathbb{R}^D$ with the goal of the model set to minimize the cross-entropy loss between the softmax of y_w and the 1-hot vector corresponding to the masked token (h_w):

$$loss = -log(\frac{exp(y_w[w])}{\sum_{w'} exp(y_w[w'])}) \tag{2}$$

4.2 Enhancing Contextual Embeddings with Clinical Knowledge

We updated the Masked LM procedure for training the language model of UmlsBERT to take into consideration the associations between the words in the UMLS database.

Semantic group embeddings Firstly, we introduced a new embedding matrix called $SG \in \mathbb{R}^{d\times D_s}$ into the input embedding of the BERT model, where d is BERT’s transformer hidden dimension and $D_s = 6$ is the number of unique UMLS semantic groups that could be identified in the vocabulary of our model. In particular, in this matrix, each row represents the unique semantic group in UMLS that a word can be identified with (for example the word ‘heart’ is associated with the semantic group ‘Anatomy’ in UMLS).

To incorporate the SG embedding matrix into the input embedding of our model, all the words that have a clinical meaning (in UMLS) are identified. In addition, for each word that is identified, its concept unique identifiers (CUI) and semantic group type were extracted. We used $s_w \in \mathbb{R}^{D_s}$ as a 1-hot vector corresponding to the semantic group of the medical word w. The identification of the UMLS terms and their UMLS semantic group was accomplished using the open-source Apache clinical Text Analysis and Knowledge Extraction System (cTakes) (Savova et al. 2010). Thus, by introducing the semantic group embedding, the input vector (equation 1) for each word was updated to:

$$u^{(j)\prime}_{input} = u^{(j)}_{input} + SGs_w \tag{3}$$

where the semantic group vector SGs_w was set to a zero-filled vector for words that were not identified in UMLS.

We hypothesized that incorporating into the input tensor the clinical information of the semantic groups is beneficial for the performance of the model. This is because semantic group vectors can enrich the input embeddings by forcing the embeddings of words that are associated with the same semantic group to become more similar (as we show in subsection 5.3). In addition, the semantic group representation can be used to enrich the input vector of words that were rare in the training corpus and thus the model did not have the chance to learn meaningful information for their representation. Figure 1 presented an overview of the insertion of the semantic group embeddings into the standard BERT architecture.

Updating the loss function of Masked LM task Secondly, we updated the loss function of the Masked LM pretraining task to take into consideration the connection between words that share the same CUI. As described in subsection 4.1, the loss function of the Masked LM pre-training task of a BERT model is cross-entropy loss between the softmax vector of the masked word and the 1-hot vector that indicates the actual masked word. We proposed to 'soften' the loss function and updated it to a multi-label scenario by using information from the CUIs.

More specifically, instead of using a 1-hot vector (h_w) that corresponds only to the masked word w, we used a binary vector indicating the presence of all the words which shared the same CUI of the masked word (h'_w). Finally, in order for the model to properly function in a multi-label scenario, the cross entropy loss (equation 2) was updated to a binary cross entropy loss:

$$loss = \sum_{i=0}^{D} -h_w^{'}[i]log(y_w[i]) + (1 - h_w^{'}[i])log(1 - y_w[i]) \tag{4}$$

These changes forced UmlsBERT to learn the underlying semantic relations between words (that are associated with the same CUI) in a biomedical context.

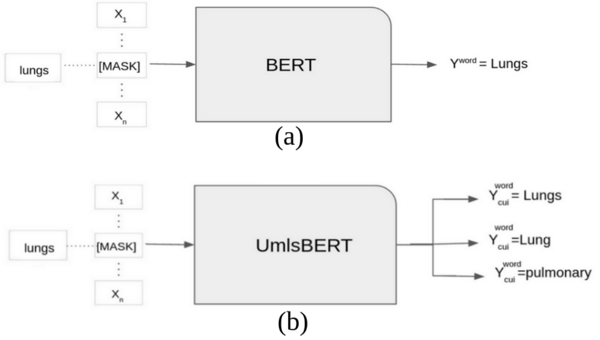


Figure 2: An example of predicting the masked word 'lungs' **(a)** the BERT model tries to predict only the word lungs **(b)** whereas the UmlsBERT tries to identify all words that were associated with the same CUI (e.g lungs, lung, pulmonary).

An example of predicting the masked word 'lungs' with and without the clinical information was presented in Figure 2. As seen in this figure, the UmlsBERT model tries to identify the words 'lung', 'lungs' and 'pulmonary' because all three words are associated with the same *CUI: C0024109* in the UMLS metathesaurus.

4.3 UmlsBERT Training

The conceptual outline of our system is as follow: Firstly, we initialized UmlsBERT with the pre-trained Bio_ClinicalBERT model (Alsentzer et al. 2019), and then, we further pre-trained it with the updated Masked LM task on MIMIC-III notes. Afterwards, to perform the downstream tasks, we added a single linear layer on top of UmlsBERT and 'fine-tuned' it to the task at hand, using either the associated embedding for each token or the embedding of the *[CLS]* token. The same fine-tuning method was applied to all other models used for comparison. In order to keep the experiment controlled, we used the same vocabulary and the WordPiece tokenization (Wu et al. 2016) across all the models. WordPiece divides words not in the vocabulary into frequent sub-words.

We did not experiment with a more complicated layer on top of UmlsBERT (such as the Bi-LSTM layer in (Si et al. 2019)), which might have hindered the performance of our model. This was because our goal was to demonstrate that incorporating domain knowledge was beneficial for the performance of the model, by showing that UmlsBert outperformed the other medical-based BERT models on a variety of medical NLP tasks (section 5).

Finally, it should be noted that we chose the UMLS metathesaurus in our process of augmenting the UmlsBERT model for two reasons: (i) The UMLS metathesaurus is a compendium of many popular biomedical vocabularies (e.g. MeSH (Dhammi and Kumar 2014) or ICD-10 (Organization 2004)) and by choosing to utilize the domain (medical) knowledge of UMLS, we are actually incorporating domain knowledge of the all the previous mention vocabularies. (ii) Our current work was focused on creating a clinical contextual embedding model capable of integrating domain (medical) knowledge. However we acknowledge that our method could even be applied to more general setting (e.g. by using WordNet (Miller 1995) which is a large lexical database of the English language)).

In the pre-training phase, UmlsBERT was trained for $150,000$ steps with batch size of 32, maximum sequence length of 128 and learning rate of $5 \cdot 10^{-5}$. All other hyperparameters were kept to their default values. Finally, it should be noted that the pre-training took 5 days using 2 nVidia V100 16GB GPU's with 224 GB of system RAM running Ubuntu 18.04.3 LTS.

5 Results

In this section, we presented the results of an empirical evaluation of the UmlBERT model that was proposed in section 4.2. In particular, we provided a comparison between different available BERT models in order to show the efficiency of our proposed model on different clinical NLP tasks. In addition, we conducted a qualitative analysis of the embedding of each model in order to illustrate how medical knowledge improves the quality of medical embeddings. Finally, we provided a visualized comparison of the embeddings of the words in each semantic group to show their effect on the creation of the input embedding of the UmlsBERT model.

Dataset		BERT$_{\text{based}}$	BioBERT	Bio_ClinicalBERT	UmlsBERT
MedNLI	Test Acc.	77.9 ± 0.6	**82.2 ±0.5**	81.2 ± 0.8	**82.2 ± 0.1**
	Val. Acc.	79.0 ± 0.5	83.2 ± 0.8	83.4 ± 0.9	83.1 ± 0.5
	Run. time(sec)	308	307	269	310
i2b2 2006	Test F1	**93.5 ± 1.4**	93.3 ± 1.3	93.1 ± 1.3	93.4 ± 1.2
	Val. F1	94.2 ± 0.6	93.8 ± 0.3	93.4 ± 0.2	93.9 ± 0.2
	Run. time(sec)	12508	12807	12729	13179
i2b2 2010	Test F1	85.2 ± 0.2	87.3 ± 0.1	87.7 ± 0.2	**88.3 ± 0.2**
	Val. F1	83.4 ± 0.3	85.2 ± 0.6	86.2 ± 0.2	87.5 ± 0.3
	Run. time(sec)	5325	5244	5279	5221
i2b2 2012	Test F1	76.5 ± 0.2	77.8 ± 0.2	78.9 ± 0.1	**79.3 ± 0.1**
	Val. F1	76.2 ± 0.7	78.1 ± 0.5	77.1 ± 0.4	77.1 ± 0.3
	Run. time(sec)	2413	2387	2403	2429
i2b2 2014	Test F1	**95.2 ± 0.1**	94.6 ± 0.2	94.3 ± 0.2	94.7 ± 0.1
	Val. F1	94.5 ± 0.4	93.9 ± 0.5	93.0 ± 0.3	93.8 ± 0.4
	Run. time(sec)	16738	17079	16643	16476

Table 2: Results of mean ± standard deviation of five runs from each model on the test and the validation test; best values are **bolded**; average running time is also provided for each model.

5.1 Downstream Clinical NLP Tasks

In this section, we reported the results of the comparison of our proposed model with the other BERT-based models on different downstream clinical NLP tasks that were described in section 3. All the BERT-based models were implemented using the transformers library (Wolf et al. 2019) on PyTorch 0.4.1. All experiments were executed on a Tesla P100 16.3 GB GPU with 32G GB of system RAM running Ubuntu 18.04.3 LTS.

Hyperparameter Tuning In order to address the concerns of the NLP community regarding reproducibility (Dodge et al. 2019), we provided the search strategy and the bound for each hyperparameter as follows: the batch size was set between 32 and 64, and the learning rate was chosen between the values 2e-5, 3e-5 and 5e-5. For the clinical NER tasks, we took a similar approach to (Lee et al. 2019) and set the number of training epochs to 20 to allow for maximal performance. The exception was for MedNLI, for which we trained the models on 3 and 4 epochs. The best values where chosen based on validation set F1 values (using the seqevals python framework for sequence labeling evaluation which was chosen for the fact that it can provide an evaluation of a named-entity recognition task on the entity-level)[2] for the i2b2 task and based on validation set accuracy (which is the standard metric for this task)[3] for MedNLI dataset.

In the interest of providing a fair comparison we also tuned the hyperparameters of each model in order to demonstrate its best performance in each task.

For the **MedNLI** dataset, we selected: 4 epochs, a batch size of 16 and a learning rate of 5e-5 for BERT$_{\text{base}}$, 4 epochs, a batch size of 16 and a learning rate of 3e-5 for BioBERT, 4 epochs, a batch size of 32 and a learning rate of 3e-5 for Bio_ClinicalBERT and 4 epochs, a batch size of 16 and a learning rate of 3e-5 for UmlsBERT.

For the **i2b2 2006** dataset, we selected: 20 epochs, a batch size of 32 and a learning rate of 2e-5 for BERT$_{\text{base}}$, 20 epochs, a batch size of 16 and a learning rate of 2e-5 for BioBERT, 20 epochs, a batch size of 16 and a learning rate of 2e-5 for Bio_ClinicalBERT and 20 epochs, a batch size of 16 and a learning rate of 2e-5 for UmlsBERT.

For the **i2b2 2010** dataset, we selected: 20 epochs, a batch size of 16 and a learning rate of 3e-5 for BERT$_{\text{base}}$, 20 epochs, a batch size of 32 and a learning rate of 3e-5 for BioBERT, 20 epochs, a batch size of 32 and a learning rate of 5e-5 for Bio_ClinicalBERT and 20 epochs, a batch size of 32 and a learning rate of 5e-5 for UmlsBERT.

For the **i2b2 2012** dataset, we selected: 20 epochs, a batch size of 16 and a learning rate of 3e-5 for BERT$_{\text{base}}$, 20 epochs, a batch size of 32 and a learning rate of 3e-5 for BioBERT, 20 epochs, a batch size of 16 and a learning rate of 5e-5 for Bio_ClinicalBERT and 20 epochs, a batch size of 16 and a learning rate of 5e-5 for UmlsBERT.

For the **i2b2 2014** dataset, we selected: 20 epochs, a batch size of 16 and a learning rate of 2e-5 for BERT$_{\text{base}}$, 20 epochs, a batch size of 16 and a learning rate of 2e-5 for BioBERT, 20 epochs, a batch size of 32 and a learning rate of 5e-5 for Bio_ClinicalBERT and 20 epochs, a batch size of 16 and a learning rate of 3e-5 for UmlsBERT.

Finally, in order to achieve more robust results, we ran our model on five different (random) seeds (6809, 36275, 5317, 82958, 25368) and we provided the average scores and standard deviation for the testing and the validation set.

BERT-based model comparison The mean and standard deviation (SD) of the scores for our model and the other competing models on different NLP tasks are reported in Table 2. UmlsBERT was shown to achieve the best F1 score in two i2b2 tasks (2010, 2012) (88.3% and 79.3%) and the best accuracy on the MedNLI tasks (82.2%). As our model was initialized with Bio_ClinicalBERT model and was pretrained on the MIMIC-III dataset, it was not surprising that it did not achieve the best performance on i2b2 2006 and i2b2

[2] https://github.com/chakki-works/seqeval

[3] https://github.com/huggingface/transformers/blob/master/src/transformers/data/metrics/__init__.py

	ANATOMY		**DISORDER**		**GENERIC**	
	heart	kidney	mass	bleeding	school	war
$\text{BERT}_{\text{based}}$	chest	liver	masses	bleed	college	battle
	cardiac	lung	massive	sweating	university	conflict
BioBERT	cardiac	liver	masses	bleed	college	wartime
	liver	lung	weight	strokes	schooling	battle
Bio_ClinicalBERT	cardiac	liver	masses	bleed	college	warfare
	liver	lung	weight	bloody	university	wartime
UmlsBERT	cardiac	**Ren**	masses	bleed	college	warfare
	heartbeat	liver	**lump**	**hem**	university	wartime

Table 3: The 2 nearest neighbors for 6 words in three semantic categories (two clinical and one generic). It can be observed that only UmlsBERT could find word associations based on the CUIs of the UMLS Metathesaurus that have clinical meaning whereas in the generic category there were not large discrepancies between the behavior of the models.

2014 (The $\text{BERT}_{\text{base}}$ model achieved 93.5% on i2b2 2006 and 95.2% on i2b2 2014). This is probably due to the nature of the de-ID challenges which was described in details in the Bio_ClinicalBERT paper (Alsentzer et al. 2019). In summary, *protected health information* (PHI) are replaced with a unique sentinel 'PHI' markers in the MIMIC dataset but in the de-ID challenge datasets (i2b2 2006, i2b2 2014) the PHI are replaced with different synthetic masks. However, it should be noted that even in these tasks UmlsBERT outperformed the other biomedical BERT models. These results demonstrate that augmenting contextual embeddings through domain (biomedical) knowledge can indeed be beneficial for the model's performance in a variety of biomedical down-stream tasks.

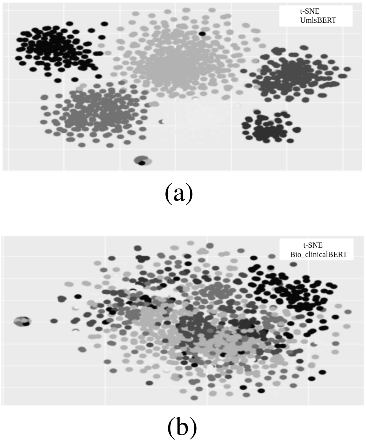
5.2 Qualitative Embedding Comparisons

Table 3 shows the nearest neighbors for 6 words from 3 semantic categories under UmlsBERT, Bio_ClinicalBERT, BioBERT and BERT. The first two categories ('ANATOMY' and 'DISORDER') were chosen to demonstrate the ability of the models to identify similar words in a clinical context and the third category ('GENERIC') was used to validate that the medical-focus BERT models can still find meaningful associations between words in general domain even if they were trained on medical-domain text datasets. This analysis demonstrated that augmenting the contextual embedding of UmlsBERT with Clinical Metathesaurus (UMLS) information was indeed beneficial for discovering associations between words with similar meanings in a clinical context. For instance, only UmlsBERT found the connection between 'kidney' and 'ren' (from the latin word 'renes', which means kidneys), between 'mass' and 'lump', and between 'bleeding' and 'hem' (a commonly used term to refer to blood). These associations were the result of changing the nature of the Masked LM training phase of UmlsBERT to a multi-label scenario by connecting different words which share a common CUI in UMLS. For the previously mentioned examples, 'kidney' and 'ren' have *CUI: C0022646*, 'mass' and 'lump' have *CUI: C0577559* and 'bleeding' and 'hem' have *CUI:C0019080*.

It should also be noted that for the term 'heart', BioBERT and Bio_ClinicalBERT identified the term 'liver' as its second nearest neighbor but UmlsBERT identified the term 'heartbeat'. Finally, the results in the generic list of words indicated that the medical-focused BERT models did not trade their ability to find meaningful associations in a general domain in order to be more precise in a clinical context as there was no meaningful difference observed in the list of neighbour words that the four models identified.

5.3 Semantic Group Embedding Visualization

As we described in subsection 4.2, one of the main novelties of UmlsBERT was the introduction of the semantic group embedding in the input embedding of the BERT architecture.



(a)

(b)

Figure 3: t-SNE visualization of the input embedding of the BERT architecture **(a)** semantic group clustering of the UmlsBERT input embedding (word embedding + semantic group embedding) **(b)** semantic group clustering of the Bio_ClinicalBERT input embedding (word embedding).

By pre-training the UmlsBERT model (on MIMIC-III), we attained more meaningful word-embedding inputs where words in the same semantic group were deemed more similar to each other in the embedding space.

In Figure 3 we presented a t-SNE dimensionality reduction (van der Maaten and Hinton 2008) mapping comparison between Bio_ClinicalBERT and UmlsBERT. In particular, we compare the input embedding (word embedding) of Bio_ClinicalBERT with the input embedding (word embedding + semantic group embedding) of UmlsBERT for all the clinical terms that UMLS could identify in the standard BERT vocabulary. It is evident that the clear clustering according to the semantic group that exists in the UmlsBERT embeddings (Figure 3a) cannot be found in the Bio_ClinicalBERT embedding (Figure 3b). Thus we concluded that by augmenting the input layer of the BERT architecture, more meaningful input embeddings could be provided to the model, which was one of the reasons that our model surpassed all the previous medical BERT-based models in the downstream tasks (subsection 5.1).

6 Conclusion and Future Work

In this paper, we presented UmlsBERT, a novel BERT-based architecture that included domain (biomedical) knowledge into the pre-training process of a contextual word embeddings model. We demonstrated that UmlsBERT has the ability to learn the association of different clinical terms with similar meaning in the UMLS metathesaurus. The proposed UmlsBERT can also create more meaningful input embeddings by incorporating information from the semantic group of each (biomedical) word. Furthermore, we demonstrated that these modifications could improve the performance of the model as UmlsBERT outperformed other biomedical BERT models in various downstream tasks.

As for future work, we plan to extend our work by addressing some of our system's limitation including: (i) Examining the effect of augmenting contextual embeddings with medical knowledge when more complicated layers (for down-stream tasks) are used atop of the output embedding of UmlsBERT. (ii) Exploring UMLS associations that extend the six semantic groups that are currently used. We hypothesize this will be particularly beneficial for down-stream tasks such as automatic clinical coding tasks. (iii) Investigating the effect of injecting medical knowledge onto larger models, such as $\text{BERT}_{\text{large}}$ (Devlin et al. 2019).

References

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; and McDermott, M. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. Minneapolis, Minnesota, USA: Association for Computational Linguistics. doi:10.18653/v1/W19-1909. URL https://www.aclweb.org/anthology/W19-1909.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32(Database issue): D267–270.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5. doi:10.1162/tacl_a_00051.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Dhammi, I. K.; and Kumar, S. 2014. Medical subject headings (MeSH) terms. *Indian journal of orthopaedics vol.* 48,5. doi:doi:10.4103/0019-5413.139827.

Dodge, J.; Gururangan, S.; Card, D.; Schwartz, R.; and Smith, N. A. 2019. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of EMNLP*.

Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D. L.; and Leser, U. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 33(14): i37–i48. ISSN 1367-4803. doi:10.1093/bioinformatics/btx228. URL https://doi.org/10.1093/bioinformatics/btx228.

Huang, L.; Chi, S.; Qiu, X.; and Huang, X. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. 3500–3505. doi:10.18653/v1/D19-1355.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3: 160035.

Lauscher, A.; Vulic, I.; Ponti, E. M.; Korhonen, A.; and Glavas, G. 2019. Informing Unsupervised Pretraining with External Linguistic Knowledge. *ArXiv* abs/1909.02339.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4): 1234–1240. ISSN 1367-4803. doi:10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Levine, Y.; Lenz, B.; Dagan, O.; Ram, O.; Padnos, D.; Sharir, O.; Shalev-Shwartz, S.; Shashua, A.; and Shoham, Y. 2020. SenseBERT: Driving Some Sense into BERT. *ArXiv* abs/1908.05646.

McCray, A.; Burgun, A.; and Bodenreider, O. 2001. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. *Studies in health technology and informatics* 84: 216–20. doi:10.3233/978-1-60750-928-8-216.

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

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38(11): 39–41. ISSN 0001-0782. doi:10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

Organization, W. H. 2004. ICD-10 : international statistical classification of diseases and related health problems : tenth revision.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Romanov, A.; and Shivade, C. 2018. Lessons from Natural Language Inference in the Clinical Domain. 1586–1596. doi:10.18653/v1/D18-1187.

Savova, G. K.; Masanz, J. J.; Ogren, P. V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K. C.; and Chute, C. G. 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.

Si, Y.; Wang, J.; Xu, H.; and Roberts, K. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association : JAMIA* 26. doi:10.1093/jamia/ocz096.

Stubbs, A.; Kotfila, C.; and Uzuner, Ö. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of biomedical informatics* 58 Suppl: S11–9.

Sun, W.; Rumshisky, A.; and Uzuner, O. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association : JAMIA* 20. doi:10.1136/amiajnl-2013-001628.

Uzuner, O.; Luo, Y.; and Szolovits, P. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association : JAMIA* 14: 550–63. doi:10.1197/jamia.M2444.

Uzuner, O.; South, B.; Shen, S.; and DuVall, S. 2011. 2010 i2B2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA* 18: 552–6. doi:10.1136/amiajnl-2011-000203.

van der Maaten, L.; and Hinton, G. E. 2008. Visualizing Data using t-SNE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *ArXiv* abs/1706.03762.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G. S.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv* abs/1609.08144.

Zhu, H.; Paschalidis, I. C.; and Tahmasebi, A. 2018. Clinical Concept Extraction with Contextual Word Embedding. *ArXiv* abs/1810.10566.