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

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, Alex Wong

University of Waterloo, {gmichalo, yuanxin.wang, hussam.kaka, helen.chen, alexander.wong}@uwaterloo.ca

Summary

We introduced UmlsBERT, a contextual embedding model that integrates domain knowledge during its pre-training process via a novel knowledge augmentation strategy.
The augmentation on UmlsBERT with the Unified Medical Language System (UMLS) Metathesaurus was performed in two ways:

- connecting words that have the same underlying 'concept' (CUI) in UMLS
- leveraging semantic type knowledge in UMLS to create clinically meaningful input embeddings

UmlsBERT can encode clinical domain knowledge into word embeddings and outperform existing domain-specific models on common named-entity recognition (NER) and on the MedNLI natural language inference clinical tasks.

Introduction

Current biomedical applications of transformer-based models [1][2] have yet to incorporate structured expert domain knowledge from a knowledge-base into their embedding pre-training process.
We proposed the usage of clinical knowledge from the UMLS Metathesaurus, a compendium of many biomedical vocabularies (Figure 1), in the pre-training phase of a BERT-based model (UmlsBERT) in order to build 'semantically enriched' contextual representations.

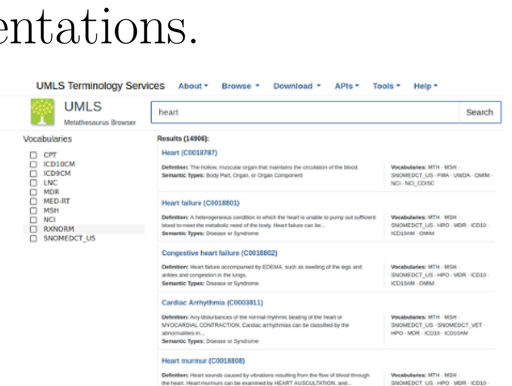

Figure 1:An example search of the word 'heart' in the UMLS Metathesaurus

UmlsBERT

Semantic type embeddings

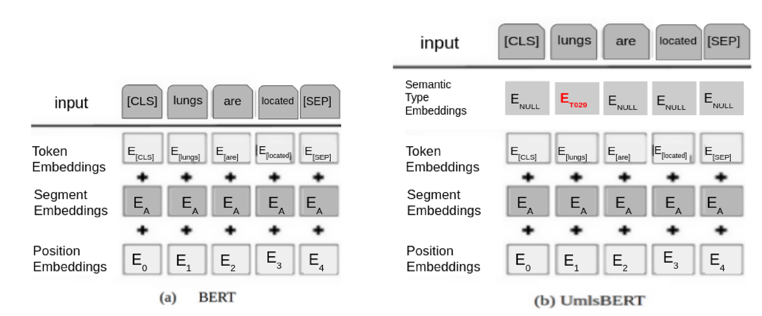

Figure 2:Examples of: **(a)** Original input vector of BERT model [3]. **(b)** Augmented input vector of the UmlsBERT where the semantic type embeddings were available.

Firstly, we introduced a new embedding matrix called $ST \in \mathbb{R}^{d \times D_s}$ into the input embedding of the BERT model, where d is BERT's transformer hidden dimension and $D_s = 44$ is the number of UMLS semantic types that could be identified in the vocabulary of our model (Figure 2).

Updating the loss function of Masked LM task

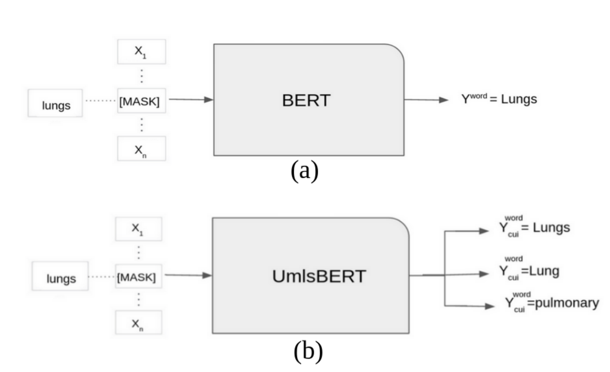

Figure 3:An example of predicting the masked word 'lungs' **(a)** the BERT model tries to predict only the word lungs **(b)** whereas the UmlsBERT tries to identify all words that were associated with the same **CUI:C0024109** (e.g lungs, lung, pulmonary).

Secondly, we updated the loss function of the Masked LM pre-training task to a binary cross entropy loss in order take into consideration the connection between words that share the same CUI (Figure 3).

Results

Dataset		$BERT_{based}$	BioBERT	Bio_ClinicalBERT	UmlsBERT
MedNLI	Test Ac.	77.9 ± 0.6	82.2 ±0.5	81.2 ± 0.8	**82.3** ± **0.1**
	Val. Ac.	79.0 ± 0.5	83.2 ± 0.8	83.4 ± 0.9	**83.8** ± **0.4**
	Run. time(sec)	308	307	269	305
	#parameters	108,312,579	108,312,579	108,312,579	108,346,371
i2b2 2006	Test F1	93.5 ± 1.4	93.3 ± 1.3	93.1 ± 1.3	**93.6** ± **0.5**
	Val. F1	94.2 ± 0.6	93.8 ± 0.3	93.4 ± 0.2	**94.4** ± **0.2**
	Run. time(sec)	12508	12807	12729	13167
	#parameters	108,322,576	108,322,576	108,322,576	108,356,368
i2b2 2010	Test F1	85.2 ± 0.2	87.3 ± 0.1	87.7 ± 0.2	**88.6** ± **0.1**
	Val. F1	83.4 ± 0.3	85.2 ± 0.6	86.2 ± 0.2	**87.7** ± **0.5**
	Run. time(sec)	5325	5244	5279	5219
	#parameters	108,315,655	108,315,655	108,315,655	108,349,447
i2b2 2012	Test F1	76.5 ± 0.2	77.8 ± 0.2	78.9 ± 0.1	**79.4** ± **0.1**
	Val. F1	76.2 ± 0.7	78.1 ± 0.5	77.1 ± 0.4	**78.3** ± **0.4**
	Run. time(sec)	2413	2387	2403	2432
	#parameters	108,320,269	108,320,269	108,320,269	108,354,061
i2b2 2014	Test F1	**95.2** ± **0.1**	94.6 ± 0.2	94.3 ± 0.2	94.9 ± 0.1
	Val. F1	**94.5** ± **0.4**	93.9 ± 0.5	93.0 ± 0.3	94.3 ± 0.5
	Run. time(sec)	16738	17079	16643	16554
	#parameters	108,343,339	108,343,339	108,343,339	108,377,131

Figure 4:Results of mean ± standard deviation of five runs from each model on the test and the validation test; we use the acronym Ac. for accuracy.

- UmlsBERT achieved the best results in 4 out of the 5 tasks (Figure 4).
- It achieved the best F1 score in three i2b2 tasks (**2006**, **2010** and **2012**) (93.6%, 88.6% and 79.4%) and the best accuracy on the **MedNLI** task (82.3%).

Qualitative Embedding Comparisons

	ANATOMY		**DISORDER**		**GENERIC**	
	feet	kidney	mass	bleeding	school	war
$BERT_{based}$	ft	liver	masses	bleed	college	battle
	foot	lung	massive	sweating	university	conflict
BioBERT	foot	liver	masses	bleed	college	wartime
	wrists	lung	weight	strokes	schooling	battle
Bio_ClinicalBERT	foot	liver	masses	bleed	college	warfare
	legs	lung	weight	bloody	university	wartime
UmlsBERT	foot	**Ren**	**lump**	bleed	college	warfare
	pedal	liver	masses	**hem**	students	military

Figure 5:The 2 nearest neighbors for 6 words in three semantic categories (two clinical and one generic).

Only UmlsBERT found the connections between the highlighted and the initial words. These associations were the result of changing the Masked LM training phase of UmlsBERT to a multi-label scenario by connecting different words which share a common CUI (Figure 5).

Semantic Type Embedding Visualization

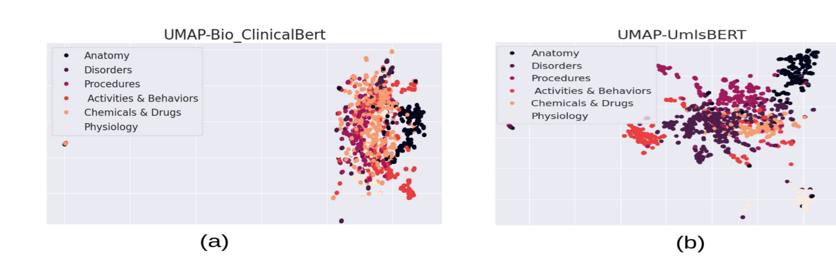

Figure 6:UMAP visualization of the clustering of the input embeddings **(a)** of Bio_ClinicalBert **(b)** of UmlsBert.

We observed that more meaningful input embeddings can be created, as the embeddings of the words, that are associated with the same semantic group, are forced to become more similar (Figure 6).

Conclusion

We presented UmlsBERT, a novel BERT-based architecture that included biomedical knowledge into its pre-training process. Our experiments demonstrated that including domain knowledge is beneficial for our model as it outperformed other biomedical BERT models in various downstream tasks.

Acknowledgements

We acknowledge the generous support from Microsoft AI for Health Program, MITACS Accelerate grant (IT19239), Semantic Health Inc., NSERC and Canada Research Chairs program.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott.Publicly Available Clinical BERT Embeddings. ClinicalNLP Workshop 2019.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.