Saama Research at MEDIQA 2019: Pre-trained BioBERT with Attention Visualisation for Medical Natural Language Inference

Kamal Raj Kanakarajan , Suriyadeepan Ramamoorthy, Vaidheeswaran Archana, Soham Chatterjee, Malaikannan Sankarasubbu*
SAAMA AI Research Lab, Chennai, India
{kamal.raj, suriyadeepan.ramamoorthy, archana.iyer, soham.chatterjee,
malaikannan.sankarasubbu}@saama.com

Abstract

Natural Language inference is the task of identifying relation between two sentences as entailment, contradiction or neutrality. MedNLI is a biomedical flavour of NLI for clinical domain. This paper explores the use of Bidirectional Encoder Representation from Transformer (BERT) for solving MedNLI. The proposed model, BERT pre-trained on PMC, PubMed and fine-tuned on MIMIC-III v1.4, achieves state of the art results on MedNLI (83.45%) and an accuracy of 78.5% in MEDIQA challenge. The authors present an analysis of the attention patterns that emerged as a result of training BERT on MedNLI using a visualization tool, bertviz.

1 Introduction

Natural Language Inference (NLI) is a fundamental task in Natural Language Processing in which the objective is to determine if the hypothesis is true (entailment), false (contradiction) or undetermined (neutral), given a premise. Entailment, Contradiction and Neutral (semantic independence) are semantic concepts that represent the relationship between sentences. The ability to infer these relations between sentences or pieces of text, is crucial in tasks like Information Retrieval, Semantic Parsing, Commonsense Reasoning, etc. NLI, like most NLP tasks, is challenging due to the ambiguous nature of natural language. A particular meaning can be expressed in multiple linguistic forms. This calls for methods that can capture meaningful semantic concepts from text.

Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) is a collection of sentence pairs labeled for entailment, contradiction, and semantic independence. It contains approximately 550,000 labeled hypothesis/premise pairs. Multi-Genre Natural Language Inference (Multi-NLI) corpus (Williams et al., 2017) contains 433,000 samples, covering a wide range of (10) genres of written and spoken English. Multi-NLI, in its complexity, is closer to Natural Language than SNLI.

MedNLI (Romanov and Shivade) is a dataset for natural language inference in clinical domain, analogous to SNLI. Romanov et al in (Romanov and Shivade), used InferSent (Conneau et al., 2017), a bidirectional LSTM based model for achieving an accuracy of 73.5% in MedNLI. In (Jin et al., 2019), Jin et al make use of BioBERT (Lee et al., 2019), a biomedical version of BERT along with pre-trained LMs(Language Models) as feature extractors, to achieve an accuracy of 81.7% on MedNLI.

This work uses BERT pre-trained on PMC and PubMed corpus, and fine-tuned on MIMIC-III v1.4 data (BioBERT) to solve MedNLI. This approach achieves new state of the art results when evaluated on MedNLI test set (83.4%). Evaluation on MEDIQA (Ben Abacha et al., 2019) test set (Shivade, 2019) results in an accuracy of 78.5 %.

2 Data

2.1 MedNLI

MedNLI or Medical Natural Language Inference is a publicly annotated dataset in the clinical domain. MedNLI was created as a NLI dataset comparable to SNLI, adjusted for the clinical domain (Table 1).

*Equal Contribution: Kamal had sole access to MIMIC and MEDIQA data, focussed on the algorithm development and implementation. Suriyadeepan and Archana focussed on the attention visualisation and writing. Soham and Malaikannan focussed on reviewing

Proceedings of the BioNLP 2019 workshop, pages 510–516
Florence, Italy, August 1, 2019.

#	Premise	Hypothesis	Label
1	ALT, AST and lactate were elevated as noted above	patient has abnormal fits	entailment
2	Chest x-ray showed mild congestive heart failure	The patient has complaints of cough	neutral
3	During Hospitalisation, patient became progressively more dyspnic requiring BiPaP and then a NRB	The patient is on room air	contradiction

Table 1: Examples from Development set of MedNLI

Dataset Size	
Training Pairs	11232
Development Pairs	1395
Test Pairs	1422
MEDIQA	405
Average Sentence Length in Token	
Premise	20.0
Hypothesis	5.8
Maximum Sentence Length in Tokens	
Premise	202
Hypothesis	20

Table 2: Data Statistics

2.2 Deriving from MIMIC-III v.1.4

While adapting the structure of SNLI, MedNLI derives its data from the MIMIC III v.1.4 dataset (Johnson et al., 2016). The MIMIC-III v.1.4 dataset consists of around 2,078,705 clinical notes written by healthcare professionals. These notes contain the de-identified records of 38,597 patients.

Annotations were done by two board-certified radiologists and two additional clinicians pursuing their residency programs.

2.3 Dataset Statistics

The MedNLI dataset used over 4 clinicians working on a total of 4,683 premises over a period of 6 weeks with 14,049 unique sentence pairs. The dataset was then split into training, development, and test sets. The class distribution is even across all classes, throughout training, development and test sets (Table 2).

2.4 MEDIQA Shared Task

MEDIQA(Ben Abacha et al., 2019) is a shared task which is part of BIONLP 2019. It was created by using an annotation technique similar to MedNLI. It serves as an additional test for the MedNLI data. It contains 405 premise-hypothesis pairs. These pairs were randomly sampled from records, segmented from *Past Medical History* section with a simple rule-based method.MedNLI train set is used to train the model and hyper parameter are tuned based MedNLI development and test set accuracy. MedNLI and MEDIQA test set follows the same label mapping.

3 BERT

3.1 Description

Bidirectional encoder representation from transformer (Vaswani et al., 2017) is a language representation model which performs on a wide range of NLP tasks such as question answering and language inference. The architecture of the BERT leverages the use of pre-trained deep bidirectional representations. Existing pre-trained language representations include feature-based (ELMO) (Peters et al., 2018) and fine-tuning approach (OpenAI GPT) (Radford et al., 2018) . However, these models are severely restricted due to their unidirectional nature. BERT uses masked language models to enable pre-trained deep bidirectional representations.

The BERT model the authors experimented with, is $BERT_{BASE}$. The model is composed of 12 transformer blocks with a hidden size of 768 and 12 attention heads. The feed-forward/filter size is 4 times the hidden size. For fine tuning on MedNLI, a classification layer is added and all the parameters of the final model are fine-tuned jointly as per the original paper (Devlin et al., 2018).

3.2 BERT on MedNLI

BERT displays a clear supremacy over contemporary architectures (Radford et al., 2018) (Peters et al., 2018) on several NLP tasks. BERT's use of bidirectional encoders is a characteristic feature that separates it from other architectures.

Natural language inference requires learning the relationship between two sentences which is not supported by naive language models. Thus, BERT which is pre-trained on binarized next sentence prediction is vital for NLI.

MedNLI is built based on GLUE (General Language Understanding Evaluation) dataset (Wang et al., 2018). The goal as of before, with inference is to predict how the first sentence is related to the next in terms of entailment, contradiction or neutral. MedNLI is a sequence level task. The model needs to learn a minimum number of parameters and is used with an additional output layer with BERT.

4 Experiments

All the experiments in this paper are done with BERT pre-trained on unlabelled biomedical data-BioBERT (Lee et al., 2019). Three pretrained models are available: One model is trained only on PubMed articles, one is trained on PMC articles and one trained on both PubMed and PMC articles.

BioBERT trained on PubMed articles was also finetuned with MIMIC III v1.4 III v1.4-III (Johnson et al., 2016) notes. MIMIC III v1.4-III is a de-identified biomedical corpus compared to PubMed articles. All the 18 HIPAA (Atchinson and Fox, 1997) identifiers are removed and masked with unique PHI (Protected Health Information) tags in MIMIC III v1.4. The reason for fine-tuning BERT on MIMIC III v1.4 is because the MedNLI is a small subset derived from MIMIC III v1.4 database. No special preprocessing for PHI elements present in MIMIC III v1.4 data was done. Furthermore, MedNLI training data also contain sentences with PHI mask similar to MIMIC III v1.4. Finally the fine-tuned model is trained on MedNLI dataset. Evaluation is performed on MedNLI test and dev sets. The trained model has also been evaluated on MEDIQA test set. The results are presented in table 4.

Fine tuning on BioBERT was done using TensorFlow with three GeForce GTX 1080Ti GPUs for 2 weeks. The model on MIMIC III v1.4 is trained with maximum sequence length 128 with batch size 32 and learning rate 2e-5 for 200,000 steps. The sequence length is limited such that it can fit into GPU memory. The pretraining data from MIMIC III v1.4 is prepared using scripts from the original BERT github repository (Devlin et al., 2018) with the default parameters. Further fine tuning on MedNLI task is done with one GeForce GTX 1080Ti GPU with 11 GB of RAM. One epoch on MedNLI takes around 3 minutes on a single GPU[1].

4.1 Hyperparameter Search

Learning Rate	2e-5,3e-5,5e-5 .
Max Sequence Length	128
Batch Size	16, 32
Warmup Proportion	0.1-0.3
Number of Epochs	3, 4, 5

Table 3: Hyperparameters

All of hyperparameter search is done with a fixed random seed of 42. Each iteration took an average of 3-4 minutes. A variant of Adam optimizer which selectively avoids applying weight decay to normalization layers, proposed in BERT(Devlin et al., 2018) paper is used. Only learning rate is tuned while all the other hyperparameters like β_1, β_2, L_2 weight decay are fixed at 0.9, 0.999 and 0.01 respectively.

4.2 Results

The results of the experiments with BERT pre-trained on PubMed, PMC and fine-tuned on MIMIC III v1.4, are tabulated in Table 4. Pre-training on PubMed and PMC gives similar results. Pretraining on both PubMed and PMC gives a slight increase in accuracy in both dev and test sets. Finally, BioBERT-MIMIC III v1.4, BERT pre-trained on PubMed, fine-tuned on MIMIC III v1.4 outperforms other models by roughly a 2% margin, and marks a new state-of-the-art for MedNLI. The same model when evaluating on MEDIQA (Ben Abacha et al., 2019) test set, gives an accuracy of 78.5%.

5 Visualizations

Vig et al (Vig, 2019) uses a visualization tool, bertviz (Vig, 2019), presents 6 patterns of attention observed in BERT. Each attention pattern as explained in (Vig, 2019), provides with intuition regarding the underlying mechanics of the model. In deep learning models which are notoriously

[1] The code is available at https://github.com/kamalkraj/biobert

Model	**Dev(%)**	**Test(%)**	**MEDIQA(%)**
BioBERT-PubMed	83.42	80.74	78.3
BioBERT-PMC	83.05	81.07	77.8
BioBERT-PubMed + PMC	83.22	81.92	78.1
BioBERT-MIMIC III v1.4	**85.16** **(SOTA)**	**83.45** **(SOTA)**	**78.5**

Table 4: Comparision of Results

known for their opaque nature, these intuitions offer a peek behind the curtains. bertviz, was subsequently used to visualize BioBERT-MIMIC III v1.4 before and after training on MedNLI task. In this section, some of the interesting patterns are presented which were observed by comparing and contrasting attention patterns before and after fine-tuning on MedNLI task. The distinct patterns that emerge from fine-tuning are heavily dependent on the nature of the task (NLI).

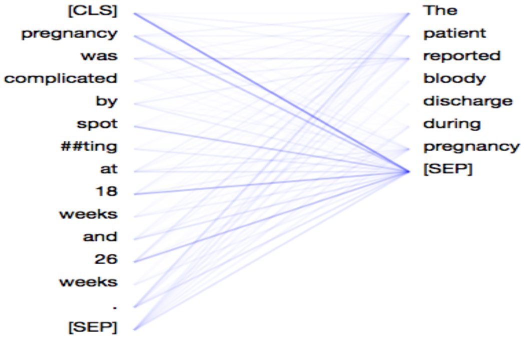

Figure 1: Distribution of Attention-Before

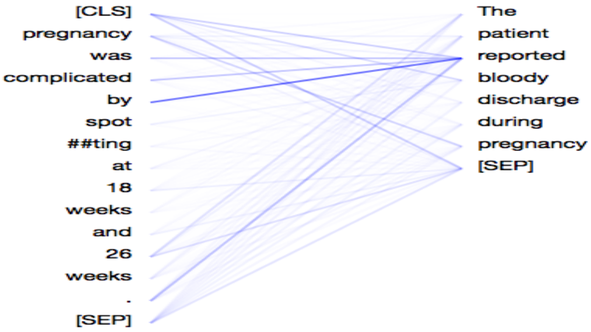

Figure 2: Distribution of Attention-After

1. **Distribution of Attention** Before training, majority of the attention is focused on the delimiter token of the second sentence [SEP], as seen in figure 1. After fine-tuning on MedNLI, the attention is distributed all over the second sentence as observed in figure 2. The dense connections seen in the figure, could be perceived as a natural consequence of fine-tuning the network on a NLI task, where establishing connections between two sentences is crucial.

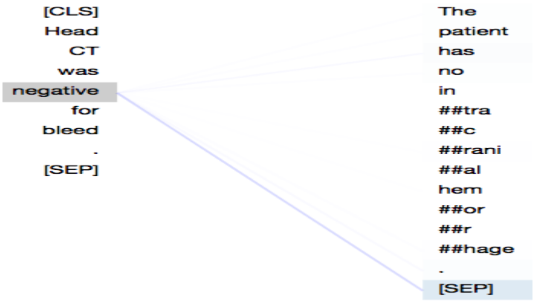

Figure 3: Word Similarity-Before

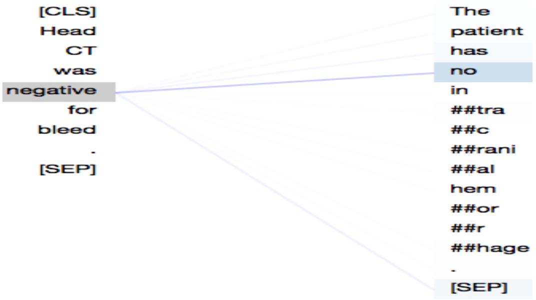

Figure 4: Word Similarity-After

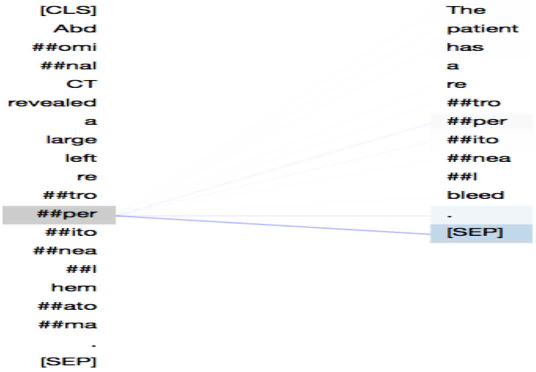

Figure 5: Tokenized Words-Before

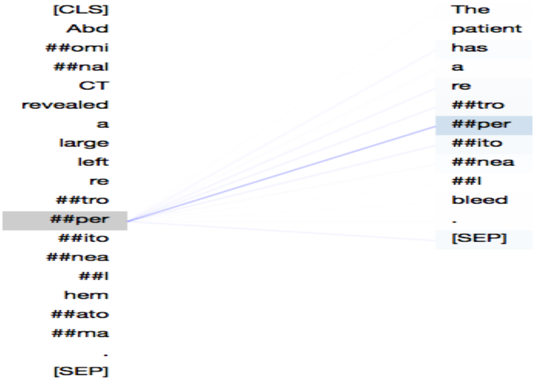

Figure 6: Tokenised Words-After

2. **Word Similarity** It can be observed that

words similar to source word gets more attention. Notice the words *negative* and *no*, expressing similar sentiments (negative), connected via attention flow in figure 4. Word-level similarity, although not always, is a good indicator of entailment. Upon encountering sentences with similar words, it is reasonable for a network to be biased towards entailment.

3. **Tokenized Words** In BERT, OOV (Out of Vocabulary) words are identified and split into segments. This way, the morphological information is maintained, which comes in handy in tasks such as textual entailment where word-level similarity is an important aspect to notice. Before fine-tuning, the OOV (Out of Vocabulary) words split into multiple tokens receive weak attention from source tokens, as observed in figure 5. After fine-tuning on MedNLI, a strong attention flow between the tokenized words across two sentences can be seen. As mentioned above, these connections as seen in figure 6, help in identifying word-level similarity between sentences.

The authors have presented a error analysis study based on attention patterns in Appendix A. Based on the intuitions gained from error analysis, the authors propose a list of changes that could improve the performance of the model. A limitations of the proposed approach and a list of possible improvements are presented in Appendix B.

6 Conclusion

In this paper, a variant of BERT, fine-tuned on MIMIC III v1.4, is proposed to solve the task of MedNLI, a Natural Language Inference task designed for clinical domain. The experiments include evaluation of BERT pre-trained on PMC, PubMed and MIMIC III v1.4, on MedNLI test and dev sets, and MEDIQA test set for MedNLI. State-of-the-Art results (83.45%) in MedNLI is achieved by pre-training BERT on PubMed followed by fine-tuning on MIMIC III v1.4. The same model achieves an accuracy of 78.5% on MEDIQA test set. The authors have identified distinct attention patterns in BERT trained on MedNLI and have explored the origin and significance of those patterns in the context of NLI.

Acknowledgments

The authors would like to thank Bhuvana Kundumani for reviewing the manuscript and for providing her technical inputs. The authors would also like to extend their gratitude to Saama Technologies Inc. for providing the perfect research and innovation environment.

References

Brian K Atchinson and Daniel M Fox. 1997. From the field: The politics of the health insurance portability and accountability act. *Health Affairs*, 16(3):146–150.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, 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 EP –. Data Descriptor.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain.

Chaitanya Shivade. 2019. Mednli for shared task at acl bionlp 2019,physionet.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig. 2019. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

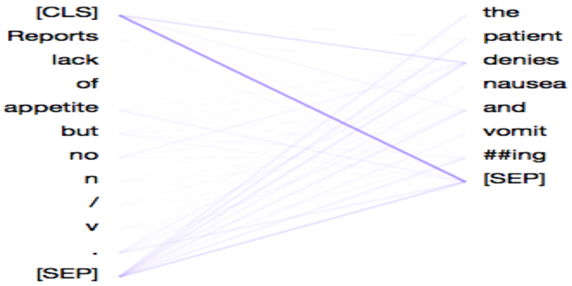

Figure 7: Attention distribution pattern for the example presented in section A without preprocessing

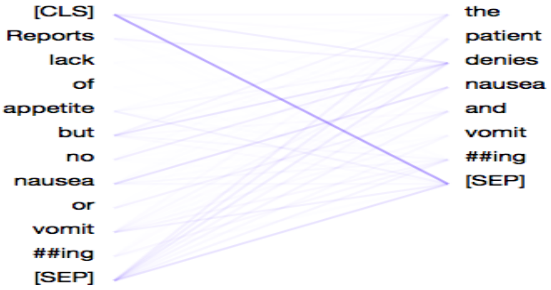

Figure 8: Attention distribution pattern for the example presented in section A after preprocessing

A Error Analysis

The authors have studied the misclassified examples in MedNLI (test set) and MedQA (task set). 70% of the misclassified examples are falsely labelled as Contradiction. The confusion matrix consisting of the count of misclassified examples for both the sets are presented in figures 9 and 10. The common pattern that exists in misclassified examples, is the model's lack of understanding of certain tokens that are crucial for relating the premise to the hypothesis. Consider the example presented below.

Premise : ”Reports lack of appetite but no n/v.” Hypothesis : ”the patient denies nausea and vomiting”

The abbreviation *n/v* in the premise expands to nausea and vomiting. The hypothesis contains the expanded form *nausea and vomiting*. It is clear from observing the attention pattern (figure 7) that the model doesn't identify *n/v* and *nausea and vomiting* as same concepts. When the abbreviation in the premise was expanded to nausea and vomiting, the model identified them as same concepts which is clearly evident from figure 8. Based on this intuition, the authors suggest a preprocessing step, that expands and normalizes abbreviated terms.

B Limitations and Future Work

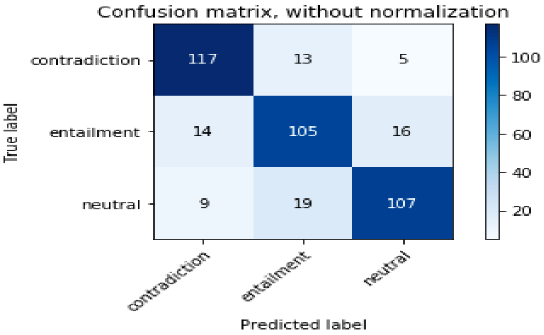

Figure 9: Confusion matrix of misclassified examples from MEDIQA test set

One of the limitations of this work is the lack of text preprocessing. The only preprocessing step followed by the authors is tokenization. In domain-specific tasks like Medical NLI, it would be beneficial to identify and normalize medical concepts which could be represented in more than one form. The other significant limitation is that the sentences are tokenized based on a 30,000 size vocabulary derived from Wikipedia corpus. Al-

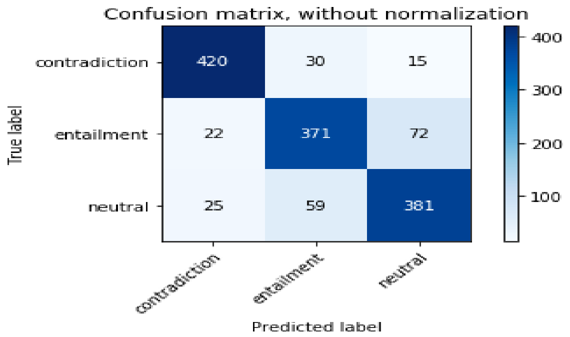

Figure 10: Confusion matrix of misclassified examples from MedNLI task set

though the fine-tuning is done on Pubmed, the commonly occurring medical terms are identified as unknown words and split into tokens.

The authors suggest a preprocessing step that identifies and normalizes medical concepts. The vocabulary could be built based on PubMed corpus which ensures that most common medical terms are part of the vocabulary. Along those lines, the authors suggest the use of entity embeddings to learn medical concepts and make use of the information contained in them.