Incorporating domain knowledge into Medical NLI using Knowledge Graphs

Anonymous EMNLP-IJCNLP submission

Abstract

Recently, models pretrained on medical text such as BioElMo have shown state-of-the-art results for the textual inference task in the medical domain. In this paper, we explore how to incorporate structured domain knowledge, available in the form of knowledge graphs, for the Medical NLI task. Specifically, we experiment with fusing knowledge graph embeddings with the state-of-the-art approaches. We also experiment with fusing the domain specific sentiment information for the task. Experiments suggest that this strategy improves the baseline BioELMo architecture for the Medical NLI task.

1 Introduction

Natural language inference (NLI) is one of the basic natural language understanding tasks which deals with detecting inferential relationship such as entailment or contradiction, between a given premise and a hypothesis. In recent years, with the availability of large annotated datasets like SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), researchers have come up with several neural network based models which could be trained with these large annotated datasets and are able to produce state-of-the-art performances (Bowman et al., 2015, 2016; Munkhdalai and Yu, 2017; Sha et al., 2016; Chen et al., 2017; Tay et al., 2017). Even though with these attempts NLI in domains like fiction, travel etc. has progressed a lot, NLI in medical domain is yet to be explored extensively. With the introduction of MedNLI (Romanov and Shivade, 2018), an expert annotated dataset for NLI in the clinical domain, researchers attempt the problem of clinical NLI.

Recently, with the emergence of advanced contextual word embedding methods like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), performance of many NLP tasks have improved, setting state-of-the-art performances. Following this stream of literature, Lee et al. (2019) introduce BioBERT, which is a BERT model pretrained on English Wikipedia, BooksCorpus and fine-tuned on PubMed (7.8B tokens in total) corpus, PMC full-text articles and Jin et al. (2019) propose BioELMo which is a domain-specific version of ELMo trained on 10M PubMed abstracts, and attempt to solve medical NLI problem with these domain specific embeddings, leading to state-of-the-art performance. These two attempts show a direction towards solving medical NLI problem where the pretrained embeddings are fine-tuned on medical corpus and are used in the state-of-the-art NLI architecture. Another line of solution tries to bring in the extra domain knowledge from sources like Unified Medical Language System (UMLS) (Bodenreider, 2004). One such attempt is made by Lu et al. (2019) by incorporating domain knowledge in terms of the definitions of medical concepts from UMLS with the state-of-the-art NLI model ESIM (Chen et al., 2017) and vanilla word embeddings of Glove (Pennington et al., 2014) and fastText (Bojanowski et al., 2017). Even though, the authors achieve significant improvement by incorporating only concept definitions from UMLS, the features of this clinical knowledge graph are yet to be fully exploited. Motivated by the emerging trend of embedding knowledge graphs to encode useful information in a high dimensional vector space, we propose the idea of applying state-of-the-art knowledge graph embedding algorithm on UMLS and use these embeddings as a representative of additional domain knowledge with the state-of-the-art medical NLI models like BioELMo, to investigate the performance improvement on this task. Additionally, we also incorporate the sentiment information for medical concepts given by MetaMap (Aronson and Lang, 2010) leading to further im-

provement of the performance. Note that, as state-of-the-art baselines we use the models proposed by Jin et al. (2019) and Lu et al. (2019) since both of these studies consider ESIM as the core NLI model which makes it more convenient for us to incorporate extra domain knowledge and to have a fair performance comparison with these state-of-the-art models. Our contributions are two-fold.

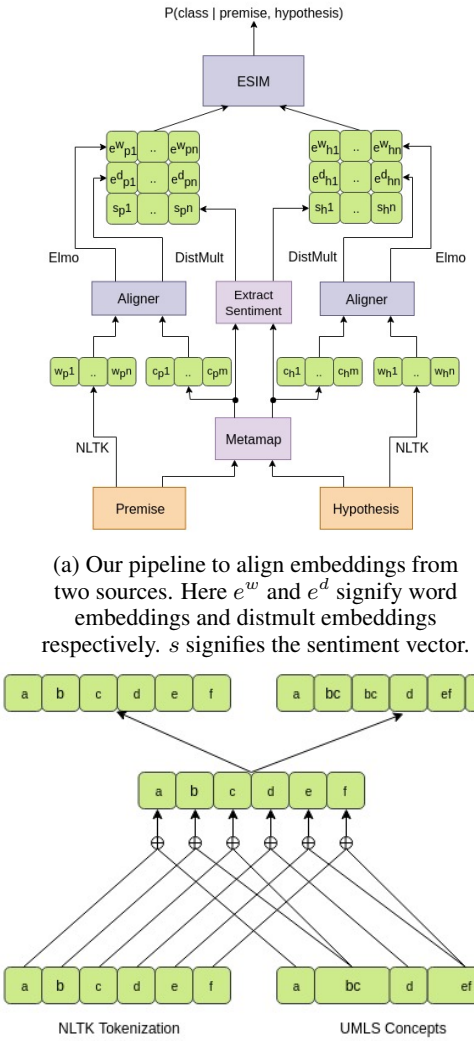
- We incorporate domain knowledge via knowledge graph embeddings applied on UMLS. We propose an intelligent path-way to combine embeddings from two domains and feed them to the state-of-the-art NLI models like ESIM which is otherwise a difficult task to deal with. This helps to improve the performance of the base architecture.
- We further show the usefulness of the associated sentiments per medical concept from UMLS in boosting the performance further, which in a way shows that if we can carefully use the domain knowledge present in sources like UMLS, it can lead to promising results as far as the medical NLI task is concerned.

2 Dataset

In this study, we use the MedNLI dataset (Romanov and Shivade, 2018), a well-accepted dataset for natural language inference in clinical domain. The dataset is sampled from doctors' notes in the clinical dataset MIMIC-III (Alistair EW Johnson and Mark., 2016) and is arguably the largest publicly available database of patient records. The entire dataset consists of 14,049 premise-hypothesis pairs divided into 11,232 train pairs, 1,395 validation pairs and 1,422 test pairs. Each such pair consists of a gold label which could be either entailment (true), contradiction (false), or neutral (undetermined). The average (maximum) sentence lengths of premises and hypotheses are 20 (202) and 5.8 (20), respectively.

3 Proposed Approach

The task is to classify the given premise (p) and hypothesis (h) sentence pair into one of the three classes: entailment, contradiction and neutral. As a core architecture, we reuse the model BioELMo (Jin et al., 2019) where authors bring in contextual information in terms of embeddings obtained via applying ELMo trained on 10M PubMed abstracts, and use these with the ESIM model (Chen et al., 2017) for the NLI



(a) Our pipeline to align embeddings from two sources. Here e^w and e^d signify word embeddings and distmult embeddings respectively. s signifies the sentiment vector.

(b) Sentence Aligner. Takes as input unaligned NLTK Tokenization and UMLS Concepts gotten from MetaMap and outputs aligned tokenizations for both

Figure 1: ELMo w/ KG pipeline

task. ESIM (Chen et al., 2017) is a state-of-the-art model for the NLI task. The architecture includes two sentence encoders each of which takes in as input the word embeddings of p and h. The inputs are run through a bi-directional LSTM encoder layer. Pairwise attention matrix is computed between p and h, which forms the attention layer followed by a second bi-directional LSTM layer. Max and average pooling are performed over the outputs of LSTM layers and the output of pooling operations is run through a softmax model. We feed this architecture an additional domain knowledge from UMLS as vector representations obtained via knowledge graph embeddings, the details of which are described below.

UMLS: Unified Medical Language System (UMLS) is a compendium which includes many health and biomedical vocabularies and standards.

It provides a mapping structure between these vocabularies and is a comprehensive thesaurus and ontology of biomedical concepts. UMLS contains 3 knowledge sources: Metathesaurus, Semantic Network, and Specialist Lexicon and Lexical Tools. We use two of these sources: the Metathesaurus and the Semantic Network. The Metathesaurus comprises of over 1 million biomedical concepts and 5 million concept names. Each concept has numerous relationships with each other. Each concept in the Metathesaurus is assigned one or more Semantic Type linked to other Semantic Types through a semantic relationship. This information is provided in the Semantic Network of UMLS. There are 127 semantic types and 54 relationships in total. Semantic types include "disease", "symptom", "laboratory test" and semantic relationships include "is-a", "part-of", "affects".

MetaMap: MetaMap is a tool for effective mapping of biomedical text to the UMLS Metathesaurus. On feeding a sentence to MetaMap, it divides the sentence into phrases based on medical concepts found in the sentence and for each medical concept it provides its ID in Metathesaurus, its position in the sentence, the list of semantic types the concept is mapped to, the preferred medical name and ID for the preferred concept (such as "chest pain" would be "angina"). We also get a boolean value denoting whether the medical concept occurs in a negative sentiment (1) or not (0). For example, in the sentence, "The patient showed no signs of pain", medical concept 'pain' would appear with a negative sentiment. Note that, for each extracted phrase, there may be more than one related medical concepts and each concept may have more than one possible mapping. For our study, we only consider the mapping with the highest MetaMap Indexing (MMI) score, a metric provided by MetaMap. As a result, every word has zero or one corresponding medical concept.

Constructing the appropriate knowledge graph: We use the Metamap tool to process the complete MedNLI dataset and extract the relevant information from UMLS into a smaller knowledge graph. First, we use Metamap to extract medical concepts from p and h, and map them to the standard terminology in UMLS. We choose to map each medical concept to its preferred medical term. E.g., "blood clots" would map to "thrombus". This helps us to map different synonymous surface forms to the same concept. This results in 7,496 unique medical concepts from UMLS matched to various words and phrases in the MedNLI dataset. Each unique concept in UMLS becomes a node in our knowledge graph. The relations in our knowledge graph come from two sources: The Metathesaurus and the Semantic Network of UMLS. Using relations extracted from these two sources, we connect the filtered medical concepts from UMLS to build a smaller Knowledge Graph (subgraph of UMLS). We get 117,467 triples from the Metathesaurus and 23,824,105 triples from the Semantic Network, which constitute the edgelists in the prepared knowledge graph.

Knolwedge Graph Embeddings: To obtain the embedding from this graph, we use state-of-the-art Distmult model (Bishan Yang and Deng, 2015). The choice is inspired by Kadlec et al. (2017), which reports that an appropriately tuned DistMult model can produce similar or better performance while compared with the competing knowledge graph embedding models.

Combining Knolwedge Graph Embeddings with BioELMo: As explained in Figure 1b, each sentence (p or h) is tokenized using NLTK as well as processed using MetaMap to get UMLS concepts. To align these, we copy the UMLS concept for a phrase to all the constituent words. Once we have aligned the tokens obtained via NLTK and MetaMap, we apply ELMo and Distmult to get the embedding vectors, $e_{ELMo,w}$ and $e_{distmult,w}$ for each word w. We concatenate these vectors as $e_w = e_{ELMo,w} \oplus e_{distmult,w}$ to obtain the word representation for w. We call the proposed model which uses these embeddings as *BioELMo w/ KG*.

Combining Sentiment Information: We further enhance the domain knowledge by incorporating sentiment information for a concept separately. For that purpose, we use the sentiment boolean provided by MetaMap and create a 1-d vector $(sent_w)$ containing 0 for positive medical concepts or non-medical concept and 1 for negative concept. We concatenate this single dimension with our concatenated resultant embeddings. Thus $e_w = e_{ELMo,w} \oplus e_{distmult,w} \oplus sent_w$. We call the proposed model which uses these embeddings as *BioELMo w/ KG + Sentiment*.

We use the vanilla ESIM model (Chen et al., 2017) and feed the model the obtained concate-

Model	Accuracy
BL_1 (Jin et al., 2019)	78.2%
BL_2 (Lu et al., 2019)	77.8%
BioELMo w/ KG	78.76%
BioELMo w/ KG+Sentiment	**79.04%**

Table 1: Performance of our models along with the state-of-the-art baseline models

nated embeddings for each word in the premise and hypothesis, to be trained for the inference task (see Figure 1a).

4 Experimental Results and Analysis

As discussed earlier, we mainly consider the models presented by Jin et al. (2019) [BL_1] and Lu et al. (2019) [BL_2] as our baselines. We report accuracy as the performance metric. Table 1 represents the performance comparison of our proposed models and the baselines, which shows that incorporation of knowledge graph embeddings helps to improve the model performance. Further, incorporating sentiment of medical concepts gives further improvements, achieving an overall 1% improvement over the baseline model.

We also see from (Jin et al., 2019) that BERT and BioBERT show an accuracy of 77.8% and 81.7%, respectively. However, they also showcase through a probing task that BioELMo is a better feature extractor than BioBERT, even though the latter has higher performance when fine tuned on MedNLI. Due to this reason, we take BioELMo as our base architecture and use our enhancements over BioELMo instead of BioBERT.

We also experimented with replacing contextual embeddings (BioELMo) with non-contexualized word embeddings (Glove, fastText). However, the accuracies for fastText (73.67%) and Glove (74.46%) were much lower than that for ELMo.

Training Details: For *DistMult*, we use word embeddings dimensions to be 100. SGD was used for optimization with an initial learning rate of 10^{-4}. The batch size was set to be 100. For *ESIM*, we take the dimension of hidden states of BiLSTMs to be 500. We set the dropout to be 0.5 and choose an initial learning rate of 10^{-3}. We choose a batch size of 32 and run for a maximum of 64 epochs. The training is stopped when the development loss does not decrease after 5 subsequent epochs.

Qualitative Analysis: We explain the efficacy of our model with the help of a few examples. Consider the sentence pair, p: "History of CVA" and h: "patient has history of stroke". In medical terms, 'CVA' means 'Cerebrovascular accident' which is another term for 'stroke'. By Using MetaMap, we are able to find that the preferred term for 'stroke' is 'Cerebrovascular accident' and hence our model classified the sample pair correctly as entailment. To take another example, consider the pair p: "Blood Glucose 626" and h: "Patient has normal A1c". The level of blood glucose indicated is higher than normal. 'A1c' is a common blood test used to diagnose type 1 and type 2 diabetes. Since the patient has higher blood glucose level, the patient having normal 'A1c' would be a contradiction and is thus classified as such.

Even though our model produces a decent performance, there are cases which our model is not able to capture. For example, for the sentence pair p: "She was speaking normally at that time" and h: "The patient has no known normal time where she was speaking normally," contradicting each other, our model predicts this to be entailment. The probable reason could be that the ESIM model fails to capture the inverse relationship in the hypothesis. In another example case, p: "He had no EKG changes and first set of enzymes were negative." and h: "the patient has negative enzymes," our model classifies this pair as entailment while the gold label is neutral. While the premise says that the first set of enzymes was negative, it gives no information about the current state. This leads us to believe that a sense of timeline is extremely important for examples like this which is not already being captured by our model. Taking care of these cases would be our immediate future work.

5 Conclusion

In this paper, we showed that knowledge graph embeddings obtained through applying state-of-the-art model like Distmult from UMLS could be a promising way towards incorporating domain knowledge leading to improved state-of-the-art performance for the medical NLI task. We further showed that sentiments of medical concepts can contribute to medical NLI task as well, opening a new direction to be explored further. With the emergence of knowledge graphs in different domains, the proposed approach can be tried out in other domains as well for future exploration.

References

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

A. R. Aronson and F.-M. Lang. 2010. An overview of metamap: Historical perspective and recent advances. *J. Amer. Med. Inform. Assoc.*, 17:229–236.

Xiaodong He Jianfeng Gao Bishan Yang, Wen-tau Yih and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. *ICLR*.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R Bowman, Raghav Gupta, Jon Gauthier, Christopher D Manning, Abhinav Rastogi, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*. Association for Computational Linguistics (ACL).

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

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*.

Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. 2017. Knowledge base completion: Baselines strike back. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 69–74, Vancouver, Canada. Association for Computational Linguistics.

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

Mingming Lu, Yu Fang, Fengqi Yan, and Maozhen Li. 2019. Incorporating domain knowledge into natural language inference on clinical texts. *IEEE Access*.

Tsendsuren Munkhdalai and Hong Yu. 2017. Neural tree indexers for text understanding. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 11. NIH Public Access.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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)*, pages 2227–2237.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Lei Sha, Baobao Chang, Zhifang Sui, and Sujian Li. 2016. Reading and thinking: Re-read lstm unit for textual entailment recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2870–2879.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. *arXiv preprint arXiv:1801.00102*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.