MSIT_SRIB at MEDIQA 2019: Knowledge Directed Multi-task Framework for Natural Language Inference in Clinical Domain

Sahil Chopra[1], Ankita Gupta[2], and Anupama Kaushik[1]

[1]Maharaja Surajmal Institute of Technology, Delhi .
[2] Samsung Research Institute, Bangalore.
{*sahilchopra, anupama*}*@msit.in*
gupta.ankita@samsung.com

Abstract

In this paper, we present Biomedical Multi-Task Deep Neural Network (Bio-MTDNN) on the NLI task of MediQA 2019 challenge (Ben Abacha et al., 2019). Bio-MTDNN utilizes "transfer learning" based paradigm where not only the source and target domains are different but also the source and target tasks are varied, although related. Further, Bio-MTDNN integrates knowledge from external sources such as clinical databases (UMLS) enhancing its performance on the clinical domain. Our proposed method outperformed the official baseline and other prior models (such as ESIM and Infersent on dev set) by a considerable margin as evident from our experimental results.

1 Introduction

The task of natural language inference (NLI) intends to determine whether a given hypothesis can be inferred from a given premise. This task also referred to as recognizing textual entailment (RTE), is one of the most prevalent tasks among NLP researchers . It has been one of the significant components for several other language applications such as Information Extraction (IE), Question Answering (QA) or Document Summarization. For example, Harabagiu and Hickl (2006) argue that RTE can enable QA systems to identify correct answers by allowing filtering and reranking them w.r.t a given question. Another approach is proposed by Ben Abacha and Demner-Fushman (2016), whereby the authors employ RTE in IE/QA domain to answer a given question (queried by a consumer) by retrieving similar questions that are already well responded by professionals.

In order to address this simple yet challenging task of NLI, several open domain datasets have been proposed, with Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) being the most popular ones. They serve as a standard to assess recent NLI systems. However, there have been only a few resources available in specialized domains such as biomedical or medicine. Language inference in the medical domain is extremely complex and remains less explored by the ML community. This scantiness of adequate resources (in terms of datasets) can be attributed to the fact that patient's data is sensitive, is accessible to authorized medical professionals only, and requires domain experts to annotate it, unlike generic domains where one can rely on crowdsourcing based techniques to acquire annotations.

To this end, Ben Abacha et al. (2019) released a new dataset made available through MIMIC-III derived data repository, named MedNLI, for NLI in the clinical domain which has been annotated by experts. Along these lines, the MediQA 2019 challenge aims to foster the development of appropriate methods, techniques and standards for inference/entailment in the medical domain, specifically on MedNLI dataset through a shared task. The task intends to recognize three inference relations between two sentences: Entailment, Neutral and Contradiction.

Previous research associated with the present task, such as work by Romanov and Shivade (2018) analyzed several state-of-the-art open domain models for NLI on the MedNLI dataset. The same has been utilized as a baseline for comparison in the above mentioned shared task. Prior to this, efforts have been made towards the automatic construction of RTE datasets (Ben Abacha and Demner-Fushman, 2016; Abacha et al., 2015), application of active learning on small RTE data (Shivade et al., 2015).

Our approach to solving the NLI task on the

Proceedings of the BioNLP 2019 workshop, pages 488–492
Florence, Italy, August 1, 2019.

MedNLI data is based on leveraging transfer learning paradigm integrated with direct incorporation of domain-specific knowledge from medical knowledge bases (KB). Unlike Romanov and Shivade (2018) which utilizes transfer learning to utilize standard NLI models (such as InferSent and ESIM trained specifically on NLI task only) in the clinical domain, we employ Mutli-task learning (MTL) framework with domain adaptation to learn representations across multiple natural language understanding (NLU) tasks. This approach not only leverages vast amounts of cross-task data but also benefits from a regularization effect that leads to better generalization and facilitates adaptation to new tasks and domains. Besides domain adaptation, we also directly infuse domain specific knowledge from database of medical terminologies so as to enable the system to perform well in the clinical domain.

The rest of the paper is organized as follows: Section 2 describe the details of our approach. Section 3 demonstrates the experimental results. We conclude in Section 4.

2 Approach

This section elaborates on the various methods we experimented with for the NLI task. In order to establish a simple baseline first, we utilize a feature-based system. The extracted features include word containment (Lyon et al., 2001) and Jaccard similarity (unigram, bigram, and trigram) based features. We also use similarity measure of distributed sentence representations obtained using universal sentence encoder (Cer et al., 2018). We consider Levenshtein, and Euclidean distance, negations and cosine function as similarity measures. In order to find the n-grams, we utilize NLTK and scispaCy tokenizer (Neumann et al., 2019). We train a 3-class logistic regression classifier with above-mentioned features to output the inference relations. Apart from this baseline, We now elaborate on the transfer learning and external knowledge integration based method in the following subsections.

2.1 Transfer Learning

Given the vast amounts of data available in the open-domain NLU tasks, we leverage them to attack the NLI task on MedNLI. Given a source domain D_S, a corresponding source task T_S, as well as a target domain D_T and a target task T_T, the objective of transfer learning is to learn the target conditional probability distribution $P(Y_T|X_T)$ in D_T with the information gained from D_S and T_S where $D_S \neq D_T$ and/or $T_S \neq T_T$. X and Y are feature and label space respectively.

We consider the scenario when $D_S \neq D_T$ (D_S being open-domain and D_T being clinical domain) and $T_S \neq T_T$, with two possibilities for target task T_T. In the first scenario, we consider a single related T_T and in the second scenario we leverage multi-task framework where we augment the T_T with multiple but related NLU tasks. For both the scenarios, we utilize the method of sequential transfer where a model is pre-trained on the large source domain data and fine-tuned on limited target domain data (clinical here). Next, we describe the neural network based models that we utilize.

2.1.1 Bi-CNN-MI

We leverage Bi-CNN-MI model (Yin and Schütze, 2015) to realize the single transfer task scenario. This DNN model is trained on a similar NLU task of paraphrase identification (PI) which is formalized as a binary classification task: for given two sentences, determine whether they both convey roughly the same meaning.

Bi-CNN-MI compares two sentences on multiple levels of granularity (word, short n-gram, long n-gram and sentence) and learns corresponding sentence representations using a convolutional neural network (CNN) based Siamese network. It also captures the sentence interactions between two sentences by computing an interaction matrix at each level of granularity. This model has been reported to outperform various earlier approaches on PI (Yin and Schütze, 2015).

We leverage this model for sequential transfer by learning the model parameters on the PI task and fine-tuning them on MedNLI dataset. Note that the classification task in MedNLI can also benefit by capturing interactions at various levels of granularity making it related to PI task but at the same time different from PI as the objective of MedNLI is not only to determine if a pair of sentences convey the same meaning but also segregate if they oppose each other or are unrelated.

2.1.2 MT-DNN

In the second scenario of transfer learning, we augment the target task T_T by various related NLU tasks and train the model to perform on all of them. This approach not only leverages exten-

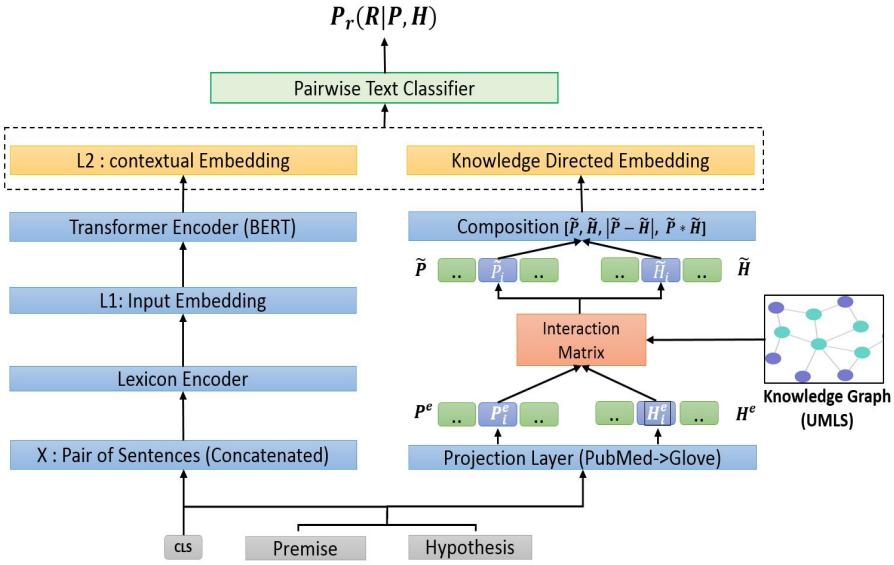

Figure 1: Architecture Diagram for Bio-MTDNN. $P_r(R|P,H)$ denotes the probability of inference relation (R) between Premise (P) and Hypothesis (H).

sive amounts of data on multiple tasks but also enables the regularization effect leading to better generalization ability. Essentially, we want to use the knowledge acquired by learning from related tasks to do well on a target task. For this approach, we utilize MT-DNN (Liu et al., 2019) which combines MTL with pre-trained language model (BERT) to improve the text representations.

The MT-DNN model combines four types of NLU tasks: single-sentence classification (sentiment classification, grammatical acceptability), pairwise text classification (NLI on several corpus and PI), text similarity scoring (STS-B), and relevance ranking (QNLI). Note, the pairwise text classification task is the NLI task that we originally intended to address in MedNLI.

The model architecture of MT-DNN involves lower layers that are shared across all tasks, while the top layers represent task-specific outputs. The input X, comprising of premise P and hypothesis H is concatenated and represented as a sequence of embedding vectors (Layer L1). The transformer encoder (BERT) then captures contextual information in the second layer (L2). This is the shared semantic representation that is trained by the multi-task objectives.

MT-DNN trained on all of the above-mentioned tasks on open-domain datasets is then fine-tuned by MedNLI dataset. In this fine-tuning step, we update the shared weights and weights associated with only the pairwise text classification task. Essentially, we first try to capture the knowledge from several related tasks in NLU followed by adapting the model to the clinical domain.

2.2 Knowledge from External Sources

Medical texts often hold relations between entities which require domain-specific knowledge for the analysis. For example, the knowledge that pneumonia is a lung disease may not be evident from the clinical text directly. In such a scenarios, incorporation of external knowledge which conveys such relationships can help. We utilize UMLS database (restricted to the SNOMED-CT terminology) represented as a graph where clinical concepts are nodes, connected by edges representing relations, such as synonymy, parent-child, etc. Next we discuss the details of the mechanism to incorporate this external knowledge, thus elaborating our Bio-MTDNN model architecture.

2.2.1 Bio-MTDNN

We propose Bio-MTDNN model which integrates domain knowledge on top of the MT-DNN model in a way similar to how interactions are captured in Bi-CNN-MI model. Specifically, we calculate the interaction matrix $I \epsilon R^{N \times M}$ between all pairs of tokens P_i and H_j in the input premise (length N) and hypothesis (length M) respectively. The value in each cell is the length of the shortest path l_{ij} between the corresponding concepts of the premise and the hypothesis in SNOMED-CT. This matrix is then utilized to generate knowledge attended representations, $\tilde{P}$ and $\tilde{H}$. Each token $\tilde{P}_i$ of the

premise is a weighted sum of the embedding H_j^e of the relevant tokens H_j of the hypothesis, weights derived from the interaction matrix. Finally, the two knowledge directed representations (averaged over the token representations) of the premise $\tilde{P}$ and hypothesis $\tilde{H}$ are composed together using elementary operations (concatenation, multiplication and subtraction) and fed to a single feed forward layer. This composed representation is then concatenated with the L2 layer of MT-DNN before passing it to the task-specific layers.

In the above process, the creation of knowledge directed representations relies upon the input token embeddings of premise (P_j^e) and hypothesis (H_j^e). One of the simplest options for token embeddings is to use GloVe embeddings (Pennington et al., 2014). However, these embeddings are not specific to the clinical domain and may result in many tokens being mapped to the embedding of the unknown (UNK) token. To alleviate this issue, we learned a non-linear transformation (Sharma et al., 2018) that maps words from PubMed (Pyysalo et al., 2013) to GloVe subspace. We train the DNN using the common words in both the embeddings. We obtain the transformed embeddings for all the words in the PubMed that are not present in the GloVe by using inference step of the learned DNN.

Note that, here we cannot utilize the embeddings learned in the first layer (L1) of MT-DNN as they incorporate segment embeddings of the premise and hypothesis concatenated together. Thus, the L1 layer of MT-DNN learns the interactions between premise and hypothesis in an end-to-end manner. However, what we are trying is to learn these interactions which are directed by the knowledge obtained from UMLS enabling Bio-MTDNN to incorporate external information.

3 Experiments and Results

3.1 Setup and Implementation Details

For the feature-based system we used Logistic Regression classifier from the scikit-learn library (Pedregosa et al., 2011). We use publicly available implementations for Bi-CNN-Mi[1] and MT-DNN[2]. For external knowledge integration, the required medical concepts in SNOMED-CT were identified in the premise and hypothesis sentences using MetaMap by Aronson and Lang (2010). We used glove and PubMed word embeddings and used DNN (Sharma et al., 2018) for non-linear projection. In all experiments we report the average result (on the dev set) of 5 different runs, with the same hyperparameters and different random seeds. For the best performing systems, we also report the results on the test set.

[1]https://github.com/chantera/bicnn-mi
[2]https://github.com/namisan/mt-dnn

Model	Dataset	
	Dev	Test
MT-DNN + External Knowledge	81.2	81.3
MT-DNN	80.1	80.5
Infersent	73.5	-
ESIM	73.1	-
Official Baseline		71.4
Features Baseline	51.9	49.4
Bi-CNN-MI	54.1	53.6

Table 1: Experimental Results

3.2 Results and Discussions

Table 1 mentions the experimental results for all the systems. Bio-MTDNN performs best among all the systems with 81.2% accuracy on the dev set. Integration of external knowledge in Bio-MTDNN helped the system to outperform the MT-DNN performance (with 80.1% accuracy). The multi-task learning framework boosted the performance of both the systems. We submitted results from Bio-MTDNN for the challenge which obtained 81.3% accuracy on the test set.

In order to compare against other transfer learning based approaches (Romanov and Shivade, 2018), we also mention the results of Infersent and ESIM (note that for both these models, $D_S \neq D_T$ and $T_S = T_T$, unlike the scenarios we considered). It can be observed that Bio-MTDNN outperforms both ESIM and Infersent with significant margins. This can be attributed to the external knowledge incorporation and ability of MTL framework which empowers the model to learn better shared representations. However, contrary to the expectations, Bi-CNN-MI model performs very poorly on the dev dataset with only 54.1% accuracy, only slightly better than feature based baseline which achieves 51.9 % accuracy. This may be attributed to the possibility that the knowledge gained by Bi-CNN-MI when trained on PI task (although a related task to NLI) is not sufficient for the model to be able to segregate contradicting premise and hypothesis.

4 Conclusion

In this paper, we introduce Bio-MTDNN , which is a knowledge directed, multi-task learning based language inference model for biomedical text mining. While MT-DNN was built for general purpose language understanding, Bio-MTDNN effectively leverages domain specific knowledge from UMLS as demonstrated by our experimental study. We presented our results on the MedNLI dataset under MediQA challenge. Incorporation of knowledge from external sources such as UMLS gives performance advantage to Bio-MTDNN. Our proposed system outperformed the official baseline and other prior models (ESIM and Infersent on dev set) by a great margin.

References

Asma Ben Abacha, Duy Dinh, and Yassine Mrabet. 2015. Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 238–242. Springer.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association.

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*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905–912. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504.

Caroline Lyon, James Malcolm, and Bob Dickerson. 2001. Detecting short passages of similar text in large document collections. In *Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP)*, pages 118–125.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Vasu Sharma, Nitish Kulkarni, Srividya Pranavi, Gabriel Bayomi, Eric Nyberg, and Teruko Mitamura. 2018. Bioama: Towards an end to end biomedical question answering system. In *Proceedings of the BioNLP 2018 workshop*, pages 109–117.

Chaitanya Shivade, Courtney Hebert, Marcelo Lopetegui, Marie-Catherine De Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Textual inference for eligibility criteria resolution in clinical trials. *Journal of biomedical informatics*, 58:S211–S218.

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.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.