Surf at MEDIQA 2019: Improving Performance of Natural Language Inference in the Clinical Domain by Adopting Pre-trained Language Model

Jiin Nam
AI Core Team
Samsung Research
Seoul, Korea
jiin.nam@samsung.com

Seunghyun Yoon
Dept. ECE
Seoul National University
Seoul, Korea
mysmilesh@snu.ac.kr

Kyomin Jung
Dept. ECE
Seoul National University
Seoul, Korea
kjung@snu.ac.kr

Abstract

While deep learning techniques have shown promising results in many natural language processing (NLP) tasks, it has not been widely applied to the clinical domain. The lack of large datasets and the pervasive use of domain-specific language (i.e. abbreviations and acronyms) in the clinical domain causes slower progress in NLP tasks than that of the general NLP tasks. To fill this gap, we employ word/subword-level based models that adopt large-scale data-driven methods such as pre-trained language models and transfer learning in analyzing text for the clinical domain. Empirical results demonstrate the superiority of the proposed methods by achieving 90.6% accuracy in medical domain natural language inference task. Furthermore, we inspect the independent strengths of the proposed approaches in quantitative and qualitative manners. This analysis will help researchers to select necessary components in building models for the medical domain.

1 Introduction

Natural language processing (NLP) has broadened its applications rapidly in recent years such as question answering, neural machine translation, natural language inference, and other language-related tasks. Unlike other tasks in NLP area, the lack of large labeled datasets and restricted access in the clinical domain have discouraged active participation of NLP researchers for this domain (Romanov and Shivade, 2018). Furthermore, the pervasive use of abbreviations and acronyms in the clinical domain causes the difficulty of text normalization and makes the related tasks more difficult (Pakhomov, 2002).

In building NLP models, a word embedding layer that transforms a sequence of tokens in text into a vector representation is considered as one of the fundamental components. In recent studies, it has been shown that the pre-trained language models by using a huge diversity of corpus (i.e. BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018)) generate deep contextualized word representations. These methods have shown to be very effective for improving the performance of a wide range of NLP tasks by enabling better text understanding and have become a crucial part of the tasks since they have published.

To stimulate the research in the clinical domain, researchers have further investigated to transform the pre-trained language models from general purpose version into the medical domain-specific version. Lee et al. (2019) propose BioBERT that utilizes large-scale bio-medical corpora, PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC), to obtain a medical domain specific language representation through fine-tuning the BERT. Similarity, a PubMed-ELMo[1], trained with medical domain corpus, is released as one of the contributed ELMo models for medical domain researchers. However, these models are not yet fully explored in medical domain tasks.

Besides these general efforts in building better word representations, Romanov and Shivade (2018) introduce a large and publicly available natural language inference (NLI) dataset, called MedNLI, for the medical domain (see table 1). Considering the expensive annotation cost of medical text due to the sparsity of the clinical-domain experts, the medical NLI task plays an import role in boosting existing datasets for medical question answering systems by retrieving similar questions that are already answered by human experts. Along with this effort, ACL-BioNLP 2019 committee announced a shared task, NLI for the medical domain, motivated by a need to develop

---

[1]https://allennlp.org/elmo

#	Premise	Hypothesis	Label
1	She was treated with Magnesium Sulfate, Labetalol, Hydralazine and bedrest as well as betamethasone.	The patient is pregnant.	entailment
2	Denied headache, sinus tenderness, rhinorrhea or congestion.	Patient has history of dysphagia	contradiction
3	Type II Diabetes Mellitus 3.	The patient does not require insulin.	neutral
4	Ruled in for NSTEMI with troponin 0.11.	The patient has myocardial ischemia.	entailment
5	Her CXR was clear and it did not appear she had an infection.	Chest x-ray showed infiltrates	contradiction
6	CHF, EF 55% 6.	complains of shortness of breath	neutral

Table 1: Examples from the development set of MedNLI.

relevant methods, techniques and gold standards for inference and entailment (Ben Abacha et al., 2019). The newly released dataset is larger in size than that of any other previous medical domain NLI dataset, however, it is still not enough to train complicated neural network based models.

To fill this gap, we propose a combination approach of NLP models and machine learning methods to tackle the medical domain NLI task. Our contributions are summarized as follows:

- We adopt the pre-trained language models (BioBERT, PubMed-ELMo) to overcome the shortage of training data which is a common problem in the clinical domain.

- We apply the transfer learning method with two general domain NLI datasets and show that a source task in a domain can benefit learning a target task in a different domain.

- We show the independent strengths of the proposed approaches in quantitative and qualitative manners. This analysis will help researchers to select necessary components in building models for the clinical domain.

2 Related Work

Researchers have investigated NLI tasks. Most of the works employed a recurrent neural network to encode each pair of sentences and to compute the similarity between them (Conneau et al., 2017; Subramanian et al., 2018). Recently, Liu et al. (2019) proposed multi-task learning for natural language tasks and achieved the best results on NLI tasks. In the medical domain, Romanov and Shivade (2018) adopted the ESIM (Chen et al., 2017) model to the MedNLI task. The ESIM

model employs two bidirectional LSTM to encode each sentence independently and to calculate a matching score between the sentences by using alignment and pooling methods. They also applied transfer learning with SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) datasets to improve model performance in the MedNLI task.

Recently, pre-trained language models were proposed (Peters et al., 2018; Devlin et al., 2018). The multi-task benchmark for natural language understanding (Wang et al., 2018) has shown that these pre-trained language models brought additional performance gain by providing deep contextualized word representations. Upon this success, researchers further extended previous pre-trained language models to medical domain-specific versions such as BioBERT (Lee et al., 2019) and PubMed-ELMo (Peters, 2018).

However, none of these researches directly applied the pre-trained language models of the medical domain to the MedNLI task.

3 Dataset and Problem

MedNLI (Romanov and Shivade, 2018), a large publicly available and expert annotated dataset, has been recently published for the MEDIQA 2019 shared task. This dataset comprises of tuples $<P, H, Y>$ where: P and H are a clinical sentence pair, (premise and hypothesis, respectively); Y indicates whether a given hypothesis can be inferred from a given premise. In particular, Y is categorized as one of three classes: "*entailment*", "*contradiction*", and "*neutral*". Table 1 shows examples of the MedNLI dataset. A total of 14,049 pairs, (11,232, 1,395, 1,422 for training, development, and test, respectively), are created based on the past medical history section of MIMIC-
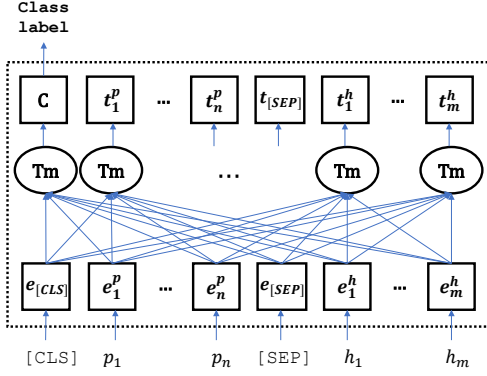
Figure 1: Overview of the BERT model.

III (Johnson et al., 2016).

In this research, we are interested in building a model that classifies the given sentence pair into the corresponding category. First, we consider a **point-wise** approach that classifies each pair of data independently into one of the three classes. Next, we re-organize the dataset into the set of a list that contains one of each class sentence pair. Then we apply **list-wise** classification that classifies three sentence pair into each "*entailment*", "*contradiction*", and "*neutral*" class exclusively.

4 Methods

As the size of the MedNLI dataset is limited to train the whole weight parameters in complicated neural network based models, we first choose a BERT (Devlin et al., 2018) based model that provides pre-trained model parameters from a large corpus. To further explore the performance of modern neural network based models, we extend the compare aggregate model (Wang and Jiang, 2016) with another type of pre-trained word-level embedding, ELMo (Peters et al., 2018). Additionally, we apply transfer learning from similar NLI tasks (Bowman et al., 2015; Williams et al., 2018), and we try to expand medical abbreviations to deal with the general problem in the medical domain.

4.1 BioBERT

As a baseline model, we choose BioBERT (Lee et al., 2019) since MedNLI is a bio-domain specific NLI task. It shows strength in understanding medical domain text as it is fine-tuned with bio-datasets such as PubMed and PMC. The BioBERT adopts the same architecture as BERT, as shown in figure 1, that takes WordPiece embeddings from textual input and generates a language representation using a transformer model (Vaswani et al.,

2017).

WordPiece embedding: BioBERT utilizes the WordPiece dictionary of BERT generated from general domain corpus. Each premise **P** and hypothesis **H** turn into sub-word embeddings, $\mathbf{E}^P \in \mathbb{R}^{n \times d_e}$ and $\mathbf{E}^H \in \mathbb{R}^{m \times d_e}$, using the dictionary where d_e is a dimension of sub-word embedding vectors and n and m are the length of the sequences of **P** and **H**, respectively.

$$\mathbf{E}^P = \text{WordPiece\_embedding}(\mathbf{P}),$$
$$\mathbf{E}^H = \text{WordPiece\_embedding}(\mathbf{H}). \quad (1)$$

BioBERT adds the special classification embedding "[CLS]" as the first token of every sentence and separates $\mathbf{E}^P$ and $\mathbf{E}^H$ with a special token "[SEP]". The final input representation fed to transformer blocks is the sum of the token embeddings ($\mathbf{E}^T$), position embeddings ($\mathbf{E}^{Po}$), and segmentation embeddings ($\mathbf{E}^S$) as follow.

$$\mathbf{E} = \mathbf{E}^T + \mathbf{E}^{Po} + \mathbf{E}^S,$$
$$\mathbf{E}^T = [\mathbf{E}_{[CLS]}, \ \mathbf{E}^P, \ \mathbf{E}_{[SEP]}, \ \mathbf{E}^H]. \quad (2)$$

Transformer encoder: The transformer encoder consists of multiple transformer blocks. Each block uses Multi-Head Attention (MHA) generating h different attentions. All the attention heads calculated with different weights are concatenated. A linear layer with a weight matrix $\mathbf{W}^H \in \mathbb{R}^{(h \times d_v) \times d_e}$ computes the MHA ($\mathbb{R}^{\text{input\_length} \times d_e}$) with the concatenated attention heads as follows:

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\text{concat}\{hd_1, .., hd_n\})\mathbf{W}^H,$$
$$hd_i = \text{attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i),$$
$$\text{attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax}(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}})\mathbf{V}_i, \quad (3)$$

where $Q = [Q_1, ..., Q_h], Q_i \in \mathbb{R}^{n \times \frac{d_e}{h}}$, $K = [K_1, ..., K_h], K_i \in \mathbb{R}^{n \times \frac{d_e}{h}}$, $V = [V_1, ..., V_h], V_i \in \mathbb{R}^{n \times \frac{d_e}{h}}$.

4.2 Compare Aggregate (CompAggr)

As we focus on the task that classifies the relationship between two sentences **P** and **H** (premise and hypothesis) into one of three classes (entailment, contradiction, or neutral), we adopt the compare aggregate (CompAggr) model that is widely used for a text sequence matching task (Wang
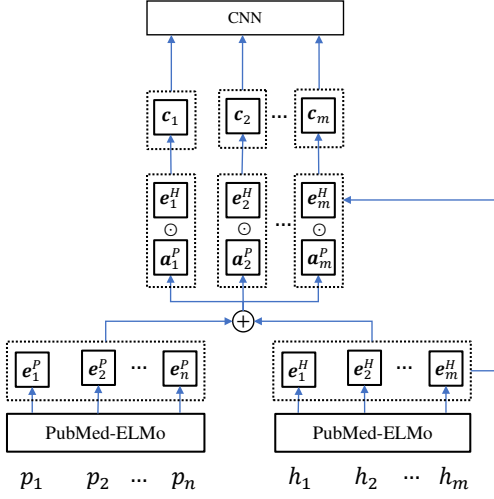
Figure 2: Overview of the CompAggr model.

and Jiang, 2016). In addition to the CompAggr model, we adopt PubMed-ELMo, that is trained with medical domain corpus and released as one of contributed ELMo models (Peters et al., 2018; Peters, 2018), to alleviate the lack of training corpus for the shared task. The final model consists of four parts which are shown in figure 2.

Word representation: Premise $\mathbf{P} \in \mathbb{R}^{d \times n}$ and hypothesis $\mathbf{H} \in \mathbb{R}^{d \times m}$, (where d is a dimensionality of word embedding and n, m are length of the sequences in $\mathbf{P}$ and $\mathbf{H}$, receptively), are processed to capture contextual information within the sentence by using pretrained PubMed ELMo (Peters et al., 2018) as follows:

$$\begin{aligned} \mathbf{E}^P &= \text{PubMed-ELMo}(\mathbf{P}), \\ \mathbf{E}^H &= \text{PubMed-ELMo}(\mathbf{H}). \end{aligned} \quad (4)$$

Attention: The soft aliment of the $\mathbf{E}^P$ and $\mathbf{E}^H$ are computed by applying an attention mechanism over the column vector in $\mathbf{E}^P$ for each column vector in $\mathbf{E}^H$. Using an attention weight α_i for each column vector in $\mathbf{E}^P$, we obtain a corresponding vector $\mathbf{A}^P \in \mathbb{R}^{d \times m}$ from weighted sum of the column vectors of $\mathbf{E}^P$.

$$\mathbf{A}^P = \mathbf{E}^P \cdot \text{softmax}((\mathbf{W}\mathbf{E}^P)^{\mathsf{T}}\mathbf{E}^H), \quad (5)$$

where $\mathbf{W}$ is a learned model parameter matrix.

Comparison: We use an element-wise multiplication as a comparison function to combine each pair of $\mathbf{A}^P$ and $\mathbf{E}^H$ into a vector $\mathbf{C} \in \mathbb{R}^{d \times m}$.

Aggregation: Finally Kim (2014)'s CNN with n-types of filters is applied to aggregate all the information followed by another fully connected layer

to classify the $\mathbf{P}$ and $\mathbf{H}$ pair as follow:

$$\begin{aligned} \mathbf{R} &= \text{CNN}(\mathbf{C}), \quad (R \in \mathbb{R}^{nd}) \\ \hat{y}_c &= \text{softmax}((\mathbf{R})^{\mathsf{T}}\mathbf{W} + \mathbf{b}), \end{aligned} \quad (6)$$

where $\hat{y}_c$ is the predicted probability distribution for the target classes and the $\mathbf{W} \in \mathbb{R}^{nd \times 3}$ and bias $\mathbf{b}$ are learned model parameters.

Our loss function is cross-entropy between predicted labels and true-labels as follow:

$$\mathcal{L} = -\log \prod_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}), \quad (7)$$

where $y_{i,c}$ is the true label vector, and $\hat{y}_{i,c}$ is the predicted probability from the softmax layer. C is the total number of classes (entailment, contradiction, and neutral for this task), and N is the total number of samples used in training.

4.3 Transfer learning

Pan and Yang (2010) provide definitions of transfer learning as follows:

Definition 1 (*Transfer Learning*) Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, *transfer learning* aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_s \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$.

While MedNLI has a relatively large amount of training data in the clinical domain, NLI tasks in general domain such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) have way larger training data than MedNLI has. Since a source and a target task in different domains can improve a model performance if they are related to each other we decide to use the two general domain NLI tasks to train BERT and BioBERT to transfer their knowledge for MedNLI. Our case is $\mathcal{D}_S \neq \mathcal{D}_T$ where the feature spaces between the domains are different or the marginal probability distributions between domain datasets are different ($P(X_S) \neq P(X_T)$).

4.4 Abbreviation expansion

Not unlike other medical text, abbreviations and acronyms are easily found throughout the text in MedNLI as table 1 shows from # 4 to 6. In order to understand the effect of expanded forms for clinical abbreviations, we replace the abbreviations with corresponding expanded forms. As Liu et al. (2015) mentions that no universal rules

Dataset	Accuracy	
	dev	test
+PMC	80.50	78.97
+PubMedd	81.14	78.83
+PubMed+PMC	**82.15**	**79.04**

Table 2: The BioBERT performance on the MedNLI task. Each model is trained on three different combinations of PMC and PubMed datasets (top score marked as bold).

Model	Accuracy	
	dev	test
BioBERT	82.15	79.04
CompAggr	80.40	75.80
BioBERT (transferred)	83.51	**82.63**
BioBERT (expanded)	**83.87**	79.95

Table 3: The model performance of four different methods (top score marked as bold). BioBERT (transferred) and BioBERT (expanded) refer to the best results of transfer learning experiments and the result of MedNLI with abbreviation expansion on BioBERT respectively.

or dictionary for clinical abbreviations is available we gather and exploit the public medical abbreviations from Taber's Online[2].

5 Experiments

We explore three kinds of BioBERT that are fine-tuned from the original BERT with PMC, PubMed, and PMC+PubMed datasets. As shown in table 2, BioBERT trained on PubMed+PMC performs the best. Thus we select it as a base BioBERT model for the rest of the experiments. Depends on a need for comparison or better understanding, we also include original BERT in the experiments and report the results. The overall results of MedNLI are shown in table 3.

5.1 Experimental Setup

All experiments based on BioBERT and BERT have a fixed learning rate 2e-5. We add early stopping to stop the models from learning if evaluation loss has not decreased for 4 steps where 1 step is defined 20% of the whole training data. Other than the learning rate and early stopping, all settings are the same as they are in BioBERT and BERT.

For the CompAggr model, we use a context projection weight matrix with 100 dimensions. In the aggregation part, we use 1-D CNN with a total of 500 filters, which involved five types of filters $K \in \mathbb{R}^{\{1,2,3,4,5\} \times 100}$, 100 per type. The weight matrices for the filters were initialized using the Xavier method (Glorot and Bengio, 2010). We use the Adam optimizer (Kingma and Ba, 2014) including gradient clipping by norm at a threshold of 5. For the purpose of regularization, we applied dropout (Srivastava et al., 2014) with a ratio of 0.7.

5.2 Performance evaluation

Transfer learning: We conduct transfer learning on four different combinations of MedNLI, SNLI, and MNLI as it shown in the table 4 (line 4 to 7) and also add the results of general domain tasks (MNLI, SNLI) for comparison. As expected, BERT performs better on tasks in the general domain while BioBERT performs better on MedNLI which is in the clinical domain.

In overall, positive transfer occurs on MedNLI. There are three things we can observe from the results. First of all, even though BioBERT is fine-tuned on general domain tasks before MedNLI, transfer learning shows better results than that fine-tuned on MedNLI directly. It implies that the same tasks in different domains have overlapping knowledge and transfer learning between the tasks effects positively on each other as the definition of transfer learning mentions in section 4. Second, the domain specific language representations from BioBERT are maintained while fine-tuning on general domain tasks by showing that the transfer learning results of MedNLI on BioBERT have better performance than the results on BERT (line 4 to 7). Lastly, the accuracy of MNLI and SNLI on BioBERT is lower than the accuracy on BERT. The lower accuracy indicates that BioBERT captures different features such as medical terms and generate different representations than what BERT does which are helpful for the clinical domain task, MedNLI, but not for the other two tasks.

The best combination is SNLI $\rightarrow$ MNLI $\rightarrow$ MedNLI on BioBERT. We refer to the best result of transfer learning as BioBERT (transferred).

Results analysis for different models: There are fundamental differences between the two models we apply. BioBERT tokenizes an input sentence

---

[2]https://www.tabers.com/tabersonline/view/Tabers-Dictionary/767492/all/Medical˙Abbreviations

Dataset	BERT		BioBERT	
	dev	test	dev	test
MedNLI	79.56	77.49	82.15	79.04
MNLI (M)	83.52	-	81.23	-
SNLI (S)	90.39	-	89.10	-
M → MedNLI	80.14	**78.62**	82.72	80.80
S → MedNLI	80.28	78.19	83.29	81.29
M → S → MedNLI	80.43	78.12	83.29	80.30
S → M → MedNLI	**81.72**	77.98	**83.51**	**82.63**
MedNLI (expanded)	79.13	77.07	**83.87**	79.95
S → M → MedNLI (expanded)	**82.15**	**79.95**	83.08	**81.85**

Table 4: All experiment results of transfer learning and abbreviation expansion (top-2 scores marked as bold). MedNLI (expanded) denotes MedNLI with abbreviation expansion.
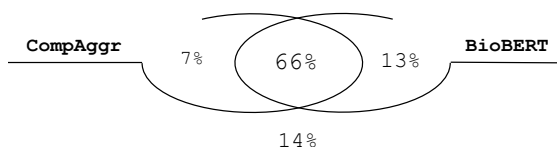

Figure 3: Venn diagram for the test results of Compare Aggregate model and BioBERT.

to sub-word level and uses the transformer model while CompAggr uses word-level embeddings and Compare&Aggregate model. In light of the dissimilar nature, we expect each model captures different features and generates different language representations.

Figure 3 shows the percentage for each area takes of the test set. CompAggr correctly classifies 97 examples (7% of the test set) which BioBERT classifies them incorrectly while BioBERT classifies 188 examples correctly (13% of the test set) which CompAggr does not. It demonstrates that both models have different strength on the MedNLI task.

We manually examine all promise and hypothesis pairs of each portion of 7% and 13% of the test set with high confidence and "element" label. For CompAggr, we pick pairs with the probability higher than 0.80 which are 6 pairs. For BioBERT, we select pairs with top 10 probabilities. Interestingly, each pair from CompAggr does not have overlapping words between premise and hypothesis. It appears that CompAggr's strength is in it's ability to capture the relationship between two sentences even though there is no word overlap while BioBERT labels them "*neutral*" except one

pair as you can see in table 5. In contrast, the majority of the pairs, 7 out of 10, from BioBERT have overlapping words between them. Biobert shows strong confidence when premise and hypothesis have overlapping words as below.

- (*Premise*) En route to the Emergency Department, she developed worsening substernal chest pain without any radiation.

- (*Hypothesis*) patient has chest pain

Lastly, we compute the average conditional probability of the correct results to check the confidence of each model. The results are 0.87 and 0.82 for BioBERT and CompAggr showing that BioBERT predicts labels with higher confidence.

Abbreviation expansion: We refer to the dataset of MedNLI with abbreviation expansion as MedNLI (expanded). The inconsistency of the experiment results on MedNLI (expanded) makes it difficult to observe their effects. MedNLI (expanded) shows better performance than MedNLI on BioBERT while MedNLI works better on BERT (see table 4). Furthermore, the performance of MedNLI (expanded) with transfer learning is higher on BERT and lower on BioBERT than the performance of MedNLI with transfer learning.

We examine the test results to figure out the inconsistency and observe an interesting phenomenon that the abbreviation expansion changes the conditional probability distribution $P(Y|X)$, where X and Y represent input texts and their expected labels, respectively. The same input texts with no expansion are classified into different

Premise	Hypothesis	CompAggr	BioBERT
He denies any fever, diarrhea, chest pain, cough, URI symptoms, or dysuria.	He denies any fever, diarrhea, chest pain, cough, URI symptoms, or dysuria.	entailment	neutral
This quickly became ventricular fibrillation and he was successfully shocked X 1 360J with return of rhythm and circulation.	Patient has NSR post-cardioversion	entailment	contradiction
PAST MEDICAL HISTORY: Coronary artery disease status post MI [**09**] years ago, status post angioplasty.	History of heart attack	entailment	neutral
A MRA prior to discharge showed increased ... of single and rector spinal muscles at T3-4 adjacent to facets and anterior within the right psoas.	the patient has degenerative changes of the spine	entailment	neutral
The patient now presents with metastatic recurrence of squamous cell carcinoma of the right mandible with extensive lymph node involvement.	The patient has oropharyngeal carcinoma.	entailment	neutral
The transbronchial biopsy was nondiagnostic.	Patient has a mediastinal mass	entailment	neutral

Table 5: Examples with the highest probabilities showing the strength of CompAggr.

Rank	Team	Accuracy
1	WTMED	98.0
2	PANLP	96.6
3	Double Transfer	93.8
4	Sieg	91.1
5	**Surf (ours)**	**90.6**
6	ARS_NITK	87.7
7	Pentagon	85.7
8	Dr.Quad	85.5
9	UU_TAILS	85.2
10	KU_ai	84.7

Table 6: Performance comparison among the top-10 participants (official) of the NLI shared task. Teams [1-4, 6-10] are from (Wu et al., 2019; Zhu et al., 2019; Xu et al., 2019; Bhaskar et al., 2019; Agrawal et al., 2019; Pugaliya et al., 2019; Bannihatti Kumar et al., 2019; Tawfik and Spruit, 2019; Cengiz et al., 2019), respectively.

classes. For instance, a pair of *Premise* and *Hypothesis* like below is not changed after abbreviation expansion since it does not contain any abbreviations or acronyms.

- (*Premise*) He denied headache or nausea or vomiting.

- (*Hypothesis*) He is afebrile.

However, the results are different. It is originally classified into "*neutral*" which is the right label for the pair but it is classified into "*entailment*" when we use MedNLI (expanded).

5.3 MEDIQA-NLI shared task

We are participating in a shared task MEDIQA-NLI of the bioNLP workshop at ACL 2019. In order to solve the task, we try four different **point-wise** approaches, CompAggr, BioBERT, transfer learning, and abbreviation expansion. We run each model several times to obtain the best result out of each. Our best result, which is ranked 5th on the leaderboard of the task, is obtained by applying **list-wise** approach (in section 3) with the best result (BioBERT (transferred)). Table 6 shows the model performance of each participant in the leaderboard.

6 Conclusion

In this paper, we study natural language inference in the clinical domain where training corpora is insufficient due to its domain nature. To tackle the problem, we propose approaches that adopts pre-trained language models, transfer learning method and data-augmentation to boost the train instances. To this end, we observe that the BioBERT pre-trained on bio-medical corpus shows better performance than that of the BERT on the general domain corpus. The CompAggr with bio-ELMO and the BioBERT behave differently in classifying the MedNLI dataset due to the difference in their own architecture. Transfer learning with NLI tasks in general domain, (MNLI, SNLI), does not hurt the ability of the BioBERT capturing language representations of the clinical domain. In addition, we

observe that it transfers positive knowledge from general NLI tasks to the MedNLI task. In contrast, a abbreviation expansion method needs particular care when adopting since it may hurt the model to predict the conditional probability distribution of the task.

Acknowledgments

References

Anumeha Agrawal, Rosa Anil George, Selvan Sunitha Ravi, Sowmya Kamath, and Anand Kumar. 2019. Ars_nitk at mediqa 2019:analysing various methods for natural language inference, recognising question entailment and medical question answering system. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Vinayshekhar Bannihatti Kumar, Ashwin Srinivasan, Aditi Chaudhary, James Route, Teruko Mitamura, and Eric Nyberg. 2019. Dr.quad at mediqa 2019: Towards textual inference and question entailment using contextualized representations. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

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.

Sai Abishek Bhaskar, Rashi Rungta, James Route, Eric Nyberg, and Teruko Mitamura. 2019. Sieg at mediqa 2019: Multi-task neural ensemble for biomedical inference and entailment. 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. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Cemil Cengiz, Ulaş Sert, and Deniz Yuret. 2019. Ku_ai at mediqa 2019: Domain-specific pre-training and transfer learning for medical nli. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

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*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the AISTATS*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*.

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

Yue Liu, Tao Ge, Kusum Mathews, Heng Ji, and Deborah McGuinness. 2015. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In *Proceedings of BioNLP 15*, pages 92–97.

Serguei Pakhomov. 2002. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 160–167. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Matthew Peters. 2018. PubMed ELMo. `https://allennlp.org/elmo`. [Online; accessed 09-May-2019].

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.

Hemant Pugaliya, Karan Saxena, Shefali Garg, Sheetal Shalini, Prashant Gupta, Eric Nyberg, and Teruko Mitamura. 2019. Pentagon at mediqa 2019: Multi-task learning for filtering and re-ranking answers using language inference and question entailment. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

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.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *ICLR*.

Noha Tawfik and Marco Spruit. 2019. Uu_tails at mediqa 2019: Learning textual entailment in the medical domain. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

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.

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*.

Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747*.

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.

Zhaofeng Wu, Yan Song, Sicong Huang, Yuanhe Tian, and Fei Xia. 2019. A hybrid approach to biomedical natural language inference. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Yichong Xu, Xiaodong Liu, Chunyuan Li, Hoifung Poon, and Jianfeng Gao. 2019. Doubletransfer at mediqa 2019: Multi-source transfer learning for natural language understanding in the medical domain. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

Wei Zhu, Xiaofeng Zhou, Keqiang Wang, Xun Luo, Xiepeng Li, Yuan Ni, and Guotong Xie. 2019. Panlp at mediqa 2019: Pre-trained language models, transfer learning and knowledge distillation. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.