NCUEE at MEDIQA 2019: Medical Text Inference Using Ensemble BERT-BiLSTM-Attention Model

Lung-Hao Lee*, Yi Lu, Po-Han Chen, Po-Lei Lee and Kuo-Kai Shyu
Department of Electrical Engineering, National Central University, Taiwan
Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan
*lhlee@ee.ncu.edu.tw

Abstract

This study describes the model design of the NCUEE system for the MEDIQA challenge at the ACL-BioNLP 2019 workshop. We use the BERT (Bidirectional Encoder Representations from Transformers) as the word embedding method to integrate the BiLSTM (Bidirectional Long Short-Term Memory) network with an attention mechanism for medical text inferences. A total of 42 teams participated in natural language inference task at MEDIQA 2019. Our best accuracy score of 0.84 ranked the top-third among all submissions in the leaderboard.

1 Introduction

Natural Language Inference (NLI) is the task of determining whether a given hypothesis is true (entailment), false (contradiction) or undetermined (neutral) by inferring a given premise. The Stanford Natural Language Inference (SNLI) corpus is a well-known dataset and serves as a benchmark for NLI system evaluations (Bowman et al., 2015). However, it is restricted to a single text genre. Therefore, the MedNLI dataset, which is annotated by doctors and grounded in patients' medical histories, was built to perform NLI tasks in the clinical domain (Romanov and Shivade, 2018). In addition to feature-based methods and bag-of-words (BOW) models, other experiments have tested several modern neural networks-based models for the specialized and knowledge intensive field of medicine, including InferSent (Conneau et al., 2017) and ESIM (Chen et al., 2017)

The MEDIQA challenge focuses on attracting research efforts in Natural Language Inference (NLI), Recognizing Question Entailment (RQE) and their applications in medical Question Answering (QA). The MEDIQA challenge includes three tasks: 1) NLI: identifying three inference relations between two medical sentences, that is, entailment, neutral and contradiction. 2) RQE: identifying entailment between two questions in the context of QA. 3) QA: filtering and improving the input ranks of retrieved answers, generated by the medical QA system CHiQA. The reuse of NLI and/or RQE systems for this task is highly recommended.

Under the policies of the MEDIQA challenge, we only participated in the first NLI task. Recently, a new method of pre-training language representations named BERT (Bidirectional Encoder Representations from Transformers) has obtained groundbreaking results on a wide array of natural language processing tasks (Devlin et al., 2018). This achievement motivates us to explore using a BERT based model to tackle the textual inference problem in the medical domain.

This paper describes the NCUEE (National Central University, Dept. of Electrical Engineering) system for the NLI task of the MEDIQA challenge at the ACL-BioNLP 2019 workshop. Our solution explores a BERT-based model, in which the BiLSTM network with attention mechanism is integrated for textual inference. The input sentence-pair is represented as a sequence of words. Each word refers to distributed vectors from a pre-trained BERT to form as an embedding matrix. The datasets provided by the task organizers are used to train the BiLSTM network with attention model for the prediction task. The output is a value from 0 to 1 representing the estimated class probability. The class with the highest probability (that is, one of entailment, neutral and contradiction) will be regarded as the inference result. Our best accuracy score of 0.84 ranked in the top-third of all 42 submissions in the leaderboard.

Proceedings of the BioNLP 2019 workshop, pages 528–532
Florence, Italy, August 1, 2019.

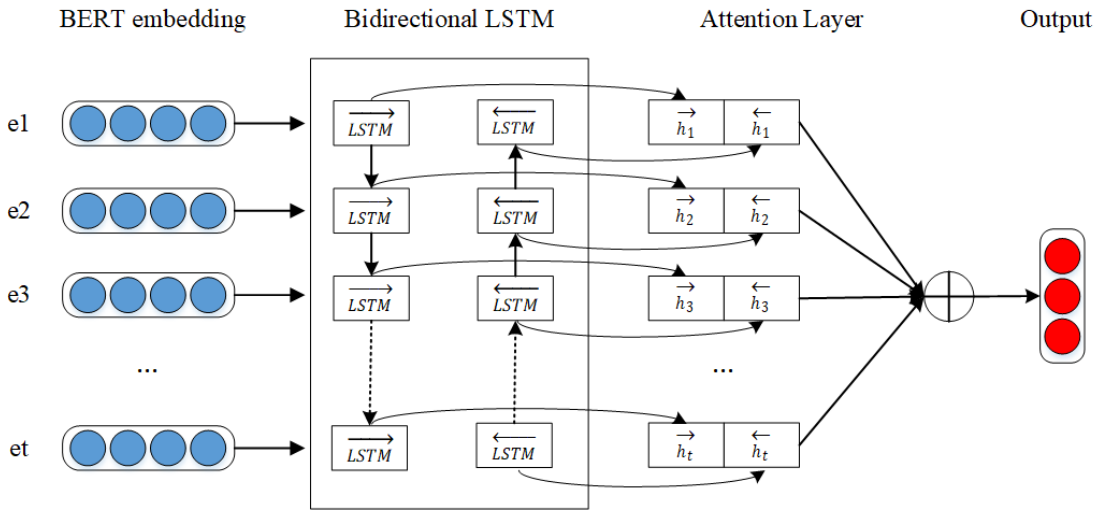

Figure 1: Our BERT-BiLSTM-Attention architecture for the NLI task.

The rest of this paper is organized as follows. Section 2 describes the NCUEE system for the NLI task. Section 3 presents the evaluation results and performance comparisons. Conclusions are finally drawn in Section 4.

2 The NCUEE System

Figure 1 shows our BERT-BiLSTM-Attention architecture for the NLI task. Our model consists of a BERT embedding layer, a BiLSTM layer and an attention layer. An input sentence-pair is represented as a sequence of words. Each word refers to a row looked up in a word embedding matrix from the last layer of the original BERT (Devlin et al., 2018). In this NLI task, we model the sentence-pairs using Bidirectional LSTM (Graves et al., 2013), an extension of the traditional LSTM, to train two LSTMs on the input pairs. The second LSTM is a reversed copy of the first one, so that we can take full advantage of both past and future input features for a specific time step. Consequently, we leverage the word attention mechanism to capture the distinguishing influence of the words and then form a dense vector (Yang et al., 2017). We then use final softmax activation function to classify the sentence-pairs to obtain the probability that belongs to each of the three classes.

During the training phase, if a sentence-pair (premise vs. hypothesis) is true (entailment), the class is assigned as 1, and 0 otherwise (contradiction). If both sentences are neutral without specific relationships, the class is assigned as 2. All training sentence pairs and their accompanying classes are used for training our BERT-BiLSTM-Attention model.

To classify a sentence pair during the test phase, we use the output probability as an indicator for classification. The class with the highest probability will be regarded as the inference result. In addition, ensemble strategies have been widely used in various research fields because of their good performance. This work uses a simple but efficient ensemble strategy called majority voting that involves selecting the class which has a majority, that is, more than half the votes from various trained models.

3 Evaluation

3.1 Data

The datasets were mainly provided by task organizers (Ben Abacha et al., 2019). The sentence-pairs for the NLI task were collected from the MedNLI dataset (Romanov and Shivade, 2018). The training, validation and test datasets were comprised of data from an independent set of sentence-pairs. During the system development phase, the training and validation sets respectively consisted of 12,627 and 1,422 sentence-pairs, for designing and implementing the system. In total, only 405 sentence-pairs in the test dataset were used for final performance evaluation.

The pre-trained word vectors are publicly available for download at the official BERT website [1]. We also used pre-trained weights of BioBERT (Lee et al., 2019), a language representation model for the biomedical domain and publicly available at the GitHub site[2].

3.2 Results

During system development phase, in addition to pre-trained word vectors, we only use the training set to train the system parameters and evaluate the result on the validation set.

In the first set of experiments, the following fine-tuning BERT models were compared to demonstrate their performance for classification.

- BERT-Base, Uncased: 12-layer, 768-hidden, 12-heads, 110M parameters.
- BERT-Base, Cased: 12-layer, 768-hidden, 12-heads, 110M parameters.
- BERT-Large, Uncased: 24-layer, 1024-hidden, 16-heads, 340M parameters.
- BERT-Large, Cased: 24-layer, 1024-hidden, 16-heads, 340M parameters.

Google Research has released the BERT-Base and BERT-Large models (12-layer/24-layer Transformer). Uncased means that the text has been lowercased before WordPiece tokenization and any accept markers have also been removed. Cased means that the true case and accent markers are preserved. The same setups are used for comparisons. The maximum sequence is 128. The training batch size is 32. The learning rate is 2e-5. The number of training epochs is 10.

Table 1 shows the results. The BERT-Large models achieved relatively better accuracy than the BERT-Base models, regardless of case-sensitivity. The best accuracy was obtained by the Uncased BERT-Large model.

In the second set of experiments, the objective is to compare the performance of both BERT and BioBERT models. The BioBERT is based on the same vocabulary as the Cased BERT-Base model. Hence, we selected the pre-trained weights of BioBERT with PubMed 200K and PMC 270K comparing with the Cased BERT-Base model. Note that both models have been fine-tuned with optimal parameter settings.

Table 2 shows the performance comparisons, where the BioBERT outperforms the BERT model, suggesting that the BioBERT model is more suitable for biomedical text mining tasks through incorporating biomedical corpora such as PubMed and PMC.

In the third set of experiments, we evaluated our proposed model based on the previous results. Since the BERT-Large and BioBERT-Base achieved better accuracy, we fine-tuned these two models and sought to identify seek the optimal system parameters. Moreover, we adopted the last layers of these two models as the word embedding to integrate the BiLSTM network with attention mechanism. The setups of the BiLSTM is follows. The hidden size is 256. The dropout rate is 0.5.

Table 3 compares the results. The BioBERT models outperformed the BERT models, regardless of whether BioBERT was used as the word embedding or fine-tuning its original model. In addition, our integrated architecture with the BiLSTM-Attention was found to produce a slight

Models	Accuracy
Fine-tuning BERT-Base, Uncased	0.759
Fine-tuning BERT-Base, Cased	0.751
Fine-tuning BERT-Large, Uncased	0.796
Fine-tuning BERT-Large, Cased	0.793

Table 1: Results of fine-tuning BERT models

Models	Accuracy
Fine-tuning BERT-Base, Cased	0.792
Fine-tuning BioBERT-Base, Cased	0.822

Table 2: Results of BERT vs. BioBERT

Models	Accuracy
Fine-tuning BioBERT-Base	0.822
BioBERT-Base + BiLSTM-Attention	0.824
Fine-tuning BERT-Large	0.809
BERT-Large + BiLSTM-Attention	0.809

Table 3: Results of BERT-BiLSTM with attention

[1] https://github.com/google-research/bert

[2] https://github.com/naver/biobert-pretrained

performance enhancement. The best accuracy was obtained by the BioBERT-Base + BiLSTM-Attention model.

3.3 Comparisons

During final testing phase of the NLI task, we used the training set to train the models and the validation set for parameter optimization. Each participating team was allowed to submit a maximum of 5 runs for each task. We submitted the four abovementioned models accompanying with the ensemble model. For our ensemble strategy, we have trained the models 5 times using the BioBERT-Base + BiLSTM-Attention. The final inference result is the majority voting of the class with the highest probability.

Table 4 shows the results of our testing models. In addition to the BioBERT-Base model, the other models achieved promising accuracy. As expected, our ensemble strategy has the better performance. Our ensemble BioBERT+BiLSTM-Attention model achieved a high accuracy score of 0.84, ranking it in the top-third of all 42 participating teams participated the NLI task in the leaderboard. After excluding invalid submissions, including those did not report their team information (name, affiliation, and so on) and/or submit their working notes papers, our best accuracy score of 0.84 ranked the 11th among all 17 valid submissions.

Models	Accuracy
BioBERT-Base	0.786
BioBERT-Base + BiLSTM-Attention	0.805
BERT-Large	0.805
BERT-Large + BiLSTM-Attention	0.808
Ensemble BioBERT+BiLSTM-Attetion	0.840

Table 4: Results of our testing modes.

The test set (405 instances) is extremely small, but the testing period (15 days) is relatively long. Human intervention can be used to manipulate the results. In addition, the test set is arranged with an obvious pattern. One premise always accompanies with three hypotheses respectively denoting each class (entailment, contradiction and neutral). Base on this observation, it's easy to upgrade the final ranking of this task through a post-editing rule. For example, if two testing instances in the three-class group are predicted as the same class, the former is changed as the class entailment. Consequently, in the same condition, if the class entailment has been determined, then the former is changed as the class contradiction. With this post-editing rule, our best model can be enhanced to achieve a high accuracy score of 0.975.

4 Conclusions

This study describes the NCUEE system in the ACL-BioNLP'19 shared task, including system design, implementation and evaluation. We present our first exploration of this research topic in medical text inference. Future work will exploit other textual features to improve performance.

Acknowledgments

This study is partially supported by the Ministry of Science and Technology, under the grant MOST 108-2218-E-008-017-MY3 and MOST 108-2634-F-008-003- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

References

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Institute of Electrical and Electronics Engineers, pages 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947

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*. Association for Computational Linguistics, pages 1586-1596. https://www.aclweb.org/anthology/D18-1187

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic 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*. Association for Computational Linguistics, pages 670-680. https://www.aclweb.org/anthology/papers/D/D17/D17-1070/

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 2019 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint*. Cornell University, https://arxiv.org/abs/1810.04805v1

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, and Chan So. BioBERT: a pre-trained biomedical language representation model for biomedical text domain. *arXiv Preprint*. Cornell University, https://arxiv.org/abs/1901.08746

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)*. Association for Computational Linguistics, pages 1657-1668. https://www.aclweb.org/anthology/papers/P/P17/P17-1152/

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*. Association for Computational Linguistics, pages 632-642. https://aclweb.org/anthology/papers/D/D15/D15-1075/

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2017. Hierarchical attention networks for document classification. In *Proceedings of the 2017 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1480-1489. https://www.aclweb.org/anthology/N16-1174