UU_TAILS at MEDIQA 2019: Learning Textual Entailment in the Medical Domain

Noha S. Tawfik
Arab Academy for Science & Technology
Alexandria 1029, Egypt
noha.abdelsalam@aast.edu
Utrecht University
3584CC Utrecht, The Netherlands
n.s.tawfik@uu.nl

Marco R. Spruit
Utrecht University
3584 CC Utrecht, The Netherlands
m.r.spruit@uu.nl

Abstract

This article describes the participation of the *UU_TAILS* team in the 2019 MEDIQA challenge intended to improve domain-specific models in medical and clinical NLP. The challenge consists of 3 tasks: medical language inference (NLI), recognizing textual entailment (RQE) and question answering (QA). Our team participated in tasks 1 and 2 and our best runs achieved a performance accuracy of 0.852 and 0.584 respectively for the test sets. The models proposed for task 1 relied on BERT embeddings and different ensemble techniques. For the RQE task, we trained a traditional multilayer perceptron network based on embeddings generated by the universal sentence encoder.

1 Introduction

Detecting semantic relations between sentence pairs is a long-standing challenge for computational semantics. Given two snippets of text: Premise *P* and Hypothesis *H*, textual entailment recognition determines if the meaning of H can be inferred from that of P (Dagan et al., 2013). The significance of modeling text inference is evident since it evaluates the capability of Natural language Processing (NLP) to grasp meaning and interprets the linguistic variability of the language. Natural language inference (NLI) tasks, also known as Recognizing Textual Entailment (RTE) require a deep understanding of the semantic similarity between the hypothesis and the premise. Moreover, they overlap with other linguistic problems such as question answering and semantic text similarity. The recent years witnessed regular organization of shared tasks targeting the RTE/NLI task, which consequently led to advances in the field. More complex models were developed that rely on deep neural networks, this was feasible with the availability of large amounts of annotated datasets such as SNLI and MultiNLI (Bowman et al., 2015; **?**). However, most models fail to generalize across different NLI benchmarks (Talman and Chatzikyriakidis, 2018). Additionally, they do not perform accurately on domain-specific datasets. This is specifically true in the medical and clinical domain. Compared to open domain data, the language used to describe biomedical events is usually complex, rich in clinical semantics and contains conceptual overlap. And hence, it is difficult to adapt any of the former models directly.

The MEDIQA challenge (Ben Abacha et al., 2019) addresses the above limitations through its three proposed tasks. The first task aims at identifying inference relations between clinical sentence pairs and introduces the medical natural language inference benchmark dataset *MedNLI* (Romanov and Shivade, 2018). Its creation process is similar to the creation of the gold-standard SNLI dataset with adaptation to the clinical domain. Expert annotators were presented 4,638 premises extracted from the MIMIC-III database (Johnson et al., 2016) and were asked to write three hypotheses with a true, false and neutral description of the premise. The final dataset comprises 14,049 sentence pairs divided into 11,232, 1,395 and 1,422 for training, development and testing respectively. An additional test batch was provided by the challenge organizers with 405 unlabelled instances. to the biomedical domain.

Similarly, the second task, Recognizing Question Entailment (RQE), tackles the problem of finding duplicate questions by labeling questions based on their similarity (Ben Abacha and Demner-Fushman, 2016). Extending the earlier NLI definition, the authors define question entailment as ”Question A entails Question B if every answer to B is also a correct answer to A exactly or partially”. The dataset is specifically designed to

Proceedings of the BioNLP 2019 workshop, pages 493–499
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

find the most similar frequently asked question (FAQ) to a given question. The training set was constructed from the questions provided by family doctors on the National Library of Medicine (NLM) platform resulting in 8,588 question pairs where 54.2% are positive pairs. For validation, two sources of questions were used: validated questions from the NLM collections and FAQs retrieved from the National Institutes of Health (NIH) website. The validation set has 302 pairs of questions with 42.7% pairs positively labelled. The test set for the challenge was balanced and comprised of 230 question pairs.

The rest of the paper is organized as follows: Section 2 briefly discusses related work. We limit our summary to textual inference research in the biomedical domain only. In Section 3, we describe our proposed model and the implementation details for both tasks. In Section 4, we show the experiment results of our proposed models. Finally, we conclude our analysis of the challenge, as well as some additional discussions of the future directions in Section 5.

2 Related Work

In (Ben Abacha and Demner-Fushman, 2016), the authors introduce a baseline model for the RQE dataset. The feature-based model relies on negation, medical concepts overlap and lexical similarity measures to detect entailment among medical question pairs. Romanov and Shivade conducted multiple experiments on the MedNLI dataset to evaluate the transferability of existing methods in adapting to clinical RTE tasks (Romanov and Shivade, 2018). The best performing was the bidirectional LSTM encoder of the inferSent. Their findings also showed that transfer learning over the larger SNLI set did not improve the results. In a previous work, we tried to model textual entailment found in biomedical literature by restructuring an existing YES/NO question-answering dataset extracted from PubMed(2019). The newly formed dataset aligned with standard NLI datasets format. Further on, we combined hand-crafted features with the inferSent model to detect inference.

To the best of our knowledge, other than the work previously mentioned, there has been minimal research conducted directly on the textual entailment task in the biomedical domain. Below, we summarize scattered attempts to extract contradictions and conflicting statements found in medical documents. Sarafraz et al. (2012), extracted negated molecular events from biomedical literature using a hybrid of machine learning features and semantic rules. Similiarly, De Silve et al. (2017), extracted inconsistencies found in miRNA research articles. The system extracts relevant triples and scores them according to an appositeness metric suggested by the authors. Alamri et al. (2016), introduced a dataset of 259 contradictory claims that answer 10 medical questions related to cardiovascular diseases. Their proposed model relied on n-grams, negation, sentiment and directionality features while in (Tawfik and Spruit, 2018), the authors exploited semantic features and biomedical word embeddings to detect contradictions using the same dataset. Zadrozny et al. (2018) suggested a conceptual framework based on the mathematical sheaf model to highlight conflicting and contradictory criteria in guidelines published by accredited medical institutes. It transforms natural language sentences to formulas with parameters, creates partial order based on common predicates and builds sheaves on these partial orders.

3 Exploratory Embedding Analysis

With the fast developmental pace of text embedding methods, there is a lack of unified methodology to assess these different techniques in the biomedical domain. We attempted to conduct a comprehensive evaluation of different text representations for both tasks, prior to submission of round 2 of the challenge. We use the *MedSentEval*[1] toolkit, a python-based toolkit that supports different embedding techniques including traditional word embeddings like GloVe and FastText, contextualized embeddings like Embeddings from Language Models (ELMO) and Bidirectional Encoder Representations from Transformers (BERT) and dedicated sentence encoders such as inferSent and Universal Sentence Encoder (USE). To evaluate the sentence representations fairly, we adopt a straightforward method that extracts embeddings from different techniques and feeds them to a logistic regression classifier. Our analysis showed that for the NLI task, embeddings from the inferSent model achieved the best performance. This is not surprising, and aligns

[1] `https://github.com/nstawfik/MedSentEval`

with the results reported by the benchmark creator (Romanov and Shivade, 2018). Moreover, we notice that embeddings acquired from language models such as ELMO and BERT, were the second best performing with minimal accuracy difference. For the *RQE* task, the transformer encoder of the USE model outperformed all other methods by a clear margin followed by inferSent trained with GloVe embeddings. This might be contributed to the multi-type training data employed by USE with questions and entailment sentence pairs among others. As observed in the General Language Understanding Evaluation (GLUE) benchmark dataset, BERT-based models are currently the state-of-the art models for the NLI task. Accordingly, we have tried to further investigate the performance of BERT in the biomedical NLI domain. We also employed USE and inferSent sentence embeddings for task 2.

Bidirectional Encoder Representations from Transformers BERT is a neural model developed by Google, that makes heavy use of language representation models designed to pre-train deep bidirectional representations (Devlin et al., 2018). It is trained in an unsupervised manner over an enormous amount of publicly available plain text data. Language Modeling (LM) serves as an unsupervised pre-training stage that can generate the next word in a sentence with knowledge of previous words in a sentence. BERT is different from other LM-based models because it targets a different training objective, it uses masked language modeling instead of traditional LM. It replaces words in a sentence randomly and inserts a ”masked” token. The transformer generates predictions for the masked words by jointly conditioning on both left and right context in all layers.

Universal Sentence Encoder USE is referred to as ”universal” since, in theory, it is supposed to encode general properties of sentences given the large size of datasets it is trained on (Cer et al., 2018). The multi-task learning encoder uses several annotated and unannotated datasets for training. Training data consisted of supervised and unsupervised sources such as Wikipedia articles, news, discussion forums, dialogues and question/answers pairs. It has two variants of the encoding architectures; The transformer model is designed for higher accuracy, but the encoding requires more memory and computational time. The Deep Averaging Network (DAN) model on the other hand is designed for speed and efficiency, and some accuracy is compromised. When integrated in any downstream task, USE should be able to represent sentences efficiently without the need for any domain specific knowledge. This is a great advantage when limited training resources are available for specific tasks.

4 Methods

4.1 Task 1: Natural Language Inference (NLI)

Experimental Settings We take advantage of two newly released BERT models trained on different biomedical data. The following models were initialized from the original bert-base-uncased setting pre-trained with 12 transformer layers, hidden unit size of d=768, 12 attention heads and 110M parameters.

- SciBERT[2] trained on a random sample of 1.14M scientific articles available in the semantic scholar repository. The training data consists of full-text papers from the biomedical and computer sciences domain with a 2.5B and 0.6B word count, respectively (Beltagy et al., 2019).

- ClinicalBERT[3] trained on approximately 2M clinical records. The training data consists of intensive care notes distributed among 15 types available in the MIMIC database (Alsentzer et al., 2019).

We combined both training and evaluation records to form a new training set of 12627 sentence pairs. The original test set was used for evaluation and development. We experimented with all models in pytorch, using the HuggingFace[4] re-implementation of the original BERT python package. We convert the SciBERT models to make it compatible with PyTorch. We use the fine-tuning script to train the model on the MEDNLI dataset in an end-to-end fashion. We trained a total of 30 models with variations of the model configuration. All models with accuracy less than 0.786

[2] The pre-trained weights for for the SciBERT model are available at `https://github.com/allenai/scibert`

[3] The pre-trained weights for the ClinicalBERT model are available at `https://github.com/EmilyAlsentzer/clinicalBERT`

[4] `https://github.com/huggingface/pytorch-pretrained-BERT`

Hyperparameter	Value
Learning rate	3e-5, 2e-5, 5e-5
Sequence length	64, 128
Number of Epochs	3
Batch Size	8, 16

Table 1: Hyperparameters values for training BERT models

on development data were discarded. The threshold value was set to the best accuracy achieved for the MedNLI dataset as reported in the paper. Table 1 list the hyperparameters for this set of experiments, the values for other parameters were kept the same as the original BERT model.

4.1.1 BERT Ensemble Model

Rather than using only a single model for predictions, ensemble techniques can be considered as a useful method to boost the overall performance. A key factor in ensembling is how to blend the results. We experimented with different systems in terms of size and fusion technique in order to increase performance accuracy:

- Drop-out Averaging: All BERT models are added into the candidate ensemble set. Iteratively, we randomly drop one model at a time. With each dropout, we test the ability of the new ensemble set to improve the overall performance by calculating the ensemble's accuracy for the development set by averaging the output probabilities for each class. The process has been repeated until no improvements were observed and the best performing set is chosen as the final ensemble set.

- STACKING BERT 1: A meta learner trained on the predictions generated from all base models and optimally combine them to form the final decision. We train three classifiers, by using five-fold cross validation, including a K-Nearest Neighbor (KNN), a linear Support Vector Machine (SVM) and Naive Bayesian (NB). The classifiers were implemented through the scikit-learn library [5]and we also apply the grid search method for parameter tuning (Pedregosa et al., 2011).

- STACKING BERT 2: We create a second level ensemble stacking. In this level, we train a logistic regression classifier on top of the combined predictions generated from the first level stacking stacking BERT phase.

4.2 Task 2: Recognizing Question Entailment (RQE)

Experimental Settings We use the transformer-based architecture of the USE encoder as it was proven to yield better results. USE was implemented through its TF hub module [6]. For all pairs, each input question was embedded separately and then their combined embedding vector is formed as $(u, v, \mid u - v \mid, u * v)$, which is a concatenation of the premise and hypothesis vectors and their respective absolute difference and hadamard product. We experiment with both logistic regression and multilayer perceptron on top of the generated input representations. The MLP consists of a single hidden layer of 50 neurons using the adam optimizer and a batch size of 64.

5 Results & Discussion

5.1 Task 1: Natural Language Inference(NLI)

The best performing single BERT model achieved 0.828 for the evaluation set. Table 2 shows results of each model ensemble used for the NLI task. For the first run, we only averaged predictions generated by the ClinicalBERT model. The drop-out ensembling resulted in 12 models in total. For the second run, we used KNN classification over predictions from all trained BERT models. The remaining 3 runs use a second level logistic regression classifier while varying the first level classification model. We can observe consistent improvement from successive ensembling from one to two stacking levels. Our five runs showed substantial improvement in the performance over the original baseline with accuracy gain ranging from 10.6% to 13.8%. By the end of the challenge, 42 teams submitted a total of 143 runs to the NLI task. our top performing submission ranked the 12th over all teams [7]. Its corresponding model could be viewed as a three-stage architecture with 2 level stacking ensemble as illustrated in figure 1.

All runs submitted relied solely on BERT text rep-

[5] https://scikit-learn.org/

[6] The TF version of the USE model is available at https://tfhub.dev/google/universal-sentence-encoder-large/3

[7] Leaderboard for the NLI task: https://www.aicrowd.com/challenges/mediqa-2019-natural-language-inference-nli/leaderboards (accessed 1st of June 2019)

Submission	Model	Accuracy	
		Dev	Test
1	Drop-out BERT AVG: 12 models with averaging ensemble	0.836	0.820
2	Stacking BERT 1: KNN	0.846	0.840
3	Stacking BERT 2: KNN followed by LR	0.847	0.847
4	Stacking BERT 2: (KNN/SVM/NB) followed by LR	0.849	0.852
5	Stacking BERT 2: Linear SVM followed by LR	0.846	0.823

Table 2: Results of our team runs on the MEDIQA challenge for the NLI task.

Submission	Model	Accuracy	
		Dev	Test
1	USE embeddings with LR Classifier	0.770	0.584
2	USE embeddings with MLP Classifier (1 hidden layer with 50)	0.778	0.580

Table 3: Results of our team runs on the MEDIQA challenge for the RQE task.

resentations without any external features. Initially, we assumed that training our models with more than just embedding features should help classification and improve overall performance. We used the predictions generated by the drop-out averaging ensemble as extra features to further fine-tune a second-level BERT model. The model hyperparameters settings were the same as the best performing single base model. We did not find this experiment to yield any gains in the evaluation phase, compared to ensemble models, with only 0.815 accuracy for the development set. This was also affirmed post submission, with the release of the gold-labels. The accuracy for the test set was only 0.812.

5.2 Task 2: Recognizing Question Entailment (RQE)

Table 3 shows our two submitted runs for task 2. Even though our approach for this task was much simpler than task 1, we still managed to achieve a considerably good accuracy outperforming the baseline by 4.3%. The final results show that our team ranked the 23rd among all 54 participants[8]. Due to time constraints we were unable to fully investigate all models described in section 3, nor conduct a suitably thorough hyper-parameter search for the MLP. However, we were able to conduct more evaluations post submission. We trained the inferSent Bi-LSTM encoder on the MedNLI data using GloVe embeddings. We then used the trained model to generate embeddings for the RQE data, and used the same MLP architecture to generate predictions. Despite the similarity of both tasks and the potential benefit from transfer learning, the model achieved an accuracy of 0.623 and 0.532 for dev and test set respectively.

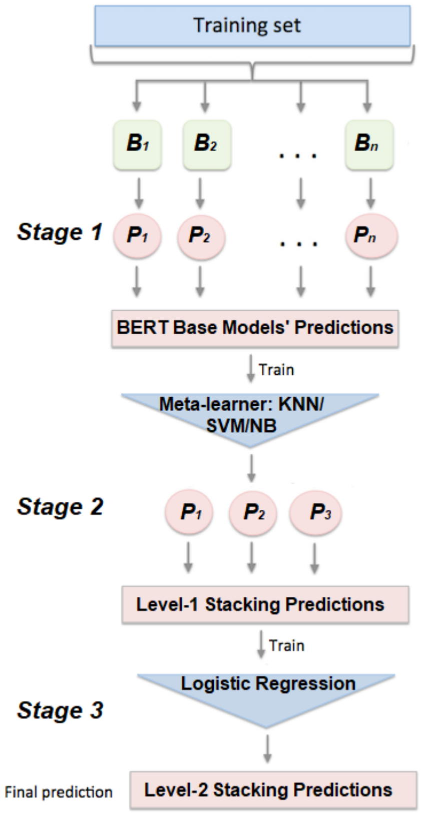

Figure 1: Overview of the ensemble architecture of the best run for the NLI task.

[8] Leaderboard for the RQE task: https://www.aicrowd.com/challenges/mediqa-2019-recognizing-question-entailment-rqe/leaderboards (accessed 1st of June 2019)

6 Conclusion

In this paper, we presented our solution for textual entailment detection in the clinical domain. Our proposed approach for the NLI task relies on BERT contextual embeddings features and machine learning algorithms such as KNN, SVM and LR for ensembling. We use two different pretrained BERT weights to train the base models and generate corresponding probabilities for the test set. Then, we adopt a 5-fold stacking strategy to learn and combine predictions. In the third and final level of the ensemble, we use a logistic regression over the outputs from level-1 stacking, to predict the final class labels. A future extension of our model is to use BERT in feature extraction mode instead of fine-tuning the end-to-end model on the MedNLI dataset. This would allow the selection of layers from which to extract embeddings and/or the combination of multiple layers. In the former scenario, different neural networks could be used to generate the base model predictions before applying ensemble techniques.

For the RQE task, we train an MLP classifier on top of USE embeddings. The results obtained were promising, given the simplicity of the model. More complex and deeper networks could be employed with the combination of USE embeddings. We also experimented with transfer learning by training the inferSent model on MedNLI before fine-tuning on the RQE corpus. While this approach did not improve the results, we aim at further investigating other inferSent architectures and training on clinical word embedding.

References

Abdulaziz Alamri. 2016. *The Detection of Contradictory Claims in Biomedical Abstracts*. Ph.D. thesis, University of Sheffield.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *arXiv e-prints*, page arXiv:1904.03323.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. SCIBERT: Pretrained Contextualized Embeddings for Scientific Text. Technical report.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing Question Entailment for Medical Question Answering. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2016:310–318.

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 *ACL-BioNLP*, Florence.

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 (EMNLP).*, Lisbon. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil Google Research Mountain View. 2018. Universal Sentence Encoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*.

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

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Noha S. Tawfik and Marco R. Spruit. 2019. Towards Recognition of Textual Entailment in the Biomedical Domain. In *International Conference on Applications of Natural Language to Information Systems*, Manchester. Springer.

Farzaneh Sarafraz. 2012. *Finding conflicting statements in the biomedical literature*. Ph.D. thesis, University of Manchester.

Nisansa de Silva, Dejing Dou, and Jingshan Huang. 2017. Discovering inconsistencies in pubmed abstracts through ontology-based information extraction. In *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)*.

Aarne Talman and Stergios Chatzikyriakidis. 2018. Testing the Generalization Power of Neural Network Models Across NLI Benchmarks. Technical report.

Noha S. Tawfik and Marco R. Spruit. 2018. Automated Contradiction Detection in Biomedical Literature. In *Proceedings of international Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 138–148. Springer, Cham.

Wlodek Zadrozny and Luciana Garbayo. 2018. A Sheaf Model of Contradictions and Disagreements. Preliminary Report and Discussion. In *International Symposium on Artificial Intelligence and Mathematics,*, Florida.