NCUEE-NLP at MEDIQA 2021: Health Question Summarization Using PEGASUS Transformers

Lung-Hao Lee, Po-Han Chen, Yu-Xiang Zeng, Po-Lei Lee, and Kuo-Kai Shyu
Department of Electrical Engineering, National Central University, Taiwan
Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan

Abstract

This study describes the model design of the NCUEE-NLP system for the MEDIQA challenge at the BioNLP 2021 workshop. We use the PEGASUS transformers and fine-tune the downstream summarization task using our collected and processed datasets. A total of 22 teams participated in the consumer health question summarization task of MEDIQA 2021. Each participating team was allowed to submit a maximum of ten runs. Our best submission, achieving a ROUGE2-F1 score of 0.1597, ranked third among all 128 submissions.

1 Introduction

Consumers increasingly use online resources to meet their health information needs. However, health information needs are usually complex and to be expressed in natural language (Kilicoglu et al., 2018). In general, health questions tend to consist of considerable contextual information that may hinder automatic Question Answering (QA) systems. Paraphrasing and summarizing the questions has been shown to substantially improve QA performance (Ben Abacha and Demner-Fushman, 2019a). Therefore, effective summarization methods for consumer health questions could play an important role in enhancing medical QA performance.

Automatic text summarization is the process of computationally shortening texts to find or generate the most informative sentences that represent the most important or relevant information within the original content. There are two general approaches to summarization: extraction and abstraction. In extractive summarization methods, a summary is extracted from the original texts, but the extracted sentences are not modified in any way. Abstractive summarization methods learn a semantic representation of the original content, and then use this representation to generate a summary that is closer to what a human might express in terms of original content.

MEDIQA 2021 is the second edition of the MEDIQA challenge collocated with the BioNLP 2021 workshop, focusing on summarization in the medical domain with three tasks: consumer health question summarization, multi-answer summarization, and radiology report summarization. We only participated the first Question Summarization (QS) task, in the domain of abstractive summarization. The goal of this task is to promote the development of new summarization methods that specifically address the challenges of long and complex consumer health questions. The recently developed transformer in NLP is a novel neural architecture that aims to solve sequence-to-sequence tasks in handling long dependencies and usually achieves promising results. This achievement motivates us to explore the use of a transformer-based model to tackle the question summarization problem in the medical domain.

This paper describes the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing Lab) system for the QS task of the MEDIQA challenge at the BioNLP 2021 workshop. Our solution explores the use of pre-trained PEGASUS Transformers (Zhang et al., 2020a) and fine-tuning on the downstream question summarization task using our collected and processed datasets. A total of 22 teams participated in this task. Each participating team was allowed to submit a maximum of ten runs. Our best submission had a ROUGE2-F1 score of 0.1597, ranking third among all 128 submissions.

Proceedings of the BioNLP 2021 workshop, pages 268–272

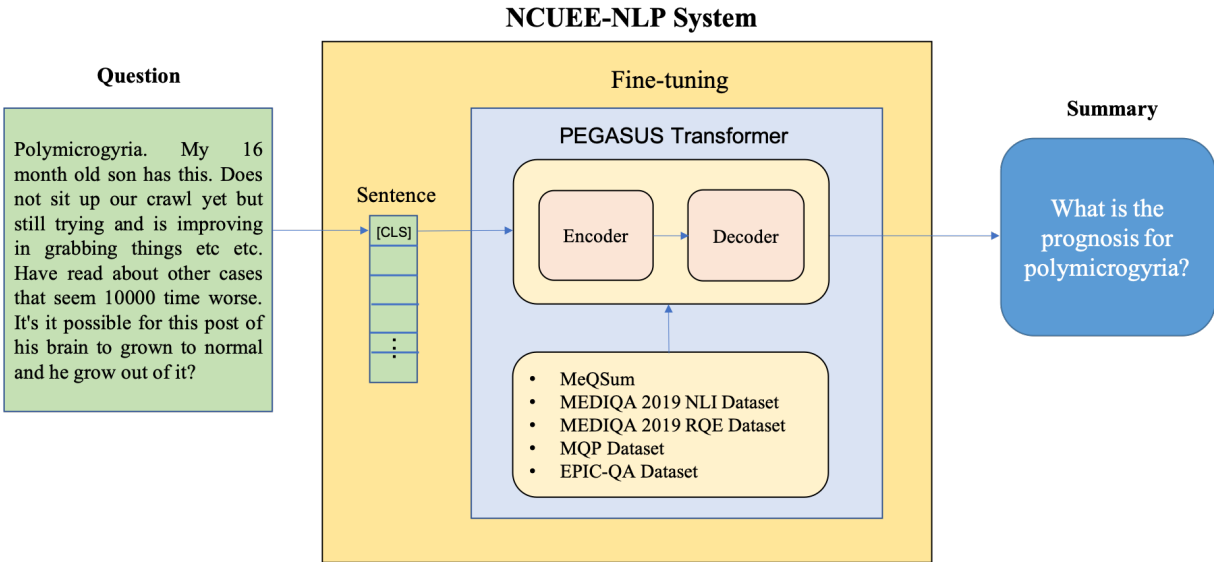

Figure 1: Our NCUEE-NLP system architecture for the QS task.

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

2 The NCUEE-NLP System

Figure 1 shows our NCUEE-NLP system architecture for the QS task. Specifically, our system is comprised of two main parts: 1) PEGASUS transformers, and 2) fine-tuning. Details are introduced as follows.

2.1 PEGASUS Transformers

Zhang et al. (2020a) proposed **PEGASUS** (**P**retraining with **E**xtracted **G**ap-sentences for **A**bstractive **SU**mmarization **S**equence-to-sequence) method that pre-trains large transformer-based encoder-decoder models on massive text corpora. New self-supervised objectives called Gap Sentences Generation (GSG) and classical Mask Language Models (MLM) were applied simultaneously as pre-training objectives. The PEGASUS model was evaluated on 12 downstream summarization tasks spanning news, science, stories, instructions, emails, patents, and legislative bills. Experimental results showed that good abstractive summarization performance can be achieved across broad domains by fine-tuning the model, outperforming previous state-of-the-art approaches on many tasks.

These achievements motivate us to explore the use of the PEGASUS transformers and fine-tuning on the downstream QS task in the medical domain.

2.2 Fine-tuning

Many summarization datasets contain original texts and their referenced summarizes written in declarative sentences. Question summaries written in interrogative sentences are relatively rare. Hence, in addition to the training set provided by task organizers, we also collected and processed the following datasets to fine-tune the QS task.

- MEDIQA 2019 – NLI Dataset (Ben Abacha et al., 2019)

The Natural Language Inference (NLI) task of the MEDIQA 2019 challenge identifies three relations between two sentences including entailment, neutral, and contradiction. We only use the entailment relation that was annotated from the training, validation and test datasets. Comparing the lengths of two the sentences in each pair, the longer sentences will be regarded as a question, while the other is used as the corresponding summary. A total of 4,683 pairs were collected from this dataset.

- MEDIQA 2019 – RQE Dataset (Ben Abacha et al., 2019)

The Recognizing Question Entailment (RQE) task of the MEDIQA 2019 challenge focuses on identifying entailments between two questions. We use the positive question-pairs (annotated as "entailment") from the training, validation and test datasets. However, some questions are not written using valid interrogative sentences such as a declarative sentences followed by "Right?". We exclude these cases and only use questions that start with wh-words, be verbs, and auxiliary verbs. Similarly, the shorter question in each question-pairs is regarded as a reference summary. This resulted in a final subset of 4,011 pairs.

- MQP Dataset (McCreery et al., 2020)

The Medical Question Pairs (MQP) dataset contains similar and dissimilar medical question pairs hand-generated and labeled by doctors. A list of 1,524 patient-asked questions were randomly sampled. Doctors as the labelers had rewritten the original question in different ways while maintaining the same intent, and used similar key words to write related but dissimilar questions for which the answer would be wrong or irrelevant. Hence, each question results in one similar and one different pair. We only use the similar question pairs to fine-tune the transformers. In the same way, the longer questions are used as original questions and the shorter ones are their reference summaries.

- EPIC-QA Dataset on COVID-19 (Goodwin et al., 2020)

In response to the COVID-19 pandemic, the Epidemic Question Answering (EPIC-QA) track in TREC 2020 conference focuses on developing systems capable of automatically answering questions about COVID-19. In the question part of EPIC-QA data, two prepared sets of approximately 45 questions were provided: one for expert-level questions and one for consumer-level questions. Without considering the question levels, we regard the longer questions as original questions and the corresponding shorter question are their summaries.

3 Evaluation

3.1 Data

The experimental datasets were mainly provided by task organizers (Ben Abacha et al., 2021). The training, validation and test sets were composed of data from an independent set of consumer health questions. The MeQSum Dataset of consumer health questions and their summaries can be used for training (Ben Abacha and Demner-Fushman, 2019b). The validation and test sets consist of consumer health questions received by the U.S. National Library of Medicine (NLM) in December 2020. Their associated summaries were manually created by medical experts for evaluation.

In summary, during the system development phase, the training and validation sets respectively consisted of 1,000 and 50 consumer health questions and their associated summaries for system designing and implementation. In total, only 100 consumer health questions in the test dataset were used for final performance evaluation.

3.2 Settings

The pre-trained PEGASUS models were downloaded from the HuggingFace (Wolf et al., 2019). A PEGASUS model was trained with sampled gap sentence ratios on both C4 (Raffel et al., 2020) and HugeNews datasets, and important sentences were sampled stochastically. We selected the PEGASUS-Large model and its mixed and stochastic model (denoting PEGASUS-Large-XSum) on the XSum (Narayan et al., 2018) datasets, containing 227k BBC news articles from 2010 to 2017 covering a wide variety of subjects along with professionally written single-sentence summarizes.

To confirm model performance, we compared the previous state-of-the-art BART method (Lewis et al., 2019) that uses a denoising autoencoder to pre-train sequence-to-sequence models. We also downloaded the pre-trained BART-Large and BART-Large-XSum models from the HuggingFace (Wolf et al., 2019).

On an Nvidia DGX-1 server using a V100 GPU with the same settings, the hyper-parameter values for our model implementation were optimized as follows: maximum sequence length 512; learning rate 0.00005; batch size 6 and gradient accumulation steps 128 for both BART models; and batch size 8 and gradient accumulation steps 512 for both PEGASUS models.

3.3 Metrics

ROUGE is used to measure summarization performance (Lin, 2004). ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation,

Models	R1-P	R1-R	R1-F1	R2-P	R2-R	R2-F1	RL-P	RL-R	RL-F1
BART-Large	0.3165	0.3255	0.3209	0.1355	0.1438	0.1395	0.3090	0.3182	0.3135
BART-Large-XSum	**0.3299**	0.3194	0.3246	**0.1435**	0.1488	0.1461	**0.3215**	0.3127	**0.3170**
PEGASUS-Large	0.3153	**0.3368**	**0.3257**	0.1307	**0.1593**	0.1436	0.3029	**0.3285**	0.3152
PEGASUS-Large-XSum	0.3159	0.3269	0.3213	0.1393	0.1553	**0.1469**	0.3017	0.3157	0.3085

Table 1: Results of summarization models on the QS validation dataset.

Models	R1-P	R1-R	R1-F1	R2-P	R2-R	R2-F1	RL-P	RL-R	RL-F1
BART-Large	0.3526	0.3159	0.3132	0.1452	0.1236	0.1268	0.3187	0.2865	0.2842
BART-Large-XSum	0.3308	0.3253	0.3116	0.1212	0.1150	0.1125	0.2976	0.2891	0.2784
PEGASUS-Large	0.3173	**0.3426**	0.2936	0.1377	0.1346	0.1217	0.2821	0.2934	0.2579
PEGASUS-Large-XSum	**0.3869**	0.3316	**0.3352**	**0.1850**	**0.1573**	**0.1597**	**0.3576**	**0.3030**	**0.3090**

Table 2: Results of summarization models on the QS test dataset.

including several automatic evaluation methods that measure the similarity between summaries. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-L accounts for the union Longest Common Sequence (LCS) in matching between a reference summary sentence and every candidate summary sentence.

In the QS task of MEDIQA 2021 challenge, ROUGE-1 (denoted as R1), ROUGE-2 (R2), and ROUGE-L (RL) were adopted as measure metrics. The F1 score, which is a harmonic mean of precision (short in P) and recall (R), of R2 was regarded as the official score to rank the participating teams' performance in the leaderboard.

3.4 Results

Table 1 shows the results on the QS validation set of MEDIQA 2021 challenge. Both PEGASUS models outperformed the BART models in a half of the metrics. The mixed and stochastic models on the XSum datasets usually outperformed than those without the XSum optimization using both BART and PEGASUS transformers. The PEGASUS-Large-XSum model obtained the best overall score of 0.1469 in R2-F1, considered as the ranking metric.

During the final testing phase of the QS task, we used the training set and collected datasets to fine-tune the models and the validation set for parameter optimization. Each participating team was allowed to submit a maximum of ten runs for each task. We submitted the four above-mentioned models. Table 2 shows the results of our testing models. The PEGASUS-Large-XSum model clearly outperformed the others than the others in almost all evaluation metrics.

A total of 22 teams participated in the QS task, each submitting at least one entry. Our best submission achieved an R2-F1 score of 0.1597, significantly outperforming the baseline model with a score of 0.1373 and ranking third place among all 128 submissions.

In addition to ROUGE metrics, task organizers also use several evaluation metrics that may be better adapted to the QS task. Our best submission also achieved a HOLMS score (Mrabet and Demner-Fushman, 2020) of 0.5783, ranking first among all 128 submissions. Our best submission had a BERTScore-F1 (Zhang et al., 2020b) of 0.6960, ranked ninth among all submissions.

4 Conclusions

This study describes the NCUEE-NLP system in the consumer health question summarization task of the MEDIQA 2021 challenge, including system design, implementation and evaluation. We used the PEGASUS transformers and fine-tuned the downstream summarization task using our collected and processed datasets. A total of 22 teams participated in the task, each submitting at least one entry. Our best submission had a ROUGE2-F1 score of 0.1597, ranking third place among all 128 submissions.

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

Asma Ben Abacha, and Dina Demner-Fushman. 2019a. On the role of question summarization and information source restriction in consumer health question answering. *AMIA Summits on Translational Science Proceedings*, 2019:117-128.

Asma Ben Abacha, and Dina Demner-Fushman. 2019b. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 2228-2234. http://dx.doi.org/10.18653/v1/P19-1215

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, pages 370-379. http://dx.doi.org/10.18653/v1/W19-5039

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain, In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing*, NAACL-BioNLP 2021, Association for Computational Linguistics.

Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatrain. 2020. Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3458-3465. https://doi.org/10.1145/3394486.3412861

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*. Association for Computational Linguistics, pages 74-81.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1-67.

Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E. Shooshan, Laritza Rodriguez, Kate Masterton, and Dina Demner-Fushman. 2018. Semantic annotation of consumer health questions. BMC Bioinformatics, 19, 34(2018). https://doi.org/10.1186/s12859-018-2045-1

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:11328-11339.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 7871-7880. http://dx.doi.org/10.18653/v1/2020.acl-main.703

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1797-1807. http://dx.doi.org/10.18653/v1/D18-1206

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's transformers: state-of-the-art natural language processing. *arXiv preprint*. https://arxiv.org/abs/1910.03771

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: evaluating text generation with BERT. Published as a conference paper at ICLR 2020. *arXiv preprint*. https://arxiv.org/abs/1904.09675

Travis Goodwin, Dina Demner-Fushman, Kyle Lo, Lucy Lu, William R. Hersh, Hoa T. Dang, and Ian M. Soboroff. 2020. EPIC-QA dataset on COVID-19. https://bionlp.nlm.nih.gov/epic_qa/

Yassine Mrabet, and Dina Demner-Fushman. 2020. HOLMS: alternative summary evaluation with large language models. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, pages 5679-5688. http://dx.doi.org/10.18653/v1/2020.coling-main.498