Utilizing Transformer for Sentence Similarity Modeling and Answer Generation in the Machine Reading Comprehension Task

Arafat Habib Quraishi
Department of CSE
Leading University, Sylhet, Bangladesh

Alak Kanti Sarma
Sylhet, Bangladesh

ABSTRACT

In the Machine Reading Comprehension (MRC) task, the objective is to understand the given text in order to answer questions about it. For this task, selecting relevant passages or sentences is an important step. In this paper, we utilize the pre-trained transformer encoder for sentence similarity modeling to select rel- evant passage(s) or sentence(s) for the MRC problem and then we fed the relevant passage(s) or sentence(s) to a question answering model for answer generation. Experimental results in the Microsoft MAchine Reading COmpre- hension (MS-MARCO) dataset for the QA task show that our proposed approach is effective to generate abstractive answers in the MRC task.

Keywords

Machine Reading Comprehension

1. INTRODUCTION

In recent years, messenger based intelligent sys- tems such as Chatbots or virtual assistants have gained significant popularity which increases the demand to build conversational agents using state- of-the-art technologies [18, 1]. To improve the performance of such conversational agents, good MRC capabilities are required in these systems. However, MRC is a challenging task since in such tasks it is required to well understand the document at first to generate the correct answer [1].

In the MRC task, to answer a given question, one straightforward approach would be to extract the relevant answer span from the source document. However, for conversational agents, just answering the question by extracting a span of text from the source document is not enough [18, 1]. Because a good conversational agent should be able to answer the given question or the query in natural language. Thus, in order to build such systems, along with understanding the document(s), the requirement is also to answer the question with a well-formed response [18, 1].

In this paper, we address the MRC problem to generate abstractive answers instead of directly ex- tracting the answer from the source document. We conduct extensive experiments with our proposed approach and obtain impressive performance for the question answering (QA) task in the MRC problem.

2. RELATED WORK

Early work in the MRC task relies on single- passage scenarios [19]. However, the single- passage MRC task is not adequate in real world. Because the real world setting involves search engines, whereas for each question the search en- gines retrieve multiple passages and the MRC mod- els generate the final answer from the most rele- vant one. With Microsoft introducing the MAchine Reading COmprehension dataset (MS-MARCO) for the MRC task in multi-passage scenarios [1], several research has been done in recent years in this dataset [22, 25, 18, 10]. One notable feature of the MS-MARCO dataset is that it also contains a well-formed QA task where the generated answers are required to be in natural language [1].

In the multi-passage MRC task for QA, the neu- ral network-based attention mechanism has been effectively used in recent years to generate answers based on relevant contexts [22, 25, 18]. How- ever, most of these models [22, 25] tend to extract answer spans from the relevant passage directly. One notable exception is the work of Nishida et al. [18], where they proposed an attention based neural encoder-decoder architecture via utilizing Pointer Generation Network [21] and Transformer

[23] which could generate the answer in natural language. They also addressed that better passage ranking model would be the key to improve the per- formance of the MRC task in multi-passage scenar- ios and demonstrated that with gold passage rank- ing, their proposed approach could significantly improve the performance. In recent years, models based on BERT showed significant improvement in different sentence ranking tasks including the an- swer ranking task [3, 15, 4, 12, 9, 8, 10]. However, such BERT-based models are not deeply investigated for passage ranking task yet.

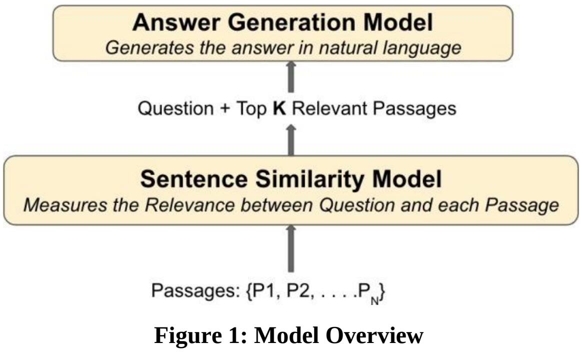
Answer Generation Model
Generates the answer in natural language
Question + Top K Relevant Passages
Sentence Similarity Model
Measures the Relevance between Question and each Passage
Passages: {P1, P2, ...P_N}

Figure 1: Model Overview

One variant of the answer generation task is the query focused abstractive text summarization (QFATS) task [17, 2, 7]. In the QFATS task, the summaries are required to be generated in natural language based on the given query. However, most of the previous work for abstractive summarization relied on some attention based encoder-decoder models for only generic summarization task with- out considering any query relevance [20, 16, 21]. In such encoder-decoder models, the encoder plays the vital role to generate the document represen- tations.

]

Whereas the decoder generates the sum- maries based on these document representations. It should be pointed out that all work stated above used different recurrent neural network-based mod- els. The recently proposed BERTSUM model [5] used BERT [3] as the encoder and the transformer decoder [23] was utilized as the decoder for ab- stractive summarization. Though the transformer- based BERTSUM model was initially used only for generic abstractive summarization, later on the QR- BERTSUM-TL model was proposed by Laskar et al., [7] in which the query relevance was also taken into account for query focused abstractive sum- mary generation. In this work, we also leverage the QR-BERTSUM-TL model for answer genera- tion in the MRC task. Our experimental findings show that ranking the relevant passage(s) using BERT followed by generating the answers using QR-BERTSUM-TL is effective to generate the abstractive answers in the MRC task.

3. PROPOSED METHOD

Suppose, we have a query $Q = q_1, q_2, ..., q_n$ containing n words and a set of passages $P = P_1, P_2, ...,P_N$ containing N passages. For the QA task in the MRC problem, if K passages among

these N passages are relevant, then the goal for this task is to use these K passages to generate an abstractive answer $y_{ag} = y_1, y_2, ...y_m$ containing$m$ words.

In Figure 1, the overall architecture of our model is shown. In the following, we first describe our similarity modeling framework, then we describethe answer generation model.

3.1 Sentence Similarity Model

Suppose, we want to measure the relevance be- tween the two sequences $X = x_1, x_2, ..., x_m$ and $Y = y_1, y_2, ..., y_n$. The sequence X can be consid- ered as the question whereas the sequence Y canbe considered as a whole passage.

To measure the relevance between two se- quences, we propose the Question Passage Sim- ilarity Modeling (QPSM) model to measure the similarity of a passage with the given question to select the relevant passages. For this task, first N independent passages are given to generate the ab- stractive answer from the most relevant passage(s) among them. Afterward, we select the top Kp rele- vant passages from these N passages based on oursimilarity modeling framework.

For the similarity modeling task, we utilize the BERT [3] architecture. Models based on BERT have shown impressive performance in different NLP tasks, including sentence similarity model- ing for the answer selection task [6, 5, 4]. Thus, we also utilize BERT architecture in our similarity modeling framework by adopting the Robustly Op- timized BERT Pretraining Approach (RoBERTa)

[15] model.

In the RoBERTa model, sentence pairs are com- bined together into a single sequence. These sen- tences are separated by a special token [*SEP*]. Moreover, an additional token [*CLS*] is added atthe beginning of the combined sequence. Onlythe representation of the first token ([*CLS*]) of the combined sequence is considered as the aggregate representation of the whole sequence. During the fine-tuning stage, new parameters are added for anadditional classification layer W. All parameters of the pre-trained RoBERTa model and the additional parameters W are fine-tuned jointly to maximize the log-probability of the correct label. The label probabilities $P \epsilon R^K$ (where K is the total number of classifier labels) are calculated as follows:

$$P = softmax(CW^T) \quad (1)$$

In the passage ranking task for answer gener- ation, there are two classifier labels (similar = 1, dissimilar = 0). In the original RoBERTa model [15], sentence pair classification task was done bypredicting the correct label (1 or 0). But in our work, we only consider the predicted score P_{tr} forthe similarity label to rank relevant passages basedon similarity score, similar to this work [12, 9, 6].

$$P_{tr} = P(C = 1|X, Y) \quad (2)$$

3.2 Answer Generation Model

To generate the abstractive answers, we adopt the BERTSUM model which could generate abstrac- tive summaries for generic summarization [14]. However, this model cannot consider the query relevance. To incorporate the query relevance in BERTSUM, we follow the work of Laskar et al. [7, 8, 10] where they concatenate the query with the document and give them as input to the BERT encoder of the model to generate query focused summaries. We utilize it similarly for the QA task via containing the question and the relevant pas- sage(s) for ANSwer GENeration (ANSGEN). We denote our proposed model as the BERTANSGEN model. Note that a similar approach of concate- nating the question with the document has been found to work well on the answer generation tasks in some other neural network models [13].

4. EXPERIMENTAL SETUP

We now describe the datasets used in this paper, followed by the evaluation metrics that we used to evaluate our approach. Finally, we describe the training parameters used in our experiments.

4.1 Dataset

We use the MS-MARCO (v2.1) dataset [1] which contains two types of task: i) QA-ALL and ii) QA- NLG (v2.1). The difference between these two versions is that the QA-ALL dataset can contain various types of answers, whereas the QA-NLG is a subset of the QA-ALL dataset in which the generated answers are well-formed and written in natural language. Since our goal is to generate ab- stractive answers, we leverage the QA-NLG dataset in which the training set contains 153725 queries. There can be upto 10 candidate passages for each query in this dataset.

Table 1: Performance in the QA-NLG Development set of MS-MARCO

Model	ROUGE-L	BLEU-1
QPSM-BERT$_{ANSGEN}$ (K=3)	63.21	57.56
QPSM-BERT$_{ANSGEN}$ (K=1)	63.50	58.70

4.2 Evaluation Metrics

For the QA task, we used the official MS-MARCO evaluation script1 and reported the ROUGE-L (F1) and BLEU-1 score, similar to the prior work on this dataset [18].

4.3 Training and Parameter Settings

For the sentence similarity model, we imple- mented RoBERTa [15] using the HuggingFace Transformer2 in PyTorch [24] and kept the hyper- parameters similar to the CETE model [12]. For the answer generation model, we used the cased3 version of the BERTBase model to fine-tune our encoder. Maximum number of tokens considered for each input sequence was 256. We used cross entropy loss function to calculate the loss. The Adam was used as the optimizer. As the encoder is the pre-trained BERT model and decoder is a randomly initialized transformer decoder, we fol- lowed the approach of the Liu and Lapata [14] to use different learning rates for encoder and decoder. The learning rate was set to 2 x 10-3 for the encoder and 0.1 for the decoder. The batch size was set to 140, with maximum 80000 training steps. In the decoder, we utilized the beam search strategy with length = 5.

5. RESULTS AND DISCUSSIONS

In this section, we analyze the effectiveness of our proposed QPSM-BERTANSGEN model. Note that we also perform an ablation test to further vali- date the effectiveness of different techniques used within our model.

5.1 Performance on the MS-MARCO dataset

We conduct experiments with two different values of K to select the top K passages: K = 1, and K = 3. From Table 1, we see that our proposed model obtains good ROUGE and BLEU scores for answer generation. We find that retrieving only one top candidate passage (K = 1) is more effective than retrieving multiple candidate passages (K = 3). We observe that our model with K = 1 outperforms its other variant K = 3 with an improvement of 0.46% in terms of ROUGE-L and an improvement of 1.63% in terms of BLEU-1.

Table 2: Ablation Test in the QA-NLG Development set of MS-MARCO. Here, 'ROUGE' is denoted by 'R', whereas 'BLEU' is denoted by 'B'.

Model	R-L	B-1
QPSM-BERT$_{\text{ANSGEN}}$ (K=1)	**0.635**	**0.587**
w/o Similarity Modeling (K=ALL)	0.540	0.519
w/o Query Relevance (K=1)	0.508	0.479

5.2 Ablation Test

We perform ablation test to further investigate the effectiveness of the QPSM model for passage ranking as well as the incorporation of the question with the relevant passage(s) in BERTSUM for answer generation. From Table 2, we find that removing the similarity modeling for passage ranking degrades the performance by 14.96% in terms of ROUGE-L. The performance deterioration is even more (by 20%) if query relevance is not incorporated.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrate how we can utilize similarity modeling in the multi-passage MRC task for abstractive answer generation. Experimental results show that our proposed approach is effective for the QA task in the MRC problem. In future, we will study how to improve the performance of the passage retrieval and the answer generation models. We will also study how we can utilize transformer models in industrial big data scenarios [11], as well as to improve performance in other tasks, such as entity linking [26], punctuation restoration [27], etc.

7. REFERENCES

[1] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, et al. Ms marco: A human generated ma- chine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[2] T. Baumel, M. Eyal, and M. Elhadad. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and sum-mary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*, 2018.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transform- ers for language understanding. In *Proceedings of the 2019 Conference of the North American Chap- ter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[4] S. Garg, T. Vu, and A. Moschitti. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. *arXiv preprint arXiv:1911.04118*, 2019.

[5] T. Lai, Q. H. Tran, T. Bui, and D. Kihara. A gated self-attention memory network for answer selection. In *Proceedings of the 2019 Conference on EmpiricalMethods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5955–5961, 2019.

[6] M. T. R. Laskar, E. Hoque, and J. Huang. Utiliz- ing bidirectional encoder representations from trans- formers for answer selection task. *The V AMMCS International Conference: Extended Abstract*, 2019.

[7] M. T. R. Laskar, E. Hoque, and J. Huang. Query focused abstractive summarization via incorporat- ing query relevance and transfer learning with trans- former models. In *Canadian Conference on Artificial Intelligence*, pages 342–348. Springer, 2020.

[8] M. T. R. Laskar, E. Hoque, and X. Huang. WSL- DS: Weakly supervised learning with distant super- vision for query focused multi-document abstractive summarization. In *Proceedings of the 28th Inter- national Conference on Computational Linguistics*, pages 5647–5654, 2020.

[9] M. T. R. Laskar, E. Hoque, and J. Xiangji Huang. Utilizing bidirectional encoder representations from transformers for answer selection. In *International Conference on Applied Mathematics, Modeling and Computational Science*, pages 693–703. Springer, 2019.

[10] M. T. R. Laskar, E. Hoque, and J. Xiangji Huang. Domain adaptation with pre-trained transformers forquery focused abstractive text summarization. *arXiv e-prints*, pages arXiv–2112, 2021.

[11] M. T. R. Laskar, J. X. Huang, V. Smetana, C. Stew- art, K. Pouw, A. An, S. Chan, and L. Liu. Extending isolation forest for anomaly detection in big data viak-means. *ACM Transactions on Cyber-Physical Sys-tems (TCPS)*, 5(4):1–26, 2021.

[12] M. T. R. Laskar, X. Huang, and E. Hoque. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5505–5514, 2020.

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettle- moyer. Bart: Denoising sequence-to-sequence pre- training for natural language generation, translation, and comprehension. In *Proceedings of the 58th An- nual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[14] Y. Liu and M. Lapata. Text summarization with pre- trained encoders. In *Proceedings of the 2019 Confer- ence on*

Empirical Methods in Natural Language Pro- cessing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3721–3731, 2019.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining ap- proach. *arXiv preprint arXiv:1907.11692*, 2019.

[16] R. Nallapati, B. Zhou, C. dos Santos, Ç. GuÌ‡lçehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Pro- ceedings of The 20th SIGNLL Conference on Compu- tational Natural Language Learning*, pages 280–290. Association for Computational Linguistics, 2016.

[17] P. Nema, M. M. Khapra, A. Laha, and B. Ravindran. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1063–1072. Association for Computational Linguistics, July 2017.

[18] K. Nishida, I. Saito, K. Nishida, K. Shinoda, A. Ot-suka, H. Asano, and J. Tomita. Multi-style generative reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2273–2284, 2019.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehen- sion of text. *arXiv preprint arXiv:1606.05250*, 2016.

[20] A. M. Rush, S. Chopra, and J. Weston. A neural at-tention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empiri- cal Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics, 2015.

[21] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator net- works. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083. Association for Computational Linguis-tics, 2017.

[22] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehen- sion. *arXiv preprint arXiv:1611.01603*, 2016.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. De- langue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Fun- towicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[25] M. Yan, J. Xia, C. Wu, B. Bi, Z. Zhao, J. Zhang, L. Si, R. Wang, W. Wang, and H. Chen. A deep cascade model for multi-document reading compre- hension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7354–7361,2019.

[26] M. T. R. Laskar, C. Chen, A. Martsinovich, J. John- ston, X.-Y. Fu, S. B. TN, and S. Corston-Oliver. Blink with elasticsearch for efficient entity linking in business conversations. arXiv preprint arXiv:2205.04438, 2022.

[27] X.-Y. Fu, C. Chen, M. T. R. Laskar, S. Bhushan, and S. Corston-Oliver. Improving punctuation restoration for speech transcripts via external data. In Proceedings of the Seventh Workshop on Noisy User- generated Text (W-NUT 2021), pages 168–174, 2021.