

# NCU-IISR/AS-GIS: Using BERTScore and Snippet Score to Improve the Performance of Pretrained Language Model in BioASQ 10b Phase B

Hao-Hsuan Ting <sup>1</sup>, Yu Zhang <sup>2</sup>, Jen-Chieh Han <sup>2</sup> and Richard Tzong-Han Tsai <sup>2,3,4,5</sup>

<sup>1</sup>*Interdisciplinary Program of Electrical Engineering and Computer Science, National Central University, Taiwan*

<sup>2</sup>*Department of Computer Science and Information Engineering, National Central University, Taiwan*

<sup>3</sup>*IoX Center, National Taiwan University, Taiwan*

<sup>4</sup>*Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan*

<sup>5</sup>*Corresponding Author*

## Abstract

This paper presents our system for the BioASQ10b Phase B task. For ideal answers, we used the fine-tuned BioBERT model on the MNLI dataset to construct sentence embeddings and combined it with BERTScore to select sentences from the provided Snippets as answers. For the exact answers, we also used the BioBERT model and used the snippet scores generated from the ideal answer selection model to predict the exact answers for factoid and list questions. The exact answers of our fifth test batch ranked second place. In addition, the ideal answers we submitted achieved first place in the ROUGE score in all test batches from batch second to fifth.

## Keywords

Biomedical Question Answer, Pre-trained Language Model, Text Similarity

## 1. Introduction

Since 2013, BioASQ has held annual biomedical semantic indexing and question answering challenges. This year, BioASQ Task 10b Phase B (QA task) provides biomedical questions and some relevant snippets, and the participants have to generate either the exact answer or the ideal answer by using the snippets. BioASQ Task 10b PhaseB task provided 4,234 training questions, including the previous year's test set with gold annotations, plus 500 new test questions for evaluation, divided into five batches of 100 questions each. All questions and answers are constructed by a biomedical expert team from across Europe. The questions are categorized into four types: Yes/no, factoid, list, and summary. Three types of questions required exact answers: yes/no, factoid, and list. Participants need to submit the ideal answer for every question. In Task 10b, each participant was allowed to submit up to five results per batch.

Four examples of QA types for BioASQ Task 10b Phase B (QA task) are illustrated in figure 1. Each BioASQ QA instance includes a question and several relevant PubMed abstract snippets. As a result, we formulated the task as query-based multi-document extraction (for the exact


---

*CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy*

✉ denis049@gmail.com (H. T. ); phoenix000.taipei@gmail.com (Y. Z. ); joyhan@cc.ncu.edu.tw (J. H. ); thtsai@csie.ncu.edu.tw (R. T. T. )



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

**Figure 1:** The QA examples of the BioASQ Task 10b Phase B (QA task)

Yes/no	Q. Does metformin interfere thyroxine? • Exact Answer: No • Ideal Answer: <b>No</b> . There are not reported data indicating that metformin reduce with thyroxine absorption.
Factoid	Q. What is the mode of inheritance of Facioscapulohumeral muscular dystrophy (FSHD)? • Exact Answer: [autosomal dominant] • Ideal Answer: Facioscapulohumeral muscular dystrophy has an <b>autosomal dominant</b> inheritance pattern.
List	Q. Which are the different isoforms of the mammalian Notch receptor? • Exact Answer: [Notch-1, Notch-2, Notch-3, Notch-4] • Ideal Answer: Notch signaling is an evolutionarily conserved mechanism, used to regulate cell fate decisions. Four Notch receptors have been identified in man: <b>Notch-1, Notch-2, Notch-3</b> and <b>Notch-4</b> .
Summary	Q. What is clathrin? • Ideal Answer: Clathrin helps build small vesicles in order to safely transport molecules within and between cells.

answer) and summarization (for the ideal answer) tasks. Last year, we used the BioBERT model combined with linear regression to achieve the best result in generating ideal answers [1].

This year, we improved our previous BioASQ 9B system in two ways. First, we selected the most relevant segments for each question by replacing ROUGE-SU4 with BERTScore, and used different linear regression layers to improve the ideal answers. In addition, we used sequential learning BioBERT to generate the exact answer. Meanwhile, we combined KU-DMIS's method [2] with the snippet scores given by the ideal selection model to select the final answer. This method is applied to factoid and list problems.

The sections are organized as follows. Section 2 briefly reviews recent works on biomedical QA. The details of our methods are described in Section 3. Section 4 is our detailed experiment procedure. Section 5 describes our configurations submitted to the BioASQ 10b Phase B challenge and the results. Section 6 discusses and summarizes our system's performance in the BioASQ QA task.

## 2. Related Work

**Transfer Learning with Pre-trained Language Model:** Since the introduction of BERT [3], the pre-trained language model has achieved excellent results in various tasks in NLP. However, the pre-trained language model has reached a bottleneck, and it is hard to improve performance in specific domains. To solve this problem, domain-adapted models are gradually being developed. TANDA [4] is an effective approach for the question answering task - Answer Sentence Selection (AS2). They used transfer learning on a pre-trained language model. In the paper, they also built a new dataset - ASNQ to fine-tune the pre-trained language model.

They outperform the previous state-of-the-art to prove that transfer learning can produce great results.

For Biomedical QA task, there are also many domain-adapted models are proposed, such as BioBERT [5], PubMedBERT [6], and SciBERT [7], and so on. In this paper, we choose BioBERT to be our basic model to do transfer learning. Sequential transfer learning [2] is another approach developed by KU-DMIS. They applied this approach to biomedical QA and demonstrated that transferring the knowledge of MNLI and SQuAD to the BioASQ task can improve the performance of BioBERT. They also mentioned that the order of datasets in sequential transfer learning is important. After BioASQ 8b, they released fine-tuned models such as BioBERT-MNLI and BioBERT-MNLI-SQuAD.

In addition, QA tasks require a large amount of annotated corpus to train the model. This is a prerequisite for deep learning. In addition to BioASQ, many QA datasets annotated by biomedical experts have been published [8]. The WikiQA dataset is collected from Wikipedia, the question is selected from the question-like queries that more than five distinct users click, and the candidate answer is the sentence from the summary section of the associated Wikipedia page.

**Extractive Summarization:** Summarization tasks can be divided into extractive and abstractive summarization. Extractive summaries are derived by selecting sentences or concatenating the most important sentences in a document. This approach is more robust and more suitable for tasks with less training data. There are many extractive abstraction methods based on neural networks have been proposed. Some of these studies were based on the RNN family of models [9, 10]. However, in the last years, with the success of BERT, many new studies have switched to using pre-trained language models. Yang Liu first proposed the BERTSUM architecture, which is based on BERT with the addition of inter-sentence Transformer layers to obtain document-level features, and surpasses other neural network models [11]. Later, some scholars have used text matching to improve the effectiveness of the BERT model in extractive summarization tasks [12]. Based on the above studies, We tried to accomplish the ideal answer task as an extractive summarization task. In contrast to other summarization datasets, for the BioASQ dataset, we can focus on extracting the relationships between questions and candidate sentences without worrying about the relationships between the candidate sentences.

## 3. Method

### 3.1. Data

To generate the exact and ideal answer, we first use BioASQ 10b and BioASQ 9b datasets to split specific training and testing data. We use BioASQ 9b as our training set and the portion of BioASQ 10b that is not included in BioASQ 9b as our testing set. Furthermore, we employed a distinct way to make the data for each type of problem into exact answers.

For Yes/no problem, we concatenate the question and each snippet into a sentence pair. Moreover, each sentence pair has one [ANSWER] label that is the answer to this question. In other words, if there are N snippets in one question, we will get the N sentence pairs with an answer label finally.

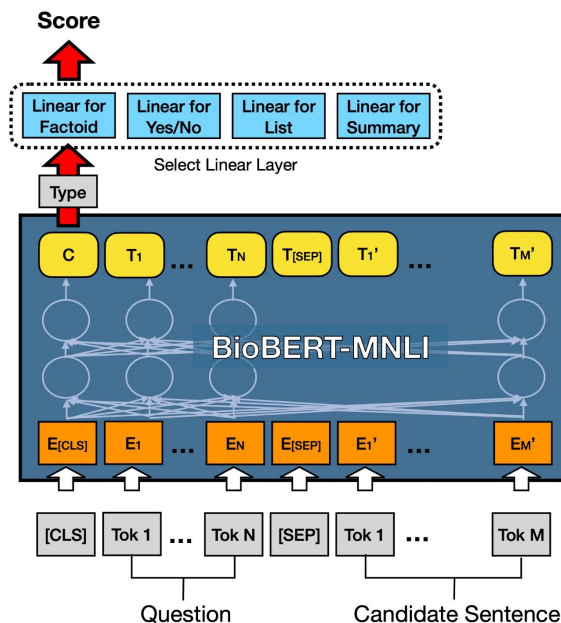
We make the question and each snippet a QA pair for the factoid and list problem. However, answer text is required in QA. We use Longest Common Subsequence (LCS) to find the answer text and index where are the answer started in each snippet. However, we found many invalid or too short answers found by the LCS algorithm, which might decrease the model performance and prediction. To solve this problem, we use **Levenshtein distance** to find the closer text to be the answer text so that all the answer in the training and testing set is meaningful.

### 3.2. Ideal Answer

For the ideal answer, we improved on our system of BioASQ 9B [1] and got better results. The improvements we made this year consist of two main parts: (1) We replaced ROUGE-SU4 with BERTScore and obtained better results. (2) Different linear regression layers were trained for different problem types, which allowed the problem type information to be integrated into the model.

Our model is shown in figure 2 . The goal is to select the sentence that best fits as an answer from multiple snippets for each question in the BioASQ question examples. The method consists of two parts. We split the input text into candidate sentences in the first step and scored each sentence according to its similarity to the golden answer. We fed the question and each candidate sentence into the pre-trained language model for fine-tuning in the second step. We apply an additional linear regression layer to the [CLS] token output of the BERT model to calculate the output score. The loss function is the mean square error between the output scores and the similar scores from the previous step. The fine-tuned model can predict scores based on the questions and candidate sentences. We selected the candidate sentences with the highest scores

**Figure 2:** Model for predicting the ideal answer



under each question as the system's ideal answer prediction result in the prediction stage.

The pre-trained language model we use is BioBERT-MNLI, which is a fine-tuned BioBERT checkpoint on the MultiNLI dataset [13], and it comes from Korea University [2].

**Text Similarity:** Many different methods can be used for the calculation of text similarity scores. The classical methods are ROUGE [14], BLEU [15], and other evaluation methods based on the difference between n-gram tokens. We used ROUGE-SU4 because it is one of the official methods for evaluating ideal answers in BioASQ. In addition, some studies have also developed evaluation methods based on pre-trained language models in recent years, the most famous and widely used of which is BERTScore [16]. BERTScore uses pre-trained contextual embeddings of BERT to match words in candidate and reference sentences by cosine similarity. It has been shown to correlate with human judgments on sentence-level evaluations. Therefore, we also used BERTScore based on the BioBERT-MNLI model this year.

**Question Type Specific Layer:** It is worth noting that there are four types of questions in the BioASQ 10b dataset, namely factoid, yes/no, list, and summary. We believe that there are differences in the ideal answers for different question types, e.g., the answer to a list question may contain multiple entities, and the answer to a factoid question is more likely to be something like a definition. Therefore, the model should have different parameter weights depending on the type of question. We can fine-tune the four BioBERTs with entirely different parameters, but this will significantly reduce the amount of training data. Therefore, we let the four different problem types share the vast majority of the model parameters and only have different parameters in the last layer of the linear regression part.

### 3.3. Exact Answer

To predict the exact answer, we used the transfer learning pre-trained BioBERT model and KU-DMIS's method [2] to get the possible answer list. The final answer list was combined with the snippet score given by the ideal answer selection model. We will introduce the system below respectively.

**For Yes/no problem,** we used the BioBERT-MNLI model, which is BioBERT fine-tuned on the MNLI dataset. MNLI (Multi-Type Natural Language Inference), published by New York University, is a text entailment task requiring determining whether a hypothesis holds given a premise (Premise) or whether the hypothesis is contradictory and neutral to the premise. The snippet score was applied to the answer list generated from the model. For each snippet in every question, if the probability of the snippet's answer is higher than 0.5, this snippet's weight is a positive weight "1"; otherwise, it is a negative weight "-1". The snippet score is then **multiplied** by its weight, yielding a new score for this snippet. We obtain the final answer to this question by adding all of the snippets' new scores together. The final answer is yes or no, depending on whether the summary is positive or negative.

**For factoid problem,** we used the BioBERT-SQuAD model, which is BioBERT fine-tuned on the SQuAD dataset. SQuAD (Stanford Question Answering Dataset) is a reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles. The answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. The snippet score was also applied to the generated answer list. In the answer list, we have the probability of each possible answer of each snippet.

**Table 1**

Results of Ablation Study of Ideal Answer Prediction on BioASQ9b Dataset

Method	ROUGE-SU4			BERTScore		
	P	R	F1	P	R	F1
ROUGE + BioBERT-MNLI	39.08	36.86	37.94	74.09	74.19	74.14
BERTScore + BioBERT-MNLI	40.12	36.57	38.26	75.45	74.59	75.02
BERTScore + BioBERT-MNLI + Type Layer	<b>41.04</b>	<b>36.99</b>	<b>38.91</b>	<b>76.05</b>	<b>74.85</b>	<b>75.45</b>

Each answer’s probability was **multiplied** by the snippet score and the resulting probability was used to re-order all snippets’ answers. We got all possible answers ranked by new probability and selected the top five answers to be the final answers for each factoid question.

**For list problem**, we also used the BioBERT-SQuAD model. The snippet score was also applied to the generated answer list. As same as the method for the factoid problem, the answer’s probability was **multiplied** by the snippet score, and the resulting probability was used to re-order all snippets’ answers. We got all possible answers ranked by new probability for each list question. However, it is not a good strategy to submit all possible answers. We set a threshold to select answers that new probability exceeds the threshold. Different thresholds are set for different models and snippet score versions.

## 4. Experiment

### 4.1. Ideal Answer

Table 1 shows the performance of our system on the BioASQ 9b dataset. The average scores were calculated from the system’s predictions and the golden answers. All the experiment data are the best results of the model under this method. We set the epoch number to 3, batch size to 20, and learning rate to  $2e-5$ .

To our surprise, the system with BERTScore outperformed the system with ROUGE on all evaluation criteria, even on the evaluations related to ROUGE. We believe that such experimental results imply that BERTScore is an overall better method for evaluating text similarity. In addition, the experimental results also show that training specific linear regression layers for different problem types can help improve system performance.

### 4.2. Exact Answer

For the following experimental procedure, we use the same testing set split by Training 10b and Training 9b to do the test to observe the experimental results.

**For Yes/no problem**, we experimented with different transfer learning models, including BioBERT, BioBERT-MNLI, BioBERT-SQuAD, and BioBERT-MNLI-SQuAD, and compared the differences between the models. As shown in Table 2, the performance of BioBERT-MNLI is substantially higher than the other models by more than 0.04 in both metrics. We consider that it is because yes/no problem is more similar to MNLI (text entailment task) than the others. Additionally, we evaluated the performances of the BioBERT-MNLI model with/without

**Table 2**

Result of Ablation Study for Yes/no problem

Model	Snippet Score	Acc	Ma F1
BioBERT	Not Applied	0.810	0.771
BioBERT-MNLI	<b>Not Applied</b>	<b>0.888</b>	<b>0.872</b>
	Applied	0.879	0.861
BioBERT-SQuAD	Not Applied	0.845	0.830
BioBERT-MNLI-SQuAD	Not Applied	0.810	0.767

**Table 3**

Result of Ablation Study for factoid problem

Model	Snippet Score	Approach	SAcc	LAcc	MRR
BioBERT	Applied	Multiply	0.342	0.563	0.420
BioBERT-MNLI-SQuAD	Not Applied		0.506	0.747	0.598
	Applied	Multiply	0.519	0.747	0.602
BioBERT-SQuAD	Not Applied		0.519	<b>0.785</b>	0.620
BioBERT-SQuAD	Applied	Multiply	<b>0.532</b>	0.778	<b>0.625</b>
		Plus	0.532	0.747	0.615

the usage of a snippet score. To our surprise, Table 2 shows that the performances of the BioBERT-MNLI on applied snippet score are decreased by roughly 0.01.

**For factoid problem**, we evaluated different transfer learning models, including BioBERT, BioBERT-SQuAD, and BioBERT-MNLI-SQuAD, as shown in Table 3. KU-DMIS [2] showed that BioBERT-MNLI-SQuAD had a higher performance than BioBERT-SQuAD. However, our experiment found that BioBERT-SQuAD outperformed BioBERT-MNLI-SQuAD in all measurements by more than 0.02. We also evaluated the performance of BioBERT-SQuAD with and without using the snippet score. There is an increase of 0.02 on SAcc, which shows that it predicts more accurate answers. Meanwhile, MRR has a small improvement of 0.005.

In addition, we conducted experiments on different snippet score calculations. We combined the snippet score and the answer’s probability by two different operations, "Multiply" and "Plus". In Table 3, using the "Multiply" approach outperformed using the "Plus" approach in all metrics.

**For list problem**, BioBERT, BioBERT-SQuAD, and BioBERT-MNLI-SQuAD are evaluated. In Table 4, we found that BioBERT-SQuAD with the "Multiply" approach outperformed BioBERT-MNLI-SQuAD on F-Measure, this result is also different from KU-DMIS [2]. We evaluated the performances of BioBERT-SQuAD with and without using the snippet score. There is an increase of 0.03 for the "Multiply" approach on the F-Measure, and Precision also has an increase of 0.06.

In addition, we also conducted experiments using different snippet score calculations. we observed that, unlike factoid, using "Plus" for the list problem might have a considerable negative

**Table 4**  
Result of Ablation Study for list problem

Model	Snippet Score	Approach	Prec.	Rec.	F-Measure
BioBERT	Applied	Multiply	0.522	0.639	0.514
BioBERT-MNLI-SQuAD	Not Applied		0.434	<b>0.754</b>	0.505
	Applied	Multiply	0.533	0.648	0.532
BioBERT-SQuAD	Not Applied		0.485	0.747	0.542
BioBERT-SQuAD	Applied	Plus	0.593	0.35	0.381
		Multiply	0.547	0.711	<b>0.572</b>
		Result of "Multiply" with Power of 1.5	0.630	0.498	0.493
		Result of "Multiply" with Power of 2	<b>0.705</b>	0.415	0.475

impact on performance. Therefore, we tried calculating the result of the "Multiply" approach in various powers. We found that the Precision increased after adding the power. However, F-Measure and Recall are decreased.

## 5. Submission

### 5.1. Ideal Answer

We participated in all but the first batch and obtained the highest score in the ROUGE evaluation (based on ROUGE-2 F1 and ROUGE-SU4 F1) from the second to fifth batches. In each batch, we submitted multiple configurations, most of which differed only in their hyperparameter settings. We simplified the description and reported the best results of each method in Table 5.

The results demonstrate the overall effectiveness of our system. The results of batch 4 show that BERTScore is a better measure of text similarity and positively affects our model training. However, Batch 5 is different from our experimental results on BioASQ 9b dataset. The model's performance became worse after adding the question type for Linear Layer selection. Because the gold answers for BioASQ 10b have not yet been published, we have not been able to analyze its possible causes.

### 5.2. Exact Answer

In the official submission, we did not submit other models' results. We only submitted the system that performed the best during the experiment.

In the fifth test batch, we used the newest snippet score version, which applied to factoid and list problems to generate an answer list, detailed as shown in Table 6. We also show the first place of each problem together in the table. Our factoid answer is ranked second in this batch, and the difference between it and first place (Ir\_sys3) is quite minimal, indicating that our solution is very effective in the factoid problem. In addition, we came in sixth place in the list problem, with UDEL-LAB providing the top five solutions. However, there is still a discrepancy



**Table 5**

Test batch results for the ideal answer

Method	Batch	Rank	ROUGE-2		ROUGE-SU4	
			Recall	F1	Recall	F1
ROUGE + BioBERT-MNLI	2	#1	47.28	42.62	48.23	41.68
ROUGE + BioBERT-MNLI	3	#1	41.93	37.61	42.36	36.89
ROUGE + BioBERT-MNLI	4	#4	40.40	40.26	40.64	39.41
BERTScore + BioBERT-MNLI	4	#1	41.79	42.29	41.57	41.65
BERTScore + BioBERT-MNLI	5	#1	42.70	40.20	42.08	39.16
BERTScore + BioBERT-MNLI + Type Layer	5	#2	39.62	37.07	39.28	36.26

**Table 6**

Test Batch fifth Result of Exact Answer

System name	Yes/no: Macro F1	factoid: MRR	list: mean F-Measure
NCU-IISR-AS-GIS-4	0.8893	0.4983 (#2)	0.5332 (#6)
lr_sys1	0.9282 (#1)	0.4195	0.5224
lr_sys3	0.8916	0.5098 (#1)	0.4620
UDEL-LAB4	0.8893	0.4017	0.6016 (#1)

between our results and those of UDEL-LAB, indicating that our approach to the list should be improved.

## 6. Discussion and Conclusion

In the BioASQ 10b phaseB task, we used the fine-tuned BioBERT model on the MNLI dataset to construct sentence embeddings and combined it with BERTScore to select sentences from the provided snippets as ideal answers. As for the exact answer, we combined the KU-DMIS approach with our ideal answer selection model to predict exact answers for factoid and list problems.

Our approach proved that replacing ROUGE-SU4 with BERTScore is a practical improvement for the ideal answer. This further validates BERTScore as a better method for text evaluation. At the same time, our proposed regression method for linear layer selection based on question types is worth further investigation. Meanwhile, the experiment results of our proposed regression method for linear layer selection based on problem type were inconsistent, but it deserves further study.

We tried to apply the snippet score to three types of problems for the exact answer. For the yes/no problem, our approach could not further improve the performance, even if it negatively impacts model performance. We found an effective approach to improve the model’s performance for the factoid problem, placing us in second place in the fifth test batch. Our approach has also resulted in a significant improvement in our performance on the test set for the list problem. We are, however, still a long way from first place in the fifth test batch. Overall, we have improved the performance of the sequential learning model for factoid and list questions.

However, our method is not optimal, and there is still potential for improvement in yes/no and list problems. We hope to try to come up with better ways to increase performance in the future.

## References

- [1] Y. Zhang, J.-C. Han, R. T.-H. Tsai<sup>123</sup>, Ncu-iisr/as-gis: Results of various pre-trained biomedical language models and linear regression model in bioasq task 9b phase b, in: CEUR Workshop Proceedings, 2021.
- [2] M. Jeong, M. Sung, G. Kim, D. Kim, W. Yoon, J. Yoo, J. Kang, Transferability of natural language inference to biomedical question answering, arXiv preprint arXiv:2007.00217 (2020).
- [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [4] S. Garg, T. Vu, A. Moschitti, Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7780–7788.
- [5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240.
- [6] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, *ACM Transactions on Computing for Healthcare (HEALTH)* 3 (2021) 1–23.
- [7] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
- [8] Y. Yang, W.-t. Yih, C. Meek, Wikiqa: A challenge dataset for open-domain question answering, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2013–2018.
- [9] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, T. Zhao, Neural document summarization by jointly learning to score and select sentences, arXiv preprint arXiv:1807.02305 (2018).
- [10] X. Zhang, M. Lapata, F. Wei, M. Zhou, Neural latent extractive document summarization, arXiv preprint arXiv:1808.07187 (2018).
- [11] Y. Liu, Fine-tune bert for extractive summarization, arXiv preprint arXiv:1903.10318 (2019).
- [12] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, X. Huang, Extractive summarization as text matching, arXiv preprint arXiv:2004.08795 (2020).
- [13] A. Williams, N. Nangia, S. Bowman, 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), Association for Computational Linguistics, 2018, pp. 1112–1122. URL: <http://aclweb.org/anthology/N18-1101>.
- [14] K. Ganesan, Rouge 2.0: Updated and improved measures for evaluation of summarization tasks, arXiv preprint arXiv:1803.01937 (2018).
- [15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of

machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

- [16] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).