Smoothing Answers for Vietnamese Machine Reading Comprehension Task

1st Do Pham Phuc Tinh
Universiti of Information Technology
VNU-HCMC
Ho Chi Minh City
20522020@gm.uit.edu.vn

3rd An Huynh Quoc Tran
Universiti of Information Technology
VNU-HCMC
Ho Chi Minh City
20520955@gm.uit.edu.vn

2nd My Ngoc Ha Nguyen
Universiti of Information Technology
VNU-HCMC
Ho Chi Minh City
20521623@gm.uit.edu.vn

4th Kiet Van Nguyen
Universiti of Information Technology
VNU-HCMC
Ho Chi Minh City
kietnv@uit.edu.vn

***Abstract*—Extraction machine reading comprehension tasks typically focus on generating answers using a single span that can be extracted from a given passage. However, in the real world, answers often consist of multiple spans at different positions. Providing only a single span as an answer can result in missing essential information, including irrelevant or incorrect details, and can lead to grammatical inaccuracies. Conversely, using multi-span answers can address these existing issues. While various datasets have been available for single-span answer tasks, datasets for multi-span answers are still limited. In this study, we constructed a comprehensive and rigorous open-domain multi-span reading comprehension dataset. After the construction phase, we obtained a dataset comprising 1,457 multi-span question-answer pairs. We experimented with various hyperparameter sets for BERT, and the highest achieved results were 43.85% for the test set, with 58.59% for ROUGE-L and 82.06% for BERTScore-F1. Additionally, we analyzed error cases to identify the underlying causes and provide insights for future improvements.**

***Index Terms*—Machine Reading Comprehension, Multi-span**

I. Introduction

Machine reading comprehension (MRC) is a common task in the field of natural language processing. The extraction-type task is extracting words or phrases from a question or passage to form an answer. The single-span answer format has been extensively studied in recent years, thanks to the availability of high-quality datasets. However, this output format has several drawbacks. The answer is limited to a single span within a significant passage, which may result in missing necessary information, containing redundant information, including irrelevant details, exhibiting ambiguity, and potentially having grammatical errors. There is still a significant gap between model-generated answers and natural human-like answers. The multi-span answer task is another form of answer extraction that addresses many of these drawbacks. The multi-span answer is a task where the answer consists of multiple spans from the input question and passage. This format allows for more comprehensive answers, avoiding ambiguity.

The multi-span task poses challenges and difficulties, especially in complex languages like Vietnamese. One of the main challenges is determining the quantity and positions of the text portions to be extracted. Identifying exact spans within the text is a challenging task that requires a deep understanding of language and content. Spans can overlap or interact in complex ways, increasing the complexity of the problem. Spans also exhibit diversity and heterogeneity, making it challenging to identify and extract the correct spans while ensuring consistency and diversity of the results. Labeling the spans is also a complex task. Determining and labeling the spans requires significant time and effort for each passage. The dataset must include an adequate variety of spans and usage scenarios while maintaining consistency in labeling and span formatting. Because of these challenges, very few datasets are available for this task, especially in Vietnamese.

Recognizing the above challenges, we have developed a dedicated multi-span dataset for the Vietnamese language. The dataset comprises 1457 multi-span question-answer pair based on 1112 passages from an open-domain dataset (using passages from UIT-ViQuAD [1]). We have provided a comprehensive process and proposed challenging reasoning forms to increase the difficulty of the dataset. Additionally, we implement the BERT model [2] and analyze the results based on the number of spans, reasoning types, question types, and the errors encountered during the experiments.

The remaining parts of the report include: section II presents the related works. Section III provides a detailed description of the dataset. Section IV introduces the baseline models. Section Vdiscusses the experimental design and results. Finally, section VI concludes the report and provides directions for future development.

II. Related Works

Automatic reading comprehension is widely studied, especially for single-span answer formats. Several datasets have been developed for this task and achieved high performance close to the human level in English, Chinese, and Vietnamese. Examples include SQuAD [3], UIT-ViQuAD [1], TriviaQA [4]. Several datasets such as MultiSpanQA [5], Quoref [6] are available for the multi-span answer task. These datasets contain both single-span and multi-span answers, with multi-span ratios ranging from 2% to 10%.

Various methods have been proposed to perform extractive MRC, including sequence tagging extraction and predicting start and end positions for each span. Sequence tagging extraction has been utilized in several works []. Predicting start and end positions for each span has also been explored using syntactic methods [7]. Recent language models such as XLM-R [8] on the UIT-ViQuAD 1.0 dataset [1] achieved 87.02% on the test set for single-span answers, while the BERT-base multilingual model in [2] achieved 86.8% F1 score.

On the other hand, our dataset is the first dataset for multi-span question answering in Vietnamese with the hope of solving this problem for Vietnamese. This dataset, obtained from ViQUAD, consists of passages and questions designed for a single-span task. To facilitate multi-span question answering, we ensured that the answer labels were grammatically correct and comprehensible by smoothing them during the data labeling. This also means that the new answer contains both the content of the question and the information extracted from the passage. It makes it difficult for modern models to detect the correct subject in question and spans containing the information asked in the given passage.

III. Dataset

A. Dataset Overview

Machine Reading Comprehension is one of the common tasks in the field of natural language processing (NLP), where each data point consists of a passage, a question, and an answer. The answer is extracted from a span (a segment) within the passage to answer the question. However, the answer is limited to a single span, sometimes failing to convey all the information or achieving high effectiveness when the answer requires synthesizing information from multiple positions in the passage. Therefore, creating a multi-span QA dataset will expand the capabilities of answer systems, allowing for handling more complex questions.

However, there has been no multi-span QA dataset for the Vietnamese language. This presents an interesting opportunity to develop a new dataset in this field. We aim to build and contribute to the "Multi-Span Question Answering in the Vietnamese Language" dataset to promote the development of natural language processing models for the Vietnamese language and expand their applications in other areas such as chatbots, information extraction, and text summarization in Vietnamese.

Our dataset consists of 1457 question-answer pairs. Each answer corresponds to the provided question of the passage. The domain of our dataset spans various subjects, including history, geography, science, etc.

B. Guidelines

Before proceeding with the annotation process, we engaged in discussions and created guidelines to facilitate better annotation and capture additional dataset attributes. The guidelines outlined principles for generating questions and answers. Below are the types of answers we have defined and will be implemented in the dataset:

- Multi-span answer with multiple positions in the passage: This type of answer can appear at multiple positions within the passage. The answer can be found in different text sections and is not limited to a single position.
- Multi-span answer with multiple positions between the question and the passage: This type of answer can appear at multiple positions within both the question and the passage. The answer may include spans from the question and the passage, and the combination creates a complete answer.
- Single-span answer at a single position: This type of answer appears at a specific position within the passage. The answer is found in a specific section of the text and does not have options or dispersion to other positions.

This paper focuses on multi-span answers. Thus, We encourage annotators to assign answers that fall into the first two types mentioned above. This is done to challenge models and create difficulty for the dataset.

By incorporating these types of answers, we aim to push the boundaries of the models and introduce complexity into the dataset. This approach challenges the models to accurately locate and identify answers that may appear in multiple positions or require synthesizing information from both the question and the passage.

C. Inter-annotator agreement

We measured inter-annotator agreement by randomly selecting 150 questions and passages and having multiple annotators assign answers to those questions and passages. The evaluation metrics used to measure agreement were BLEU1, ROUGE and BERTScore.In cases where the agreement was not sufficiently high, we engaged in discussions and updated the guidelines, creating new questions and answers. This iterative process continued until the inter-annotator agreement reached an acceptable level. Once achieved, we divided and assigned the remaining portion of the dataset. Table II presents the agreement score between

TABLE I: An Example for the Task.

INPUT:

Passage: Người gốc Mỹ Latinh hoặc Iberia là nhóm cư dân tại Texas có số dân lớn thứ hai sau người gốc Âu không có nguồn gốc Mỹ Latinh và Iberia. Có trên 8,5 triệu người tuyên bố rằng mình thuộc nhóm dân cư này, chiếm 36% dân cư Texas. Trong đó, 7,3 triệu người có nguồn gốc México, chiếm 30,7% cư dân. Có trên 104.000 người Puerto Rico và gần 38.000 người Cuba sinh sống trong bang. Có trên 1,1 triệu người (4,7% cư dân) có tổ tiên Mỹ Latinh hoặc Iberia khác nhau, như người Costa Rica, Venezuela, và Argentina. (*Latinx or Iberian-origin individuals in Texas are an ethnic group that ranks second in size, following non-Latinx individuals of European descent.. Over 8.5 million people identify themselves as part of this population, comprising 36% of Texas' population. Among them, 7.3 million have Mexican ancestry, making up 30.7% of the residents. There are over 104,000 individuals from Puerto Rico and nearly 38,000 from Cuba residing in the state. Additionally, over 1.1 million people (4.7% of the population) have diverse Latinx or Iberian ancestry, including individuals from countries like Costa Rica, Venezuela, and Argentina.*)

Question: Những người gốc Mỹ Latinh sống ở Texas những quốc gia nào? (*Which countries do individuals of Latinx origin living in Texas come from?*)

OUTPUT:

Answer: người có nguồn gốc México, Puerto Rico, Cuba, Costa Rica, Venezuela, và Argentina là nhóm cư dân gốc Mỹ Latinh sống ở Texas (*Mexican ancestry, Puerto Rico, Cuba, Costa Rica, Venezuela, and Argentina are an ethnic group of Latinx origin living in Texas*)

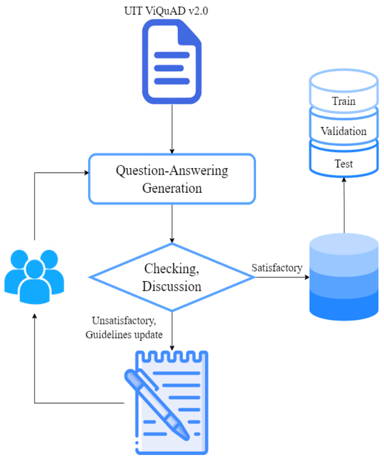

Fig. 1: Dataset Creation Process

annotators. For BLEU, we achieved a good agreement, according to Alon Lavie[1].

D. Dataset Statistics

After completing dataset construction, we obtained 1457 question-answer pairs, divided into the training, validation, and test sets in an 8:1:1 ratio (the training set has 1165 samples, the validation set - has 146 samples, and the test set - has 146 samples). Among them, 1457 passages were extracted from 1112 passages. This variation is because we intentionally reused some passages and questions while generating questions and answers to increase the model's ability to provide multiple answers. Figure 2 presents the number of spans on three datasets. The number of spans in each answer can vary widely and unevenly in the dataset, with the most concentration typically ranging from 2 to 5 spans across the dataset. The maximum is 18 spans. An answer's maximum number of spans is 18, which serves as a basis for setting parameters in the experimental setup.

TABLE II: Inter-annotator agreement.

	BLEU1	ROUGE-L	BERTScore-F1
Annotator 1	80.15	70.81	90.43
Annotator 2	70.23	65.51	86.72
Annotator 3	72.18	68.64	86.6
Average	74.19	68.32	87.92

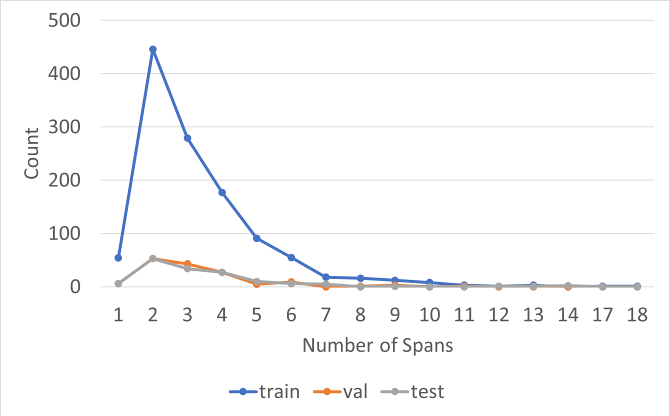

Fig. 2: Number of spans on Training, Validation, Test sets.

In addition to creating questions and answers, we determined the question types and reasoning types. In this dataset, there are 8 question types identified, namely: how, what, why, where, when, which, who, and others. Reasoning types are used to determine the direction of the inferred answer and consist of 5 types: word matching, paraphrasing, math, logic/causal relation, and coreference. The following table provides statistics on the number of question types and reasoning types in each dataset split.

From *TABLE* III, we can observe that the number of "What" and "Which" questions is higher compared to other question types. This could be due to the broad nature of these question types, which can be applied to various contexts and domains. "What" questions often seek specific information or descriptions, while "Which" questions typically require selecting from given options. Therefore, the higher count of "What" and "Which" questions reflects the diversity and popularity of these question types in the dataset.

[1] https://www.cs.cmu.edu/~alavie/Presentations/MT-Evaluation-MT-Summit-Tutorial-19Sep11.pdf

	Train	Validation	Test	Full
How	108	9	12	129
What	379	48	44	471
Why	90	10	14	114
Where	55	7	6	68
When	67	7	12	86
Which	256	38	32	326
Who	105	14	6	125
Other	105	13	20	138

TABLE III: Number of Question Types

	Train	Validation	Test	Full
Word matching	301	38	38	377
Paraphrasing	425	53	53	531
Math	69	9	8	86
Logic/causal relation	173	21	22	216
Coreference	197	25	25	247

TABLE IV: Number of Reasoning Types

As for *TABLE* IV, the "Math" reasoning type has the fewest occurrences compared to other reasoning types. This can be attributed to not all passages containing calculations-related content, resulting in a lower count for this reasoning type. Conversely, "Paraphrasing" and "Word matching" are the two most common reasoning types. We also analyzed the length in each column, including Passage, Question, Answer, and the combined Passage + Question column, in train, validation and test set to find a suitable max length for the model. The results obtained are presented in Table V.

TABLE V: Maximum Length of Each Passage, Question, Answer in the Dataset

	Train	Validation	Test
Passage	474	374	386
Question	40	43	32
Passage + Question	490	381	406
Answer	81	92	63

Based on the results in the *TABLE* V, the maximum length of the combined Passage + Question in the training set is 490 tokens. This finding suggests we can choose a suitable max len for the model as 512 tokens.

IV. Experiment

A. Baseline Model

In this paper, we use the BERT-base-multilingual-cased (mBERT) model to perform the experiment. BERT [2] is a pre-trained model on 104 languages published in 2019. BERT consists of transformer encoder layers (12 layers for BERT-base) and is pre-trained on a large corpus (about 3.3 billion words). mBERT makes a difference between "english" and "English".

For our experiments, we use the learning rates of 1E-5, 5E-5, 1E-4, 5E-4 and epochs of 3, 5. This choice is made because the maximum length of both the passage and question is 490 (according to Table V), and BERT allows a maximum of 512 tokens as input. Hence, we set the maximum input feature length to 512. Referring to Figure 2, the maximum number of spans in an answer is 18, we set it to 20.

B. Evaluation Metrics

The PyTorch framework was chosen because it's flexibility, dynamic computational graph, and active community make it an excellent choice for NLP model development. We run our model with different numbers of epochs, learning rate and batch size to find the best hyperparameters for our models.

To assess the effectiveness of models for machine reading comprehension (MRC) problems, researchers commonly compare the predicted answer to the ground truth answer. Various measures such as Exact Match, F1-score, BLEU, ROUGE, and BERTScore are commonly used in MRC research to facilitate this comparison. For instance, prior works [[5], [9], [10]] have used Exact Match and F1-score, while others [7] have employed BLEU1 and ROUGE-L. BERTScore has also been used as a primary measure in some studies [10].

This paper adopts our evaluation measures for BLEU 1, ROUGE-L, and BERTScore (F1-score). We chose these measures based on previous research highlighting their usefulness in evaluating MRC models. For example, works [[11], [7]] found that BLEU and ROUGE measures effectively assessed the efficiency and model capacity of MRC models for multi-span problems. BERTScore, on the other hand, is particularly useful for capturing the semantic similarities between words rather than just their order of appearance. Therefore, in this study, we employ a combination of these measures to evaluate our performance comprehensively.

V. Results

A. Performance on Baseline Model

We conducted experiments using the BERT-base-multilingual-cased model with many of the hyperparameter set. The best results are presented in Table VI and Figure 3. The overall model performance on the validation set achieved 45.22% with BLUE1. The ROUGE-L and BERTScore-F1 scores reached 59.13% and 82.12%, respectively. The model achieved 43.85% for the test set, with 58.59% for ROUGE-L and 82.06% for BERTScore-F1.

Additionally, we evaluated the number of spans of answer aspects in the validation and test sets. In both datasets, answers containing only one span performed significantly worse than answers with multiple spans. Specifically, in

TABLE VI: Model Performance on the Validation set and Test set.

	Validation			Test		
	BLEU1 (%)	ROUGE-L (%)	BERTScore - F1(%)	BLEU1 (%)	ROUGE-L (%)	BERTScore - F1(%)
All	**45.22**	**59.15**	82.12	43.85	58.59	82.06
Multi-span	46.31	60.22	82.39	44.57	59.23	82.32
Single-span	19.74	34.90	75.88	26.87	41.33	76.27

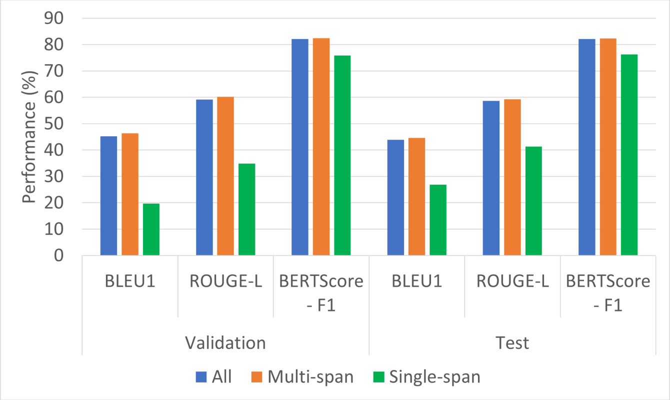

Fig. 3: Performance of Model on the Test set.

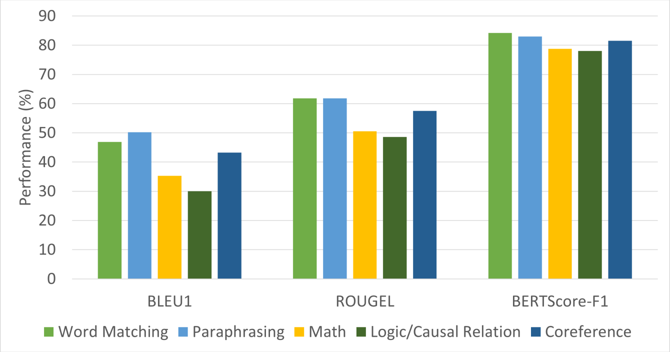

Fig. 4: Performance of Reasoning Types.

the validation set, single-span answers scored 16% lower in BLUE1, 29.52% lower in ROUGE-L, and 7.01% lower in BERTScore compared to multi-span answers. In the test set, multi-span answers outperformed single-span answers by 8.83% in BLUE1, 11.68% in ROUGE-L, and 2.23% in BERTScore.

The lower performance of single-span answers in both datasets can be attributed to the dominance of multi-span answers in the training set. As a result, the model has learned and tends to predict answers with multiple spans and more words. On the other hand, single-span answers have fewer words, leading to lower BLEU1 and ROUGE-L scores compared to multi-span answers in both the validation and test sets. Regarding BERTScore, the predicted answers contain more words, resulting in different contexts than single-span answers in the two datasets. Consequently, the performance of single-span answers is lower than that of multi-span answers.

B. Performance of Reasoning Types

As mentioned in the section, there is little difference in the distribution of reasoning and question types between the validation and test sets. Therefore, in this section and the following ones, we only analyze the results of different reasoning types on the test set. Figure 4 illustrates the detailed performance of the model on different types of reasons.

Figure 4 shows that Paraphrasing reasoning achieves the highest performance in terms of BLEU score, significantly outperforming other reasoning types such as Math, Logic/Causal Relation, and Coreference. On the other hand, for the ROUGE-L and BERTScore - F1 metrics, the Word Matching reasoning yields the highest results, surpassing Math, Logic/Causal Relation, and Coreference in performance. The superior performance of Word Matching and Paraphrasing in comparison to other reasoning types can be attributed to their higher occurrence rates in the training set, accounting for 25.84% and 36.48%, respectively (the distribution of reasoning types in the datasets present in the table IV).

Logic/Causal Relation reasoning performs poorly across all three metrics. The low performance of Logic/Causal Relation reasoning can be attributed to two reasons: its low occurrence rate in the dataset, accounting for only 11.48% in the training set, which makes it difficult for the model to learn effectively, and the dominance of Word Matching and Paraphrasing in the dataset, leading to a bias towards answers that contain more words from the question rather than being limited to the context paragraph. Due to these two reasons, the performance of Logic/Causal Relation reasoning is almost the lowest among the different reasoning types VII.

Despite having a lower occurrence rate (5.92% for Math compared to 11.48% for Logic/Causal Relation) in the training set, Math reasoning outperforms Logic/Causal Relation in terms of BLEU1, ROUGE-L, and BERTScore. This can be explained by the fact that Math reasoning typically requires shorter answers with fewer words, while the model tends to generate longer answers.

TABLE VII: The Average Length of Answer in Reasoning types.

	Train	Validation	Test
Word Matching	22.41	19.95	20.97
Paraphrasing	21.58	22.01	19.19
Math	16.72	15.33	15.75
Logic/Causal Relation	25.47	27.1	26.45
Coreference	22.45	24.48	18.2

C. Performance of Question Types

The WHY type of questions consistently yields the lowest performance and significantly underperforms compared to the other question types across all three metrics: BLEU1, ROUGE-L, and BERTScore-F1. This can be attributed to the fact that the WHY question format has a low occurrence rate in the training, validation, and test sets, which results in poor learning of the model for this particular type. Furthermore, WHY questions often correspond to Logic/Causal Relation reasoning, which also have a low occurrence rate in the dataset, and their answers tend to contain fewer words that are present in the question.

On the other hand, question types such as Word Matching and Paraphrasing have higher occurrence rates, which leads the model to generate answers that contain more words from the question. WHERE and WHEN question types have lower occurrence rates than WHY in the dataset but achieve higher results because the average number of words in their test set answers is lower than that of WHY questions and their answers tend to include more words from the question.

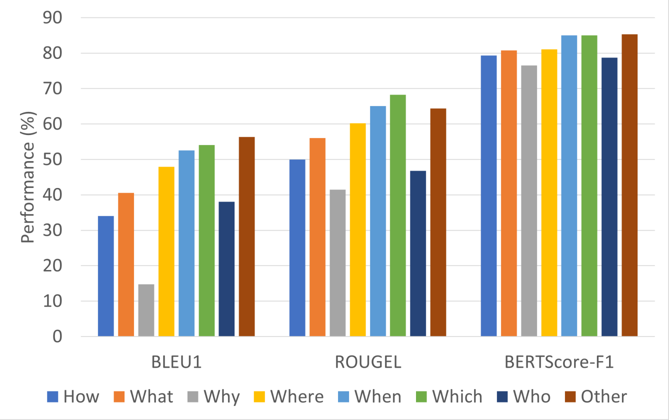

Fig. 5: Performance of Question Types.

D. Errors Analysis

We conducted a discussion and identified errors present in the prediction test set. Through this process, we found the following errors.

- Repetition errors in the question: these are answers that contain similar phrases that appear in the question.
- Errors containing question words: these are answers that include words used to differentiate between the question and the narrative (e.g., "what," "how many," "where," etc.).

These errors can be explained by the fact that the words in the answers were assigned by annotators and appeared in multiple questions, indicating the Word Matching reasoning type (where the answer includes the question concatenated with a span from the passage). The Paraphrasing reasoning type (which involves changing the sentence structure and grammar) leads to confusion in the model's learning process. Additionally, the "which" question form accounts for a significant proportion of the dataset, and as a result, the answers may repeat a portion of the question to indicate a specific choice.

VI. Conclusion and Future Work

In this study, we have initially constructed a multi-span answer dataset to facilitate smooth and comprehensive information for question answering in machine reading comprehension. Additionally, we experimented with the BERT-base-multilingual-cased model using this dataset, which served as the baseline model. The test set's highest results were 43.85% for BLEU1, 58.59% for ROUGE-L, and 82.06% for BERTScore-F1. We also analyzed the results based on the number of spans in the answers, reasoning types, question types, and predicted data to identify lingering errors and propose improvement methods.

Based on the experimental results and the analysis from the previous section, we have devised the following plans: 1) To revise and expand the dataset size to allow the model to learn more and improve performance. In this plan, we will decrease the proportion of Word Matching and Paraphrasing reasoning types while increasing other reasoning types, especially Math and Logic/Causal Relation, to enhance difficulty and create an intriguing dataset that the research community can tackle. 2) We intend to further experiment with text generation models such as various versions of GPT (GPT-2, GPT-3), Bloom and mBART to establish additional baseline models for this approach.

References

[1] K. Nguyen, V. Nguyen, A. Nguyen, and N. Nguyen, “A Vietnamese dataset for evaluating machine reading comprehension,” in *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2595–2605. [Online]. Available: https://aclanthology.org/2020.coling-main.233

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[4] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. [Online]. Available: https://aclanthology.org/P17-1147

[5] H. Li, M. Tomko, M. Vasardani, and T. Baldwin, "MultiSpanQA: A dataset for multi-span question answering," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1250–1260. [Online]. Available: https://aclanthology.org/2022.naacl-main.90

[6] P. Dasigi, N. F. Liu, A. Marasović, N. A. Smith, and M. Gardner, "Quoref: A reading comprehension dataset with questions requiring coreferential reasoning," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5925–5932. [Online]. Available: https://aclanthology.org/D19-1606

[7] J. Yang, Z. Zhang, and H. Zhao, "Multi-span style extraction for generative reading comprehension," *arXiv preprint arXiv:2009.07382*, 2020.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[9] E. Segal, A. Efrat, M. Shoham, A. Globerson, and J. Berant, "A simple and effective model for answering multi-span questions," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3074–3080. [Online]. Available: https://aclanthology.org/2020.emnlp-main.248

[10] D. Deutsch and D. Roth, "Benchmarking answer verification methods for question answering-based summarization evaluation metrics," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3759–3765. [Online]. Available: https://aclanthology.org/2022.findings-acl.296

[11] A. Chen, G. Stanovsky, S. Singh, and M. Gardner, "Evaluating question answering evaluation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 119–124. [Online]. Available: https://aclanthology.org/D19-5817