Article

Enhancing Question Answering Systems with Rephrasing Strategies: A Study on BERT Sensitivity and Refinement Techniques

Praveena Sunkara *

Posted Date: 19 May 2024

doi: 10.20944/preprints202405.1200.v1

Keywords: Answer; AI; BERT; Question; Paraphrases

Article

Enhancing Question Answering Systems with Rephrasing Strategies: A Study on BERT Sensitivity and Refinement Techniques

Praveena Sunkara, MD FACP [1]

Passion Health Primary Care, Dallas, USA; psunkara@passionhealthphysicians.com

Abstract: In this study, we aim to explore whether rephrasing enhances the performance of advanced question answering systems in terms of accuracy. Initially, we analyze the effectiveness of Bidirectional Encoder Representations from Transformers (BERT)—a widely used pre-trained language model—when presented with rephrased questions and documents containing distracting details. Our findings reveal that BERT exhibits excessive sensitivity to such instances, prompting us to investigate potential strategies for mitigating this sensitivity. To address the observed decline in performance, we propose a refinement approach for BERT, incorporating techniques such as data augmentation and multitask learning. Specifically, given a question and its context, we generate a series of paraphrases through back-translation. Alongside minimizing the loss between the predicted answer distributions for the original questions—we also minimize the supervised loss associated with the augmented questions. Furthermore, we introduce an auxiliary objective aimed at minimizing the unsupervised loss between the answer distributions of the original and augmented questions. Notably, this auxiliary loss is unsupervised, as it does not directly rely on labels corresponding to the augmented examples. To compute this unsupervised loss, we employ measures such as symmetric Kullback-Leibler divergence and the Jensen-Shannon distance, serving as regularization techniques for the model. The findings revealed that the supervised model exhibited superior performance compared to both the original and our proposed model. However, our proposed model demonstrated a nearly 2% enhancement in both Exact Match (EM) and F1 scores when tested on a paraphrased development set that we formulated. This suggests the potential effectiveness of our proposed approach in enhancing model accuracy and performance.

Keywords: answer; AI; BERT; question; paraphrases

1. Introduction

The pursuit of creating intelligent systems capable of performing tasks at a human level has captivated researchers for decades. Initially, intelligence was often defined by the mastery of complex activities, such as chess, which was considered a hallmark of intelligence. However, it soon became apparent that simple algorithms could excel at such tasks, suggesting that expertise in a specific area does not necessarily equate to general intelligence. This observation underscores the need for systems that can adapt and perform effectively across a variety of tasks and challenges. Tasks that were once used to assess human intelligence, such as numerical computation and memory tasks—have become trivial for computers, highlighting the evolving nature of intelligence metrics. Conversely, tasks involving natural language understanding and response, which are intuitive for humans, present significant challenges for machines. Symbolic manipulation alone is often insufficient for tackling these real-world tasks—leading to the development of biologically inspired learning algorithms that can better mimic human cognitive processes.

Natural Language Processing (NLP) plays a pivotal role in interpreting and representing human language using intelligent systems. Given the intricate nature of natural language—NLP is considered a critical aspect of Artificial Intelligence (AI) [1–5]. However, our study specifically focuses on Question Answering (QA), a subset of NLP dedicated to creating systems capable of extracting answers from given texts. In QA, the goal is to identify relevant information within a document that answers a given question. This serves as a metric to evaluate a machine's comprehension of textual data and its ability to understand and process natural language in a meaningful way. Through our research in QA, we aim to contribute to the advancement of NLP and the development of more intelligent and versatile AI systems [6–12]. While recent advancements in QA research have produced impressive models, questions arise regarding the true intelligence of these systems. Can they comprehend questions beyond superficial cues and adapt to various wordings or distractions in the text? Current QA systems often struggle with paraphrased questions or when tokens are replaced, highlighting their limitations. Therefore, our study aims to enhance QA systems' robustness by incorporating question as a regularization technique. In our exploration of Bidirectional Encoder Representations from Transformers (BERT)—we highlight its performance on the Stanford Question Answering Dataset (SQuAD) 1.1[1] challenge. However, we bring attention to a crucial aspect often overlooked—while BERT achieves remarkable results, its comprehension may not match its achievements. Our investigation uncovers BERT's sensitivity to variations in question phrasing and its vulnerability to adversarial input, emphasizing the necessity for abundant labeled examples during fine-tuning. Delving into innovative methods, we investigate back-translation for generating question paraphrases. Constructing four distinct development sets with 54 questions each, we utilize diverse pivot languages alongside manual paraphrasing. Our analysis reveals that leveraging two pivot languages, one from a different language family, optimizes paraphrase generation, enhancing diversity and quality. This offers valuable insights into bolstering BERT's resilience against textual variations. Moreover, we introduce an advanced fine-tuning framework for BERT, incorporating data augmentation and multitask learning as regularization techniques. Evaluating its efficacy against paraphrased questions and adversarial inputs—we compare our model with conventional approaches. While supervised data augmentation shows superior performance across development sets and paraphrased examples, our model surpasses the original BERT in assessments involving constructed paraphrases.

The study is as follows; similar papers are shown in the following section. The methods and materials are detailed in Section 3. The experimental analysis is carried out in Section 4, and in Section 5, we provide some conclusions and plans for future research.

2. Related Works

In this section, we offer a broad introduction to multitask learning—the distinctive framework provided by decaNLP [13], and an explanation of paraphrasing from the standpoint of NLP. Paraphrasing, often characterized as 'sameness of meaning' [14], can be elusive due to the varied degrees of 'sameness', leading to uncertainty in distinguishing it from other linguistic phenomena such as co-reference and inference. Within the realm of NLP, paraphrasing is primarily examined through a Machine Learning (ML) lens, with significant emphasis on identifying paraphrases for tasks like plagiarism detection [15]. Recent advancements in language models have demonstrated outstanding performance in this domain, particularly in tasks like Natural Language Inference (NLI) [16]. Evaluation often relies on corpora like the Microsoft Research Paraphrase Corpus (MSRPC) [17] and the Quora Question Pairs (QQP) dataset [18]. Moreover, diverse strategies have been devised to enhance NLP models' robustness to paraphrasing by generating variant inputs and ensembling responses [19]. For instance, [20] employ back-translation, while [13] explore an agent-based approach trained via Reinforcement Learning (RL) to optimize model performance. For decaNLP to facilitate zero-shot learning effectively, robustness to intricate paraphrasing, particularly at the semantic level, is imperative. To better comprehend paraphrase phenomena, it's essential to construct typologies based on insights from various fields, such as theoretical linguistics, discourse analysis, and computational linguistics. In computational linguistics, typologies are often crafted as lists of distinct paraphrase mechanisms, categorized into general classes for specific applications. [21] devised a typology tailored for NLP applications, which has been instrumental in tagging plagiarism corpora [22], including the influential MSRPC-A. Their hierarchical approach encompasses 20 paraphrase types, grouped based on the level of change (e.g., morphological, lexical, semantics).

[1] https://rajpurkar.github.io/SQuAD-explorer/

Multitask learning diverges from traditional single-task learning paradigms, aiming to emulate human learning by concurrently addressing multiple objectives. Methods for multitask learning encompass soft parameter sharing, hierarchical sharing, and hard parameter sharing schemes. Soft parameter sharing involves each task utilizing a subset of parameters, regularized to encourage similarity [23]. Hierarchical approaches leverage task relationships, mirroring the human cognitive hierarchy, with lower layers tackling simpler tasks and higher layers addressing more complex ones [24]. Hard parameter sharing involves sharing a portion of parameters among all tasks, typically with task-specific output layers [25].

3. Materials and Methods

In this section, we showcase QA systems that demonstrate remarkable performance on the task, with many achieving cutting-edge results on the SQuAD 1.1 challenge. Traditional QA models typically employ Recurrent Neural Networks (RNNs) supplemented with an attention mechanism. A prominent instance of such a system is the Bidirectional Attention Flow (BiDAF) model. The architecture of the BiDAF model comprises six layers that execute the following functions—character-level, word-level, and contextual embeddings. BiDAF surpassed all preceding QA models on the SQuAD 1.1 challenge. However, its recurrent structure incurs slowness, leading to prolonged training and inference times, hindering its practical utility. Therefore, an exemplar is the Question Answer Networks (QANet) architecture, exclusively grounded on convolutional and self-attention mechanisms [26]. Notably, QANet not only outpaced BiDAF in speed but also achieved the highest scores on the SQuAD 1.1 challenge. QANet's efficacy further improves with the incorporation of regularization techniques. Regularization encompasses a set of strategies aimed at refining the generalization of learning algorithms, primarily by curbing their tendency to overfit. We delve into specific regularization methods germane to QA, including multitask learning, data augmentation, and adversarial training. Multitask learning, pioneered by [27], posits that training models on multiple related tasks concurrently enhances their generalization abilities, leveraging shared parameters to benefit all tasks. This concept catalyzed the creation of evaluation benchmarks like GLUE [28] and superGLUE [29], fostering the assessment of general-purpose NLP models [30–33]. Multitask Question Answering Network (MQAN) [34] introduces a comprehensive approach of handling ten NLP tasks—eschewing the convention of extracting answers solely from the context. Conversely, Multi-Task SAN (MT-SAN) [35] challenges the notion that integrating multiple tasks uniformly into a single network invariably enhances performance. Instead, MT-SAN focuses on dynamically adjusting task weights within the loss function to optimize performance. Multi-Task Deep Neural Network (MT-DNN) [36] amalgamates the strengths of multitask learning and pre-trained language representations. The model shares lower layers across four tasks while incorporating task-specific layers, yielding multiple outputs corresponding to each task. MT-DNN achieved state-of-the-art results on eight GLUE tasks, showcasing a 2.2% absolute improvement over BERT Large. Supervised Data Augmentation (SDA) involves appending augmented examples to the training set without modifying the model architecture, a technique illustrated by QANet's back-translation approach [26]. Unsupervised Data Augmentation (UDA) [37] introduces a novel paradigm wherein model predictions between original and augmented examples are encouraged to align, effectively extending augmentation to unsupervised data. Recent research emphasizes the need to address challenges such as question paraphrasing and diverse adversarial inputs, ultimately leading to a better understanding of model behaviors and informing strategies for improving QA system performance.

3.1. Data Analysis

Datasets have been developed for solving the QA task. Our focus in this study is on the SQuAD dataset [38]—a widely used QA dataset where numerous models have been trained and assessed. We begin by briefly outlining some robust QA datasets containing factual inquiries about Wikipedia articles. Subsequently, we delve into the details of SQuAD and assess its quality. TriviaQA, introduced by [39], encompasses more than 650,000 QA-evidence tuples. This dataset is curated by gathering QA pairs from quizzes and then retrieving corresponding documents from the Web. TriviaQA poses numerous intricate questions necessitating reasoning across sentences and conducting comparisons. The Natural Questions dataset [40], developed by Google, is constructed using actual queries to the Google search engine. It comprises 307,373 training examples, 7,830 development examples, and 7,842 test examples, organized into quadruples. This dataset stands out for its authenticity, derived from real human queries, making it more applicable to real-world scenarios compared to synthetically constructed datasets. SQuAD 1.1, one of the most renowned QA datasets, comprises 107,785 sets of questions, passages, and answers, aimed at training and evaluating QA systems [38]. Passages are extracted from Wikipedia, with questions manually crafted based on these passages. Answers correspond to segments within the passages, challenging participants to locate them accurately. The dataset is divided into training (80%), validation (10%), and test (10%) sets, with only the training and validation sets publicly accessible, forming the basis of our experiments. The complexity of SQuAD 1.1 questions varies considerably, with most questions being answerable through simple key matching or token-specific searches, while only a minority demand higher-level reasoning across multiple sentences. SQuAD 2.0, an extension of SQuAD 1.1 by [41], introduces additional complexity by including questions that cannot be answered using the provided passages.

3.2. Model Analysis

In our study, we aim to refine the fine-tuning process of BERT Large for QA tasks by addressing specific challenges encountered during this phase. These challenges include the model's sensitivity to variations in question wording, its susceptibility to irrelevant sentences within passages, and the requirement for a substantial amount of annotated training data [42–47]. To overcome these obstacles, we propose a modified model architecture that integrates supervised data augmentation with a technique similar to UDA. Our approach leverages data augmentation and multitask learning as a form of regularization during the fine-tuning process. The proposed augmented model is illustrated in Fig. 1. Data augmentation is achieved through back-translation, a technique that generates additional training instances by paraphrasing the original question set. Formally, given an input tuple $(c, qorig, a)$, where '*c*' represents the context, '*qorig*' is the original question related to the context, and '*a*' is the labeled answer—we generate '*qpar*'—which serves as a paraphrased version of '*qorig*'. This process enhances the robustness of the model by exposing it to a variety of question formulations, thereby reducing its sensitivity to variations in question wording. Our integration of strategies into the fine-tuning process of BERT Large aims to bolster its performance on QA tasks while addressing its sensitivity to irrelevant information in passages and reducing its reliance on extensive annotated training data. This approach holds significant potential for enhancing the effectiveness and efficiency of QA systems in practical scenarios. Incorporating multitask learning involves an additional objective of minimizing the dissimilarity between predictions generated from two sets of questions—the original questions ('*qorig*') and their paraphrased versions ('*qpar*'). This task is supplementary to the primary objective of predicting an answer ('*a*'), thereby minimizing the supervised loss between the actual answer ('a_0') and the predicted answer ('a_1'). Unlike prior approaches that predominantly utilize Kullback-Leibler (KL) divergence for measuring the discrepancy between original predictions and their augmented counterparts—we propose the adoption of symmetric KL divergence or the Jenson-Shannon (JS) distance.

However, in the realm of QA systems, achieving an identical prediction for answer distributions of questions conveying the same meaning is an ideal goal. This scenario translates to a KL divergence of zero between the distributions, symbolized as $DKL(P||Q) = 0$, where P represents the distribution derived from the original query and Q from its paraphrased counterpart. However, our forthcoming analysis illustrates that BERT, a prevalent language model, often diverges from this ideal. Consequently, our proposed framework integrates an additional objective of minimizing this discrepancy. It's noteworthy that KL divergence lacks true symmetry since $DKL(P||Q)$ is generally not equivalent to $DKL(Q||P)$. Nonetheless, for equivalence between a question and its paraphrase, we aim to devise a symmetric loss function, ensuring the order of distribution inputs doesn't influence outcomes.

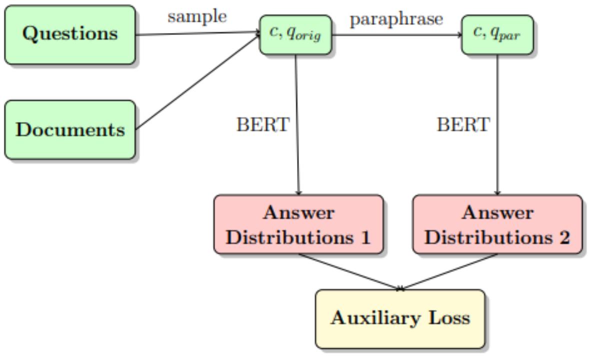

Figure 1. Augmented Model Architecture.

3.3. Evaluation Metrics

Evaluation metrics serve as benchmarks for assessing the performance of learning algorithms on a given dataset. Evaluation measures provide a method for assessing the performance of a learning method on a specific set of data. Among the various evaluation measures used for learning algorithms, accuracy stands out as particularly prominent. Accuracy is calculated by dividing the number of correct predictions by the total number of examples. Alongside accuracy, other important measures include precision, which gauges the proportion of accurately identified positive cases, and recall, which determines the proportion of true positive cases correctly identified. However, in the domain of QA, evaluation metrics play a crucial role. Two key metrics utilized in this context are Exact Match (EM) and F1 scores, both of which are the official evaluation measures for the SQuAD dataset. EM serves as a binary metric, indicating the count of predictions that exactly match the corresponding labels. F1 scores, on the other hand, offer a comprehensive evaluation by considering both precision and recall.

4. Experimental Analysis

4.1. Initial Hypothesis Testing

In this section, we investigate whether back-translation is effective for generating paraphrases and explore the influence of different intermediate languages on paraphrase quality. We propose that using multiple pivot languages, especially those from distinct language families, can result in paraphrases with varied expressions. To examine this, we create four sets of paraphrased questions based on the SQuAD 1.1:

- **Back-Translated Set 1**: We utilize Chinese (Mandarin) and Dutch as intermediate languages. The inclusion of Chinese, belonging to a different language family, is expected to introduce diverse structural variations to the paraphrased questions.
- **Back-Translated Set 2**: Vietnamese and Indonesian are employed as intermediate languages, all originating from different language families. We hypothesize that employing back-translation with multiple languages of contrasting origins might lead to significant alterations in question meanings, resulting in subpar paraphrases.
- **Back-Translated Set 3**: French is used as the sole intermediate language, aligning with prior works [26]. We speculate that employing a single pivot language closely related to the original questions' language may result in minimal changes, thus offering limited value in enriching the dataset with diverse examples.
- **Manually Paraphrased Set**: Questions are manually paraphrased to establish a human baseline for assessing the quality of other methods.

Human evaluation is conducted to compare paraphrase quality. We calculate the percentage of questions unaffected by paraphrasing (including cases of insignificant changes, like adding "the" to

doi:10.20944/preprints202405.1200.v1

a title) and the percentage of valid paraphrases. A paraphrase is deemed valid if it—i) retains the original question's meaning, ii) uses different words and/or structure, and iii) adheres to grammatical rules. Results of human evaluation are detailed in Table I. Notably, the first back-translated set yielded the highest proportion of acceptable paraphrases (over 90%, with 61.11% deemed valid and 29.63% unchanged). Conversely, the second method generated numerous incorrect paraphrases altering question meanings, while the last method failed to paraphrase most questions (75.93%), as anticipated. However, none of the methods produced notably diverse paraphrases akin to manually generated ones, suggesting potential benefits from more sophisticated paraphrasing techniques in the future.

Table I. ASSESSMENT OF PARAPHRASING METHODS BY HUMANS USING BACK-TRANSLATION.

Paraphrasing Method	Valid	Unchanged	Total	Valid/Total	Unchanged/Total
Back-Translated Set 1	33	16	54	61.11%	29.63%
Back-Translated Set 2	16	15	54	29.63%	27.78%
Back-Translated Set 3	10	41	54	18.52%	75.93%

Subsequent experiments (refer to Table II) deploy the first method, utilizing Chinese and Dutch as intermediate languages, for question paraphrase generation. The original BERT Large model is trained on the SQuAD 1.1 training dataset with default hyperparameters and evaluated on the original SQuAD 1.1 development set, achieving an EM score of 83.86 and F1 score of 90.50, consistent with prior findings. We hypothesize that BERT's performance may decrease when questions are paraphrased, as it might rely on simplistic patterns for locating answers. We investigate BERT's resilience to paraphrased questions through various metrics, including EM and F1 scores, KL divergence, and Wassertstein distance. The Wassertstein distance, a measure of dissimilarity between distributions, is computed to compare predicted answer distributions between the original and augmented sets. Additionally, BERT's confidence is assessed by computing the mean probabilities of start and end locations in predictions. We delve into the resilience of BERT to paraphrases and adversarial examples, offering deeper insights than previous studies.

Table II. EXAMINING 54 ORIGINAL AND 54 PARAPHRASED QUESTIONS, THE BERT RESULTS WERE EVALUATED.

Development Data	EM	F1	Mean KL	Mean Wasserstein	Confidence
Original Set	90.74	95.97	-	-	81.51%
Back-Translated Set 1	83.33	90.34	6.05	2.46 x 10^-4	99.88%
Manually Paraphrased Set	83.33	88.02	8.31	2.56 x 10^-4	82.94%

Similarly, in this study, we investigate the effectiveness of the unaltered BERT Large model when exposed to adversarial stimuli. Specifically, we employ an adversarial dataset comprising distracting sentences inserted into the input passages of the SQuAD dataset, as delineated by [48]. As depicted in Table III, we contrast the performance of the original BERT model when assessed on the adversarial test set against its performance on the unmodified questions. Notably, a substantial decline in performance is evident, attributable to the presence of distracting sentences. We postulate that the introduction of these additional sentences perplexes BERT due to their semantic and syntactic similarity to the questions, prompting BERT to identify simplistic patterns to locate tokens resembling answers.

Table III. BERT LARGE'S PERFORMANCE ON THE ORIGINAL AND MATCHING ADVERSARIAL QUESTIONS.

Trained on Full Training Set			Trained on 10% of Training Set	
Development Data	EM	F1	EM	F1
Original Set	83.00	89.61	75.80	83.54
Adversarial Set	51.10	56.27	39.80	45.14

4.2. Augmented Model Experiments

This section details the experiments performed using the proposed model architecture alongside data augmentation to mitigate the observed performance decline when assessing the model with paraphrased questions and adversarial passages. Initially, we set the default hyperparameters, which included batch sizes of 24 examples, 2 epochs (equating to 7299 training steps), and a maximum sequence length of 384 tokens, following guidance from the official BERT Tensorflow GitHub page. Subsequent adjustments were made to hyperparameters, such as increasing the maximum sequence length to 386 and reducing the batch size to 16, in tandem with varying epochs due to the augmented training data. The rationale behind these changes was elaborated upon when discussing semi-supervised data augmentation experiments as shown in Table IV.

Table IV. RESULTS OF THREE EPOCHS OF DATA AUGMENTATION USING 10% OF THE TRAINING DATA AND NEW HYPERPARAMETERS FOR THE SEMI-SUPERVISED MODEL.

Training Data	Training Steps	Loss	α	EM	F1	Confidence
8560 (10%)	1605	KL	0.2	70.18	80.16	26.12%
8560 (10%)	1605	KL	0.5	72.21	81.53	40.59%
8560 (10%)	1605	KL	0.8	72.93	82.09	61.32%
8560 (10%)	1605	Symmetric KL	0.2	70.06	79.95	29.53%
8560 (10%)	1605	Symmetric KL	0.5	71.80	81.15	43.88%
8560 (10%)	1605	Symmetric KL	0.8	72.03	81.43	58.97%
8560 (10%)	1605	JS	0.2	71.64	81.45	37.84%
8560 (10%)	1605	JS	0.5	72.53	82.08	54.51%
8560 (10%)	1605	JS	0.8	72.71	81.99	69.18%

The first series of experiments established a baseline using default hyperparameters, comparing it to a model trained to minimize the loss between original and augmented examples in a fully supervised manner as shown in Table V. Symmetric KL and JS distances between model predictions were computed to gauge differences between models with supervised data augmentation and the original model. All obtained EM and F1 scores fell within the range reported in [49], with slight increases observed in confidence and symmetric KL and JS distances as training steps increased. The similarity in performances between BERT with and without data augmentation suggested a lack of diversity in augmented examples, attributed to the chosen back-translation approach. To further investigate the impact of augmented answers, experiments were repeated with reduced training data, revealing noticeable improvements with supervised data augmentation as shown in Table VI.

TABLE V. OPERATION OF THE FULLY SUPERVISED MODEL FOR DIFFERENT NUMBERS OF TRAINING STEPS USING DATA AUGMENTATION AND DEFAULT HYPERPARAMETERS

Model	Training Steps	EM	F1	Confidence	KL	JS
Original BERT Large	7299	83.86	90.50	82.13%	-	-
Original BERT Large	10949	83.11	90.45	86.17%	-	-
Original BERT Large	14599	82.24	89.88	90.6%	-	-
BERT Large + DA	7299	83.88	90.42	79.18%	0.24	0.17
BERT Large + DA	10949	83.96	90.79	84.89%	0.40	0.23
BERT Large + DA	14599	82.91	90.28	89.04%	0.55	0.29

Table VI. OF THE TRAINING DATA AND DEFAULT HYPERPARAMETERS FOR VARYING NUMBERS OF TRAINING STEPS.

Model	Training Steps	EM	F1	Confidence	KL	JS
Original BERT Large	730	72.96	82.51	79.42%	-	-
Original BERT Large	1095	73.86	83.26	84.95%	-	-
Original BERT Large	1460	73.91	83.16	87.56%	-	-
BERT Large + DA	730	72.19	81.87	75.52%	0.50	0.25
BERT Large + DA	1095	78.81	86.83	95.53%	0.91	0.33
BERT Large + DA	1460	73.46	82.76	86.09%	0.84	0.28

In the subsequent experiments, our model underwent refinement to minimize losses both on original examples and labels while also addressing the unsupervised loss between predictions on original and augmented examples as shown in Table VII. However, challenges emerged, particularly concerning discrepancies in sequence lengths, which demanded meticulous adjustments to maintain consistency between original and paraphrased examples. After iterative adjustments, we settled on a maximum sequence length of 386, striking a balance between computational efficiency and model performance. Additionally, due to resource constraints, we utilized a training batch size of 16 for further experiments. The integration of both fully supervised and unsupervised losses yielded promising results, outperforming the original BERT model, especially in scenarios with limited training data. The amalgamation of various experimental approaches provided valuable insights into model performance and illuminated avenues for improvement.

TABLE VII. FINDINGS FOR THE COMBINED MODEL WITH THE SUPERVISED AND UNSUPERVISED LOSSES IN SQUAD 1.1.

Training Data	Training Steps	Loss	α	EM	F1	Confidence
8560 (10%)	1605	KL	0.2	70.00	80.03	28.54%
8560 (10%)	1605	KL	0.5	71.03	80.89	44.15%
8560 (10%)	1605	KL	0.8	71.81	81.12	65.50%
8560 (10%)	1605	Symmetric KL	0.2	69.83	79.88	28.05%
8560 (10%)	1605	Symmetric KL	0.5	70.63	80.42	42.56%
8560 (10%)	1605	Symmetric KL	0.8	71.84	81.27	66.68%
8560 (10%)	1605	JS	0.2	70.62	80.36	33.79%
8560 (10%)	1605	JS	0.5	71.51	81.18	55.61%
8560 (10%)	1605	JS	0.8	72.69	81.81	64.45%

5. Conclusion and Future Works

In this study, we embarked on a series of investigations aimed at addressing various research questions. Firstly, we examined the resilience of BERT Large when fine-tuned for reading comprehension tasks, particularly its ability to handle variations in questions and adversarial instances. While BERT consistently displayed high accuracy on the SQuAD 1.1 dataset, it struggled with altered inputs, revealing limitations in its reasoning and inference capabilities. Additionally, we explored back-translation methods for generating paraphrases, finding that approaches utilizing a single pivot language often failed to produce significant alterations in the input text. Conversely, employing two pivot languages, especially from distinct language families, resulted in more valid paraphrases, though maintaining semantic equivalence remained a challenge. To enhance BERT's resilience to question variations, we introduced a regularization technique involving data augmentation and multitask learning. However, contrary to our expectations, this approach did not outperform supervised data augmentation methods, indicating potential overfitting to the new language structures introduced by paraphrasing. Moving forward, several enhancements are proposed for future endeavors. Firstly, dividing the development set into validation and test sets would facilitate more robust evaluation metrics. Additionally, employing more sophisticated paraphrase generation methods and generating multiple augmented examples could further diversify the dataset. Furthermore, adjusting the auxiliary loss to a weighted sum of pairwise losses and incorporating weight annealing techniques could optimize the regularization process.

References

1. V. K. Kanaparthi, "Navigating Uncertainty: Enhancing Markowitz Asset Allocation Strategies through Out-of-Sample Analysis," Dec. 2023, doi: 10.20944/PREPRINTS202312.0427.V1.
2. V. K. Kanaparthi, "Examining the Plausible Applications of Artificial Intelligence & Machine Learning in Accounts Payable Improvement," *FinTech*, vol. 2, no. 3, pp. 461–474, Jul. 2023, doi: 10.3390/fintech2030026.
3. V. Kanaparthi, "Transformational application of Artificial Intelligence and Machine learning in Financial Technologies and Financial services: A bibliometric review," Jan. 2024, doi: 10.1016/j.jbusres.2020.10.012.
4. V. Kanaparthi, "AI-based Personalization and Trust in Digital Finance," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15700v1
5. V. Kanaparthi, "Exploring the Impact of Blockchain, AI, and ML on Financial Accounting Efficiency and Transformation," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15715v1
6. P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility," Feb. 2024, Accessed: Mar. 21, 2024. [Online]. Available: https://arxiv.org/abs/2402.16142v1
7. G. S. Kashyap, A. Siddiqui, R. Siddiqui, K. Malik, S. Wazir, and A. E. I. Brownlee, "Prediction of Suicidal Risk Using Machine Learning Models." Dec. 25, 2021. Accessed: Feb. 04, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4709789
8. N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, "An analysis of the robustness of UAV agriculture field coverage using multi-agent reinforcement learning," *International Journal of Information Technology (Singapore)*, vol. 15, no. 4, pp. 2317–2327, May 2023, doi: 10.1007/s41870-023-01264-0.
9. S. Wazir, G. S. Kashyap, and P. Saxena, "MLOps: A Review," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.10908v1
10. S. Naz and G. S. Kashyap, "Enhancing the predictive capability of a mathematical model for pseudomonas aeruginosa through artificial neural networks," *International Journal of Information Technology 2024*, pp. 1–10, Feb. 2024, doi: 10.1007/S41870-023-01721-W.
11. G. S. Kashyap, A. E. I. Brownlee, O. C. Phukan, K. Malik, and S. Wazir, "Roulette-Wheel Selection-Based PSO Algorithm for Solving the Vehicle Routing Problem with Time Windows," Jun. 2023, Accessed: Jul. 04, 2023. [Online]. Available: https://arxiv.org/abs/2306.02308v1
12. G. S. Kashyap *et al.*, "Revolutionizing Agriculture: A Comprehensive Review of Artificial Intelligence Techniques in Farming," Feb. 2024, doi: 10.21203/RS.3.RS-3984385/V1.
13. C. Buck *et al.*, "Ask the right questions: Active question reformulation with reinforcement learning," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, May 2018. Accessed: May 14, 2024. [Online]. Available: https://arxiv.org/abs/1705.07830v3
14. M. Vila, M. A. Martí, and H. Rodríguez, "Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology," *Open Journal of Modern Linguistics*, vol. 04, no. 01, pp. 205–218, Feb. 2014, doi: 10.4236/ojml.2014.41016.
15. M. I. and W. H., "Exploring the Recent Trends of Paraphrase Detection," *International Journal of Computer Applications*, vol. 182, no. 46, pp. 1–5, 2019, doi: 10.5120/ijca2019918317.
16. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, Jun. 2019, vol. 32. Accessed: May 14, 2024. [Online]. Available: https://arxiv.org/abs/1906.08237v2
17. W. B. Dolan and C. Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases," *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pp. 9–16, 2005, Accessed: May 14, 2024. [Online]. Available: http://www.cogsci.princeton.edu/~wn/
18. Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "Duplicate Questions Pair Detection Using Siamese MaLSTM," *IEEE Access*, vol. 8, pp. 21932–21942, 2020, doi: 10.1109/ACCESS.2020.2969041.
19. M. T. Ribeiro, S. Singh, and C. Guestrin, "Semantically equivalent adversarial rules for debugging NLP models," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, vol. 1, pp. 856–865. doi: 10.18653/v1/p18-1079.
20. L. Dong, J. Mallinson, S. Reddy, and M. Lapata, "Learning to paraphrase for question answering," in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, Aug. 2017, pp. 875–886. doi: 10.18653/v1/d17-1091.
21. H. R. M Vila, MA Martí, "Paraphrase concept and typology. A linguistically based and computationally oriented approach," pp. 93–100, 2013, Accessed: May 15, 2024. [Online]. Available: http://www.redalyc.org/articulo.oa?id=515751746010

22. A. Barrón-Cedeño, M. Vila, M. Antònia Martí, and P. Rosso, "Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection," *Computational Linguistics*, vol. 39, no. 4, pp. 917–947, Dec. 2013, doi: 10.1162/COLI_a_00153.
23. L. Duong, T. Cohn, S. Bird, and P. Cook, "Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser," in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 2015, vol. 2, pp. 845–850. doi: 10.3115/v1/p15-2139.
24. V. Sanh, T. Wolf, and S. Ruder, "A hierarchical multi-task approach for learning embeddings from semantic tasks," in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Nov. 2019, pp. 6949–6956. doi: 10.1609/aaai.v33i01.33016949.
25. R. A. Caruana, "Multitask Learning: A Knowledge-Based Source of Inductive Bias," in *Proceedings of the 10th International Conference on Machine Learning, ICML 1993*, 1993, pp. 41–48. doi: 10.1016/b978-1-55860-307-3.50012-5.
26. A. W. Yu *et al.*, "QaNet: Combining local convolution with global self-attention for reading comprehension," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, Apr. 2018. Accessed: May 15, 2024. [Online]. Available: https://arxiv.org/abs/1804.09541v1
27. R. Caruana, "Multitask connectionist learning," in *Proc. 1993 Connectionist Models Summer School*, Mar. 1993, pp. 372–379. doi: 10.4324/9781315806433-54.
28. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, Apr. 2018, pp. 353–355. doi: 10.18653/v1/w18-5446.
29. A. Wang *et al.*, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.
30. V. Kanaparthi, "Evaluating Financial Risk in the Transition from EONIA to ESTER: A TimeGAN Approach with Enhanced VaR Estimations," Jan. 2024, doi: 10.21203/RS.3.RS-3906541/V1.
31. V. Kanaparthi, "Credit Risk Prediction using Ensemble Machine Learning Algorithms," in *6th International Conference on Inventive Computation Technologies, ICICT 2023 - Proceedings*, 2023, pp. 41–47. doi: 10.1109/ICICT57646.2023.10134486.
32. V. Kanaparthi, "Examining Natural Language Processing Techniques in the Education and Healthcare Fields," *International Journal of Engineering and Advanced Technology*, vol. 12, no. 2, pp. 8–18, Dec. 2022, doi: 10.35940/ijeat.b3861.1212222.
33. V. Kanaparthi, "Robustness Evaluation of LSTM-based Deep Learning Models for Bitcoin Price Prediction in the Presence of Random Disturbances," Jan. 2024, doi: 10.21203/RS.3.RS-3906529/V1.
34. B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The Natural Language Decathlon: Multitask Learning as Question Answering," Jun. 2018, Accessed: May 15, 2024. [Online]. Available: https://arxiv.org/abs/1806.08730v1
35. Y. Xu, X. Liu, Y. Shen, J. Liu, and J. Gao, "Multi-task learning with sample re-weighting for machine reading comprehension," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Sep. 2019, vol. 1, pp. 2644–2655. doi: 10.18653/v1/n19-1271.
36. X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, Jan. 2020, pp. 4487–4496. doi: 10.18653/v1/p19-1441.
37. Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V Le, "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems*, 2020, vol. 2020-Decem, pp. 6256–6268. Accessed: May 15, 2024. [Online]. Available: https://github.com/google-research/uda.
38. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuad: 100,000+ questions for machine comprehension of text," in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, Jun. 2016, pp. 2383–2392. doi: 10.18653/v1/d16-1264.
39. M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, May 2017, vol. 1, pp. 1601–1611. doi: 10.18653/v1/P17-1147.
40. T. Kwiatkowski *et al.*, "Natural Questions: A Benchmark for Question Answering Research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, Aug. 2019, doi: 10.1162/tacl_a_00276.
41. P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Jun. 2018, vol. 2, pp. 784–789. doi: 10.18653/v1/p18-2124.

42. H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model," in *Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications*, CRC Press, 2023, pp. 93–99. doi: 10.1201/9781003190301-6.
43. G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing," *Multimedia Tools and Applications*, vol. 81, no. 25, pp. 36685–36698, Oct. 2022, doi: 10.1007/s11042-021-11558-9.
44. S. Wazir, G. S. Kashyap, K. Malik, and A. E. I. Brownlee, "Predicting the Infection Level of COVID-19 Virus Using Normal Distribution-Based Approximation Model and PSO," Springer, Cham, 2023, pp. 75–91. doi: 10.1007/978-3-031-33183-1_5.
45. G. S. Kashyap, D. Mahajan, O. C. Phukan, A. Kumar, A. E. I. Brownlee, and J. Gao, "From Simulations to Reality: Enhancing Multi-Robot Exploration for Urban Search and Rescue," Nov. 2023, Accessed: Dec. 03, 2023. [Online]. Available: https://arxiv.org/abs/2311.16958v1
46. M. Kanojia, P. Kamani, G. S. Kashyap, S. Naz, S. Wazir, and A. Chauhan, "Alternative Agriculture Land-Use Transformation Pathways by Partial-Equilibrium Agricultural Sector Model: A Mathematical Approach," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.11632v1
47. G. S. Kashyap *et al.*, "Detection of a facemask in real-time using deep learning methods: Prevention of Covid 19," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15675v1
48. R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, Jul. 2017, pp. 2021–2031. doi: 10.18653/v1/d17-1215.
49. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Oct. 2019, vol. 1, pp. 4171–4186. Accessed: May 03, 2023. [Online]. Available: https://arxiv.org/abs/1810.04805v2

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.