Analysis of QA System Behavior against Context and Question Changes

Rachid Karra
Department of Computer Science, Mohammed V University, Morocco
rachid.karra@est.um5.ac.ma

Abdelali Lasfar
Department of Computer Science, Mohammed V University, Morocco
abdelali.lasfar@est.um5.ac.ma

Abstract: *Data quality has gained increasing attention across various research domains, including pattern recognition, image processing, and Natural Language Processing (NLP). The goal of this paper is to explore the impact of data quality (both questions and context) on Question-Answering (QA) system performance. We introduced an approach to enhance the results of the QA system through context simplification. The strength of our methodology resides in the utilization of human-scale NLP models. This approach promotes the utilization of multiple specialized models within the workflow to enhance the QA system's outcomes, rather than relying solely on resource-intensive Large Language Model (LLM). We demonstrated that this method improves the correct response rate of the QA system without modification or additional training of the model. In addition, we conducted a cross-disciplinary study involving NLP and linguistics. We analyzed QA system results to showcase their correlation with readability and text complexity linguistic metrics using Coh-Metrix. Lastly, we explore the robustness of Bidirectional Encoder Representations from Transformers (BERT) and Reliable National Entrance Test (R-NET) models when confronted with noisy questions.*

Keywords: *Adversarial attacks, BERT, data quality, question answering, simplification*.

Received August 5, 2023; accepted January 3, 2024
https://doi.org/10.34028/iajit/21/2/2

1. Introduction

E-learning is increasingly being used in education, and its importance has increased even more during COVID-19. Image processing and Natural language Processing (NLP) have experienced significant technological progress due to machine and Deep Learning (DL), such as sentiment analysis through facial recognition to label positive and negative perceptions or angry and happy expressions.

Service companies, like food delivery, utilize chatbots to coordinate their offerings [29]. Some services and industries employ dialog interfaces based on Artificial Intelligence (AI) to provide after-sales service and customer interactions. Within healthcare field, these interfaces can offer personalized advice and diagnoses based on the patient's symptoms through Question-Answering (QA) interface or chatbots [33]. The objective is to develop more personalized and effective models based on their users' data. Employing a chatbot to interact with students and collect data will serve as a source to improve the proposed teaching.

Machine Learning (ML) also plays an important role in monitoring assessments and reviews. It removes bias and simplifies the assessment process [4]. Additionally, it gives learners the opportunity to practice through collaborative learning, especially with the intervention of a third party like virtual agents and intelligent moderators [18]. Introducing DL is beneficial for education. It gives personalized teaching through personal tutoring and learning activities adapted to each learner. Students can monitor their learning degree and control their progress regularly [14]. On the other hand, it allows students' reactions and interactions in real-time, which offers better learning opportunities. Students are continuously supervised by answering their questions and comments.

The primary goal of the NLP model is to allow computers to understand human language on both semantic and syntactic levels. Such deep linguistic knowledge is difficult for computers to attain. Open-domain QA system uses contexts as a resource to answer any type of question. To get the right answer, we first use a tri-gram Term Frequency-Inverse Document Frequency (TF-IDF) model to retrieve the relevant document. Then, we utilize a large language model to generate the correct response from paragraph [7]. There are many large-scale QA datasets such as Stanford Question Answering Dataset (SQuAD), MicroSoft MAchine Reading COmprehension (MS MARCO), and Wikipedia open-domain Question Answering (WikiQA) consisting of questions and context for answer extraction [5, 26]. SQuAD2.0 added 50,000 questions to the 107,785 original question-answer pairs, where each answer is a text span. It labels the right answer as Ground Truth Answers (GTA) and for multiple right answers as well-formed answers table. These datasets are the backbone of large DL models for training and testing.

Previous research endeavors have focused on creating new Large Language Models (LLMs) using unlabeled datasets containing billions of parameters. [7, 12]. Meanwhile, specific research endeavors have prioritized assessing the quality of datasets [23, 27]. Advocates of text simplification contend that the syntactic and lexical alterations employed in this process enhance the readability of the text, thereby enabling readers to better comprehend and engage with it [10].

Data quality dimensions (precision, relevance, and accessibility [27] have been delineated within the literature. These attributes are applied to specific data interacting with the model-namely logs, context, or questions to enhance its performance. QA systems' effectiveness is contingent on the quality of data they engage with, underscoring the significance of prioritizing data quality.

Monitoring serves as a valuable tool to augment the system's capacity to deliver high-quality service, ensuring alignment with educational sector standards. Adopting a data quality-driven view, meticulous monitoring employs criteria like accuracy, cohesiveness, degradation over time, and completeness. This approach aims to enhance data quality and unveil the ramifications of low-quality data propagation across phases. In our methodology, the model remains unchanged while modifications are introduced to the context from which the QA system derives its answers. This prompts considerations about the relationship between sentence structure and QA system responses. [19] demonstrate that handling contexts with intricate syntactic structures, which yield more erroneous answers compared to simpler ones, poses increased difficulty. The quality of the question significantly influences the information retrieval of the QA system. Misspelled questions pose challenges, potentially hindering the system's accurate query comprehension. Such misspellings can cause confusion, resulting in misinterpretation and consequently inaccurate responses [25]. Altering the intent of the query, misspelled questions may reduce the relevance of retrieved information, impacting the system's response accuracy [32]. Limited exposure to these variations might affect the system's real-world query performance.

Numerous researchers express interest in automatically enhancing the quality of data prior to its utilization in a model [28]. Others research the impact of poor data quality on model predictions through adversarial attacks [8]. Several models of adversarial attacks are available, Alzantot *et al*. [1] and Andriushchenko *et al*. [2] proposed a black-box algorithm to generate adversarial attacks that lure sentiment analysis and textual models with a high success rate. Zang *et al*. [35] introduced a Word-level Textual Adversarial Attacking model that crafts high-quality adversarial examples than usual methods and applies it successfully to Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Encoder Representations from Transformers (BERT). Jin *et al*. [13] introduced TextFooler, to generate adversarial attacks on text classification and textual entailment that are effective, utility-preserving, and efficient.

This paper investigates the performance of QA systems using two deep-learning architectures: Transformers and Recurrent Neural Networks (RNNs). Additionally, we aim to examine how simplifying the context impacts the accuracy of responses. To achieve this, we have simplified the context in which the QA system draws the answer with a workflow of two simplifications. Simplifying the grammatical and syntactic structures within the context results in greater clarity and simplicity. Our approach to simplification maintains information integrity through two non-invasive methods. The first method addresses pronoun coreference inspired from Winograd challenge, thereby enhancing contextual comprehension. The second method involves breaking down lengthy sentences into shorter ones, which possess simpler syntactic structures, making them easier to construct and comprehend.

Our study focuses on comparing various readability and syntactic complexity measures with QA system outputs using Coh-Metrix readability formulas and L2 Syntactic Complexity Analyzer (L2SCA) syntactic complexity indices [16]. Analysis of the results demonstrates a strong correlation between context linguistic complexity and the scoring of QA systems. We conclude that by reducing both syntactic and lexical complexity, we can obtain improved answers. Furthermore, we assess the resilience of BERT and Reliable National Entrance Test (R-NET) against the same adversarial attack (such as letter substitution or typos).

2. Experiments

2.1. R-NET and BERT

R-NET is an end-to-end neural network model for reading comprehension and question answering [34]. R-NET is based on gated matching and recurrent network, which considers the importance of each word in context, like humans do, to answer the question. Therefore, each part of the context is assigned a different weight according to its importance to the question. Subsequently, a "gated matching layer" effectively processes the question while considering every word within the context (self-matching process). As a result, it can proficiently derive the answer through comprehensive analysis of the context.

BERT is a layered series of transformers encoders. It only uses a portion of transformers made up of encoders and decoders. Transformers are composed of multiple self-attention heads. In summary, each input token in a sequence creates a weighted representation from the key, value, and request vectors. The outputs of all the heads of the same layer are concatenated and pass through a fully connected layer. Each layer is wrapped with a jump connection followed by a normalization layer [12]. Some models can only read text sequentially-left to right or right to left-but could not read bi-directionally. BERT is different because it can read simultaneously in both directions.

The representation adopted by BERT has the particularity of being contextual. A word is not represented in a unique way as in a classic embedding but according to the context of the input text. For example, the word “figure” will have a different representation in “a parental figure” and “a figure totally underestimated”.

BERT is trained on a large amount of information. It can be fine-tuned by adding a layer to BERT and retraining the model on a small amount of data in a limited time [12]. There are several versions of BERT, in our case we have used: BERT-Base, which is made up of 12 layers, 768 hidden nodes, 12 attention heads, and 110 million parameters.

2.2. Methods

The methodology involves enhancing the quality of data provided to the model, thereby improving data accessibility and facilitating effective analysis. By focusing on data quality, this approach maintains the model's integrity while introducing changes to unstructured data, from which the QA system derives its responses. Well-structured, high-quality text reveals contextual patterns and latent relationships between entities, consequently influencing the chosen answer’s accuracy [36]. Conversely, poor-quality data containing symbols, unreadable sequences, equations, or intricate grammatical structures can introduce inconsistencies and mislead the QA system. Errors within the QA system’s context, such as spelling or grammatical errors, disparate encodings, paragraph duplication or similarity, and missing data, have the potential to alter its responses significantly.

According to the findings of the paper, modifications applied to the context to enhance the QA system's performance should uphold the same level of information. Simplification within the realm of QA systems should enhance the syntactic and grammatical structure of the context while maintaining information quantity. Karra and Lasfar [19] demonstrated the effectiveness of the ‘explicit/short sentence’ method, showcasing its superiority over two general simplification models, Access and KiS, in improving the QA system’s rate of correct answers. These models tend to delete words, thus diminishing the information within the context. Our upcoming studies aim to test a variety of simplification models and classify them based on their non-invasive nature and their ability to preserve contextual information rates.

2.2.1. Context Simplification

We prepared 81 questions with their related *context* from Wikipedia. Each question corresponds to the *context* as well as a *correct* answer [17]. In the first step, we inject the question and its context into the R-NET model giving us the first answer. The produced answer is compared to the *correct* answer to the *question*. We utilized identical categorization and naming conventions found in SQuAD [26]. The obtained answers are then divided into categories: *correct* GTA, *partial* (incomplete answer), and incorrect.

$$f_{params}\colon \begin{pmatrix} Question \\ Context_{with\_changes} \end{pmatrix} \rightarrow Y \begin{cases} Correct \\ Partial \\ Wrong \end{cases} \quad (1)$$

where changes={explicit or short sentences}.

As explained before, two processes were used to help the model in its predictions and thus improve the results obtained. Anaphora resolution involves the machine’s task of identifying the antecedent of an ambiguous pronoun within a statement. This task not only falls within the domain of NLP but also necessitates the application of common-sense knowledge and reasoning. Our approach differs from methods often used for improving NLP model capabilities, especially in QA systems. Instead of increasing the size of the model with more hidden layers and neurons, at a financially unsustainable pace, we opted for improving the QA system based on BERT and R-NET models by simplifying the contexts. The “explicit” method was motivated by the Winograd test, a challenge launched in 2010 by Hector Levesque [21, 22]. Here’s an example illustrating a Winograd schema:

- The board rejected the proposal because they doubted the candidate’s qualifications.
- The board rejected the proposal because they endorsed the candidate’s qualifications.

This demonstrates the ambiguity of the pronoun “they” and how the meaning shifts based on the chosen word (“doubted” or “endorsed”), illustrating the challenge in natural language understanding that the Winograd schema aims to address. Since pronouns convey less information, they are subject to interpretation or confusion. This ambiguity requires human knowledge and common sense to resolve it and represents a challenge for neural models.

With the “explicit sentences” method, we transform all contexts by replacing the pronouns with the nouns they replace. These changes are inspired from works carried out within the framework of information theory [7] and the notion of entropy conveyed by words in a sentence. Indeed, each word contains a quantity of information that it must transmit to the reader. Thereby, by adding the subject in the sentences, we add information to it, and thus we help the model better predict.

Pronouns do not contain as much information (entropy) as the nouns they represent. Further works have been carried out by [24] to see what makes individuals determine nouns that replace pronouns. Others have proposed classification methods to determine what the pronoun replaces and underlined the difficulty that this operation can cause [11].

One of the techniques to quantify the information carried by words recommends that the word contains more information if it surprises the receiver. In other words, the information provided by the word is not expected or predictable [15]. It reinforces the fact that pronouns contain less information (since they are more common than the nouns they replace, which are specific). So, sentences containing pronouns are less rich in information than those with the corresponding nouns [24]. Consequently, the model will have more facility to determine the correspondences if pronouns are replaced (especially BERT and R-NET who use attention and embedding as a basis).

The second solution is to reduce the sentence size. In this respect, some studies deduce that a sentence should contain only the number of words essential to convey the information [30]. Indeed, the entropy of each word is quantified and the information that a sentence must contain is limited, consequently, the length of the sentence must also be limited. According to a study by the American Press Institute, the longer the text, the more difficult it is for the reader to understand [3]. Thus, this study showed that the understanding score is 100% for sentences containing less than eight words. For lengths between 9 and 14 words, the previous score decreases to 90%, and if we exceed 43 words for the sentence the rate drops to less than 10%. [20] wondered how many words a Long Short-Term Memory (LSTM) model could no longer predict accurately. They experimented by shuffling or removing contexts that were farther than a certain number, denoted as ‘k’ with varying values of k. The researchers measured both the accuracy and predictive ability of the LSTM model. The study revealed that after approximately 50 tokens, the model’s perplexity remained constant indicating that the LSTM model ceased to consider words beyond a certain distance. As the length of text increases, the amount of information it can contain makes it difficult to understand and extract the answer [10]. It’s a simple and less invasive technique where each sentence briefly describes an idea. It consists of cutting long texts into units of ideas and separating them into independent sentences.

Questions with incorrect or partial answers are subject to two parallel and distinct treatments: Explicit and short sentences, as illustrated in Figure 1. We inject the modified sentences according to the two methods into the context’s model and thus obtain new answers, which we again classify as correct GTA, partial (incomplete response), and wrong. Thus, we obtain for each sentence a couple of {original context, first response}, and {modified context, second response}.

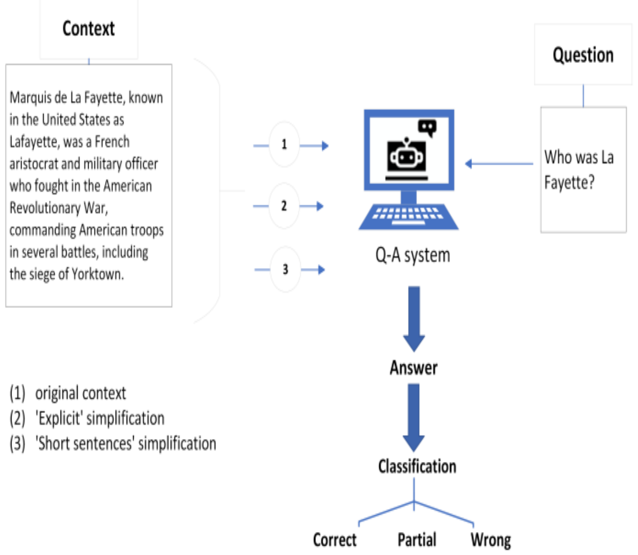

Context
Marquis de La Fayette, known in the United States as Lafayette, was a French aristocrat and military officer who fought in the American Revolutionary War, commanding American troops in several battles, including the siege of Yorktown.
1
2
3
Q-A system
Question
Who was La Fayette?
Answer
Classification
Correct
Partial
Wrong
(1) original context
(2) 'Explicit' simplification
(3) 'Short sentences' simplification

Figure 1. Context simplification process for QA system.

The achieved outcomes will serve for the comparison between the two techniques. The initial approach, known as explicit sentence alteration, substitutes all personal pronouns in the context with their respective nouns (inspired by the Winograd challenge). The second technique, termed “short sentences,” integrates concise sentences into the context. In both instances, we juxtapose the responses from the second method with the original answers.

To be able to quantify the impact of context changes on QA system performance, clear indices and measures must be used. The accuracy of the QA system is as follows:

$$Accuracy = 1 - \left(\frac{NWR}{All\ responses}\right) \tag{2}$$

where the Number of Wrong Responses is (NWR).

A separate analysis is conducted on different contexts to compare their complexity using multiple parameters and measure how much the proposed simplification (explicit, short sentences) enhances context comprehension and readability. The set of used measurements is taken from largely adopted computational tools. The measures quantify the length of clauses, sentences, and their grammatical structure. Based on some indicators, we can classify paragraphs according to their syntactic construction and differentiate between explicit and short sentence contexts. Coh-Metrix is a computational tool that automates text processing and returns text indicators [16].

$$FKGL = 0.39\left(\frac{num\ words}{num\ sentences}\right) + 11.8\left(\frac{num\ syllabes}{num\ words}\right) - 15.59 \tag{3}$$

Classically, we use Flesch-Kincaid Grade Level (FKGL) formula and Dale-Chall as measures for text readability [9]. It is important to examine how well the measurements of context complexity correlate with the QA system's correctness.

2.2.2. Question Misspelling

Adversarial attacks on QA systems entail crafting inputs to intentionally mislead the model and obtain inaccurate or unexpected answers. They can induce inappropriate responses, especially in formal settings such as education. These attacks seek to expose vulnerabilities in the model's understanding and decision-making process. We used the same questions as for the previous experiment (79) about Structured Query Language (SQL) and Python. Each question has its context as well as the corresponding correct answer. Subsequently, we developed an algorithm that systematically introduces errors into the questions in an automatic and random manner. Each question generates a series of thirty questions with varying errors (ten questions with one error, ten with two errors, and ten with three errors).

We created three databases of 790 questions each. The first contains questions with a Single Spelling Error (1SpErr), the second with Two Spelling Errors (2SpErr), and the third with Three Spelling Errors (3SpErr). Then, we fed the questions and their context into the model BERT, which produces an answer. We compare the new answers to the original answers to verify its accuracy. Finally, we classify the responses into three categories: correct, partial, and wrong.

3. Results and Discussion

3.1. Context Simplification

We have described in previous work [17] the effect of "explicit" and "short sentences" context changes on the QA system results based on R-NET model. Intersection is the percentage where the two methods have the same positive effect. Cumulative is the union of the two methods "explicit" and "short sentences". Initially, the model showed a correct response rate of 59%. If we count the partial answers as incorrect answers, this rate reaches 47%.

According to findings presented in Table 1, changing the context with the "explicit" approach improves the accuracy to 74%. Changing with the "short sentences" improves the accuracy to 75%. By combining the two methods "explicit" and "short sentences," we improve the accuracy to 85%. We repeat the same experiment with the BERT model previously described. The latter shows a rate of first correct answers of 75% by counting only the correct answers. By including responses in the partial category, this rate comes to almost 84%.

Table 1. Impact of "explicit" method on BERT model.

	Correct	Partial	Wrong	Total
Correct	60	1	0	61
Partial	0	7	0	7
Wrong	4	0	9	13
Total	64	8	9	81

The results show that the explicit method only changes three wrong answers, and the rate of correct answers increases to 79%, i.e., 4 points more. Table 2 shows that this method switches one correct answer to the partial category, and four incorrect answers have been corrected and moved to the correct category. Switching to the "short sentences" method gives 75% of correct answers, the same rate as the original data.

Table 2. Impact of "short sentences" method on BERT model.

	Correct	Partial	Wrong	Total
Correct	59	0	2	61
Partial	1	6	0	7
Wrong	1	0	12	13
Total	61	6	14	81

This result is due to the degradation of two answers which went into the wrong category against two partials, and one answers wrong which went into the "correct" category. Even if the impact of the two methods remains limited, the "explicit" method gives less important results than the "short sentences" method. The combination of the two methods gives a total of 83% correct answers, as shown in the following Table 3:

Table 3. Cumulative effect of the "explicit" and "short sentences" methods on BERT model.

	Correct	Partial	Wrong	Total
Correct	61	0	0	61
Partial	1	6	0	7
Wrong	5	0	8	13
Total	67	6	8	81

Table 4 shows the impact of context changes on the score of the QA system. The comparison of the results obtained with the two models shows two interesting results: The first is that even if BERT has the highest rate of correct answers before the changes with 77% of correct answers, R-NET obtain the best rate of answers after the cumulative application of the two methods. The impact of the two methods on R-NET is much greater ("explicit": 16 gain points, "short sentences": 17 points and 27 points for the two methods combined) than on BERT ("explicit": 4 gain points, "short sentences": 2 points and 8 for the two cumulative methods).

Table 4. Comparison of BERT and R-NET results.

Model	**Context changes**			
	No change	**Explicit**	**Short sentences**	**Cumulative**
R-NET	58%	74%	75%	85%
BERT	75%	79%	77%	83%

Even if these models use self-attention mechanism, there are some differences that explain the good result of BERT before the application of the two methods: R-NET model uses gated attention-based recurrent network to integrate question information in context representation. Text analysis is from left to right so the representation of each token can only consider the previous ones only [34]. This approach reduces the semantic representation of the model since the meaning of a word does not depend on either side. BERT uses transformers [12]. The contexts changed with the "explicit" method show a rate improvement of correct answers by 79% and those of "short sentences" to 77%.

R-NET benefits more than BERT from "short sentences" simplification. This is due to its Bi-LSTM architecture where the longer the context, with more tokens, the more the performance of the model deteriorates [20]. On the other hand, BERT based on Transformers uses the multi-attention mechanism and processes all the input tokens in parallel without differentiating the distance between the words. This is the reason its score after the "short sentences" simplification has not changed much.

However, the use of the two methods reduced the difference between these methods by 5 points for the "explicit" method and no difference for the "short sentences" method. It should be noted that R-NET records better results by combining the two methods, with 2 points more than BERT (83% against 85%).

In this part, we compare the original texts with those with "explicit" and "short sentences" changes. We explore several cohesion indicators, complexity, syntactic relations, or readability. Coh-Metrix is a popular library for the computational evaluation of text and coherence metrics. [16] define cohesion as properties of the explicit text that helps the reader in some way to mentally connect ideas in the text. It provides easy access to a wide range of syntactic complexity and easability formulas. Original context has a personal pronoun score wrdpro of 10.55 and explicit context a score of 9.70.

A context with high-frequency pronouns can produce referential cohesion problems for the QA system. The QA system can be misled by what each pronoun refers to. For L2SCA complexity per clause, we obtain a score of 1.778 for original context and 1.762 for "explicit" context (refer to Table 5). Whereas original context easability (4.518) is less than "explicit" context (4.738). As expected, we found a significant correlation between contexts' changes and their readability. The original context is rated more difficult to read and more syntactically complex. A context with fewer personal pronouns gives better results.

Table 5. The syntactic complexity measures of original and explicit contexts. We have selected questions and contexts to which we have applied the 'explicit' method.

	Indicator	Original	Short sentences
Coh-Metrix	Latent semantic analysis SS1	0.237	0.241
	Text easability PC connectivity	4.518	4.738
	Syntactic complexity SYNMEDpos	0.872	0.870
	Syntactic complexity SYNMEDwrd	0.855	0.853
	L2 readability	13.105	13.410
L2SCA	Complex nominal per clause	1.778	1.762
	Complex nominal per T-unit	2.209	2.198
	Overall sentence complexity	0.231	0.230

The results of Table 6 indicate a strong correlation between the readability and syntactic complexity scores and the "short sentences" changes. We found that the "explicit" and "short sentences" methods allow context simplification and better readability. These two methods give better performance of the QA system. The lower the complexity of the metrics, the better answers the QA system gets. The impact of the "short sentences" method is more pronounced than that of "explicit". Thus, the difference between original contexts and "short sentences" changes is more pronounced than "explicit". The results obtained confirm our previous finds [19]; the simplification of context improves the answers of the QA system.

Table 6. The syntactic complexity measures of original and "short sentences" contexts. We have selected questions and contexts to which we have applied the "short sentences" method.

	Indicator	Original	Short sentences
Coh-Metrix	Latent semantic analysis SS1	0.189	0.197
	Text easability PC connectivity	6.571	6.846
	Syntactic complexity SYNMEDpos	0.677	0.658
	Syntactic complexity SYNMEDwrd	0.894	0.882
	L2 readability	8.182	10.338
L2SCA	Complex nominal per clause	1.793	1.673
	Complex nominal per T-unit	2.345	1.991
	Overall sentence complexity	0.253	0.210

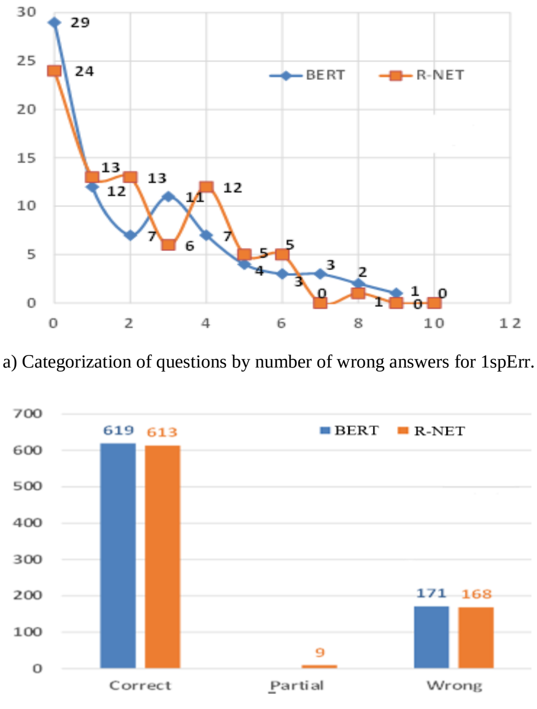
Employing a cautious simplification pipeline is advantageous for the QA system, particularly in contexts where sentences are infrequently deleted or rephrased. Our simplification method (combination of "explicit" and "short sentences") is not intrusive, does not delete data and gives better results than the original context. Thus, the difficulty lies in categorizing content as important or irrelevant. In some cases, paraphrasing results in incorrect simplification and more syntactic complex outputs [37].

Firstly, we suggest limiting context simplification for QA systems to lexical simplification using replacement. This process involves preprocessing and tokenization. Instead of removing data, we propose calculating the amount of information generated by each simplification operation (information entropy) for better assessment. R-NET has an F1-score of 80.34 and BERT 88.49 for SQuAD-v1.1. The initial results (original context) of the QA system reflect their performances. Nevertheless, BERT's advantage fades with context simplification and R-NET even outperforms BERT, after context simplification, with 85% accuracy. Again, it confirms the interest and role of simplification for less efficient models.

The results showed that the impact of the type of simplification is not the same, depending on the architecture on which the QA system is based. Simplifying text within the QA system necessitates maintaining the entropy of contextual information. Thus, emphasizing repetitions and ensuring clarity within the context is preferred over sentence deletion, which reduces entropy. Consequently, we advocate for a scalable simplification pipeline encompassing various types of simplification, including explicit changes, lexical alterations, shorter sentences, and declarative structures. This approach is deemed an effective alternative to pursuing more efficient models due to these considerations.

3.2. Questions Misspelling

Misspelling questions can be considered as an adversarial attack. Numerous studies have delved into assessing the robustness of the BERT model against adversarial attacks. [13] investigated the impact of attacks on BERT in sentence classification tasks, while [31] specifically examined its resilience against misspellings. Sun's findings suggest that BERT lacks robustness in such scenarios. Concerning the sentences with a single error, the percentage of correct answers amounts to 78.35% against 21.64% for incorrect answers and 0.9% of partial answers. Figure 2 shows that 29 sentences have no errors, twelve have only one wrong answer and 59 questions have three errors or less.



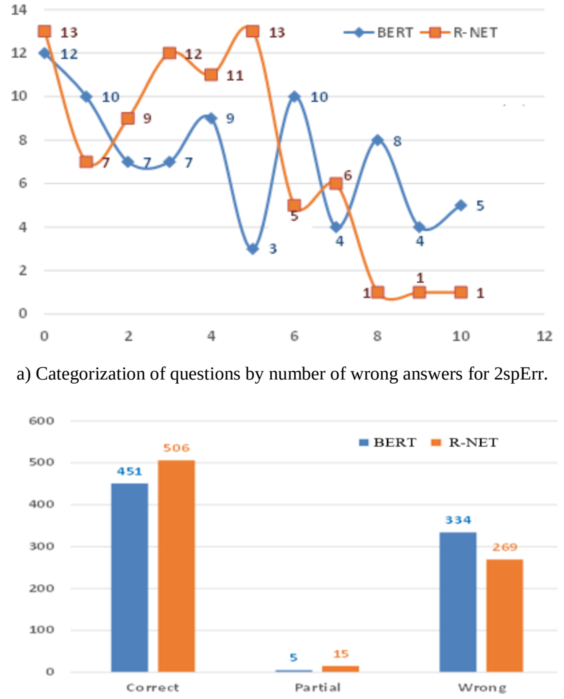
a) Categorization of questions by number of wrong answers for 1spErr.

b) Distribution of responses (BERT and R-NET).

Figure 2. Impact of 1spErr questions on QA system outputs.

The comparison between BERT and R-NET gives results close and slightly in favor of BERT. Indeed, BERT obtained 619 correct answers against 613 for R-NET. The distribution of sentences according to the wrong answers shows a similarity between models, with a slight advantage for BERT. As expected for two errors, there is a decline of 27% in the correct answers rate. It goes from 619 for the case of a single error to 451 for two errors.

The number of incorrect answers increased from 171 to 334, increasing by 95%. Figure 3 shows that even though the distribution of the number of wrong answers per sentence of R-NET and BERT has a similar trend (which indicates a positive correlation), R-NET performs better, especially in the case of 4 and 5 errors.

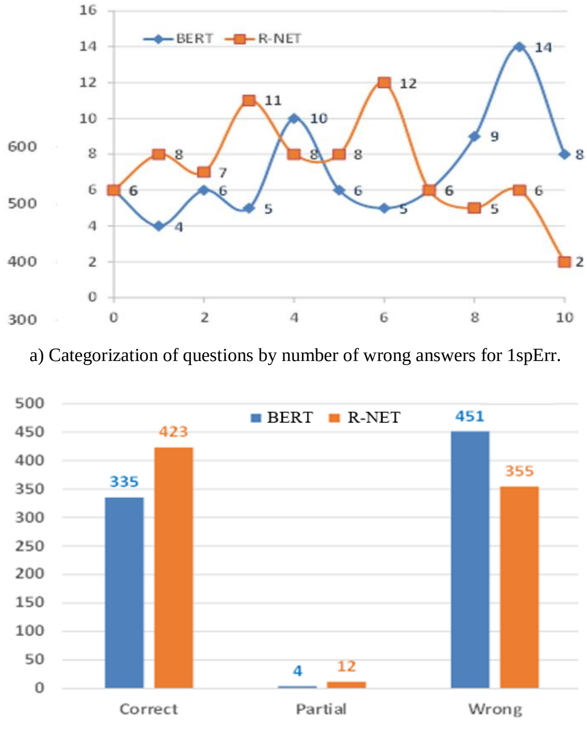


a) Categorization of questions by number of wrong answers for 2spErr.

b) Distribution of responses by the {correct-partial-wrong} categories.

Figure 3. Impact of 2spErr questions on QA system outputs.

Contrary to the results obtained for the sentences with one error, R-NET gets 506 correct answers against 451 for BERT(+12.2%). R-NET also registers fewer wrong answers with 269 against 334 for BERT(-19.5%). In the case of three errors, the number of BERT correct answers fell to 335 (-25.7%), and the number of incorrect answers increased from 334 to 451 (+35.0%). Partial answers remained almost the same from 4 to 5 (see Figure 4).



a) Categorization of questions by number of wrong answers for 1spErr.

b) Distribution of responses (BERT and R-NET).

Figure 4. Impact of 3spErr questions on QA system outputs.

R-NET performs 26.27% better than BERT in the case of three errors in terms of correct answers (423 correct answers against 335). R-NET also scores three times more partial answers and 21.29% fewer wrong answers. The previous results show that R-NET handles spelling errors better than BERT. Indeed, BERT and R-NET give similar results in the case of one error, and R-NET achieves better results (more right answers and fewer wrong answers) in the case of two and three errors. These results went against what we expected because BERT uses more parameters than R-NET, is based on Transformers architecture, and is trained on more data.

Despite BERT's superiority in F1 and EM indicators, R-NET demonstrates greater resilience to noisy questions. This resilience stems from the differences in their approaches to handling errors. BERT relies on subword embeddings, which overlook character changes. R-NET utilizes Convolutional Neural Network (CNN) character embedding, considering words as sets of characters.

However, the analysis of the architecture of each of the two models shows that these results can be explained by the embedding layers adopted by each of the models. BERT [12] adopts a token embedding, segment embeddings, and a positional embedding but no character embeddings like R-NET does, which has in addition to a word embedding, a character embedding which is useful for processing Out-Of-Vocabulary (OOV) [34]. These findings are in line with the results of [6] who used a modified version of BERT to obtain better results in the field of medicine. They thus replaced the sub-word embedding with character embeddings (character-CNN that is implemented as part of ELMo's architecture). In parallel, they obtained better results in the case of misspelling. Also [31] showed the limits of BERT against several types of word modifications and found that typing error harms performance the most because it can generate uncommon samples for sub-word embeddings.

3.3. Limitations

The primary limitation of this approach is the potential cascading effect, wherein an erroneous output from one of the upstream models could affect subsequent downstream treatments, creating a bottleneck. A secondary limitation pertains to the types of errors introduced into the questions for examining the robustness of the R-NET and BERT models. Our constraints were limited to errors involving character substitution. It is advisable to broaden the tests to encompass questions altered by character permutation, deletion, or modification of entire words. However, this necessitates further investigation and forms an interesting research path for the future.

4. Conclusions

This paper has investigated how context simplification enhances the QA system's performance. BERT and R-NET are the two base models of the QA system. Experimental results show that the BERT-based model initially performs better than the R-NET-based model, whereas R-NET catches up with context simplification. To enhance the performance of NLP models, emphasis should be placed on improving the quality and integrity of the training data rather than solely focusing on scaling NLP models based on parameter count, number of hidden layers, and dataset size. These models are specifically designed for individual tasks (such as coreference and sentence splitting) and are structured into a pipeline. It represents a cost-effective approach suitable for organizations with limited budgets, as the inference of LLMs presents a significant hurdle to their widespread implementation. We generate indicators from two automated analyzers of text complexity (Coh-Metrix and L2SCA). Their metrics are highly correlated with the QA system's correct answers. Readability and syntactic complexity indicators stand a useful way to evaluate the strength of a QA system. We have shown that R-NET is more robust to misspelling than BERT. Although BERT is based on Transformers and has superior performance, R-NET is more robust to adversarial substitution attacks. Embedding characters from a bidirectional RNN gives R-NET better resistance to attacks. Combining this characteristic with BERT performance can be beneficial for a QA system. Future work involves enhancing the evaluation of simplification methods by devising a conservative simplification metric rooted in SARI. Additionally, there is a prospect of automating the process of cycling through various types of adversarial attacks on QA systems.

References

[1] Alzantot M., Sharma Y., Elgohary A., Ho B., Srivastava M., and Chang K., "Generating Natural Language Adversarial Examples," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Brussels, pp. 2890-2896, 2018. https://aclanthology.org/D18-1316.pdf

[2] Andriushchenko M., Croce F., Flammarion N., and Hein M., "Square Attack: A Query-Efficient Black-Box Adversarial Attack Via Random Search," *in Proceedings of the European Conference on Computer Vision*, Glasgow, pp. 484-501, 2020. https://doi.org/10.1007/978-3-030-58592-1_29

[3] Ann W., Make Every Piece you Write Easier to Read and Understand, Wylie Communications Inc., https://freewritingtips.wyliecomm.com/2017-05-

09/, Last Visited, 2024.

[4] Asthana P. and Hazela B., *Intelligent Systems Reference Library*, Springer Nature, 2020. https://doi.org/10.1007/978-981-13-8759-3_16

[5] Bajaj P., Campos D., Craswell N., and Deng L., "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," *in Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, Barcelona, pp. 1-16, 2016. https://arxiv.org/abs/1611.09268

[6] Boukkouri H., Ferret O., Lavergne T., Noji H., Zweigenbaum P., and Tsujii J., "CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations from Characters," *in Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, pp. 6903-6915, 2020. https://aclanthology.org/2020.coling-main.609.pdf

[7] Brown T., Mann B., Ryder N., and Subbiah M., *Advances in Neural Information Processing Systems 33*, NeurIPS, 2020. https://proceedings.neurips.cc/paper/2020

[8] Chen L. and Chan H., "Generative Adversarial Networks with Data Augmentation and Multiple Penalty Areas for Image Synthesis," *The International Arab Journal of Information Technology*, vol. 20, no. 3, pp. 428-434, 2023. DOI: 10.34028/iajit/20/3/15

[9] Crossley S., Allen D., and McNamara D., "Text Readability and Intuitive Simplification: A Comparison of Readability Formulas," *Reading in a Foreign Language*, vol. 23, no. 1, pp. 84-101, 2011. https://files.eric.ed.gov/fulltext/EJ926371.pdf

[10] Crossley S., Allen D., and McNamara D., "Text Simplification and Comprehensible Input: A Case for an Intuitive Approach," *Language Teaching Research*, vol. 16, no. 1, pp. 89-108, 2012. DOI:10.1177/1362168811423456

[11] Denis P. and Baldridge J., "A Ranking Approach to Pronoun Resolution," *in Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, pp. 1588-1593, 2007. https://dl.acm.org/doi/10.5555/1625275.1625532

[12] Devlin J., Chang M., Lee K., and Toutanova K., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of the North American Chapter of the Association for Computational Linguistics*, Minnesota, pp. 4171-4186, 2019. https://arxiv.org/pdf/1810.04805.pdf

[13] Jin D., Jin Z., Zhou J., and Szolovits P., "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment," *in Proceedings of the 34th AAAI Conference on Artificial Intelligence,* New York, pp. 8018-8025, 2020. https://ojs.aaai.org/index.php/AAAI/issue/view/253

[14] Hien H., Cuong P., Nam L., Nhung H., and Thang L., "Intelligent Assistants in Higher-Education Environments: The FIT-EBot, a Chatbot for Administrative and Learning Support," *in Proceedings of the 9th International Symposium on Information and Communication Technology*, Danang City, pp. 69-76, 2018. https://doi.org/10.1145/3287921.3287937

[15] Ian G., Yoshua B., and Aaron C., *Deep Learning*, MIT Press, 2016. https://mitpress.mit.edu/9780262035613/deep-learning/

[16] Graesser A., McNamara D., Louwerse M., and Cai Z., "Coh-Metrix: Analysis of Text on Cohesion and Language," *Behavior Research Methods, Instruments, and Computers*, vol. 36, no. 2, pp. 193-202, 2004. https://link.springer.com/article/10.3758/BF03195564

[17] Karra R. and Lasfar A., "Effect of Questions Misspelling on Chatbot Performance: A Statistical Study," *in Proceedings of the International Conference on Digital Technologies and Applications*, Fez, pp. 124-132, 2022. https://doi.org/10.1007/978-3-031-02447-4_13

[18] Karra R. and Lasfar A., "Enhancing Education System with a Q and A Chatbot: A Case Based on Open edX Platform," *in Proceedings of the International Conference on Digital Technologies and Applications*, Fez, pp. 655-662, 2021. https://doi.org/10.1007/978-3-030-73882-2_59

[19] Karra R. and Lasfar A., "Impact of Data Quality on Question Answering System Performances," *Intelligent Automation and Soft Computing*, vol. 35, no. 1, pp. 335-349, 2023. https://doi.org/10.32604/iasc.2023.026695

[20] Khandelwal U., He H., Qi P., and Jurafsky D., "Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context," *in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, pp. 284-294, 2018. https://aclanthology.org/P18-1027

[21] Kocijan V., Davis E., Lukasiewicz T., Marcus G., and Morgenstern L., "The Defeat of the Winograd Schema Challenge," *Artificial Intelligence*, vol. 325, pp. 103971, 2023. https://doi.org/10.1016/j.artint.2023.103971

[22] Levesque H., Davis E., and Morgenstern L., "The Winograd Schema Challenge," *in Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning*, Rome, pp. 552-561, 2012. https://dl.acm.org/doi/10.5555/3031843.303190

[23] Liu Z., Ding G., Bukkittu A., and Gupta M., "A Data-Centric Framework for Composable NLP Workflows," *in Proceedings of the Empirical*

Methods in Natural Language Processing: System Demonstrations, Punta Cana, pp. 197-204, 2020. https://doi.org/10.48550/arXiv.2103.01834

[24] Nieuwland M. and Van Berkum J., "Individual Differences and Contextual Bias in Pronoun Resolution: Evidence from ERPs," *Brain Research*, vol. 1118, no. 1, pp. 155-167, 2006. https://doi.org/10.1016/j.brainres.2006.08.022

[25] Niu T. and Bansal M., "Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models," *in Proceedings of the 22nd Conference on Computational Natural Language Learning*, Brussels, pp. 486-496, 2018. https://aclanthology.org/K18-1047

[26] Rajpurkar P., Zhang J., Lopyrev K., and Liang P., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *in Proceedings of the Empirical Methods in Natural Language Processing Conference*, Texas, pp. 2383-2392, 2016. DOI:10.18653/v1/D16-1264

[27] Renggli C., Rimanic L., Gürel N., Karlaš B., Wu W., and Zhang C., "A Data Quality-Driven View of MLOps," *IEEE Data Engineering Bulletin*, vol. 44, no. 1, pp. 11-23, 2021. https://www.research-collection.ethz.ch/handle/20.500.11850/526606

[28] Schelter S., Lange D., Schmidt P., Celikel M., Biessmann F., and Grafberger A., "Automating Large-Scale Data Quality Verification," *Proceedings of the VLDB Endowent*, vol. 11, no. 12, pp. 1781-1794, 2018. https://doi.org/10.14778/3229863.3229867

[29] Singh P. and Manure A., *Natural Language Processing with TensorFlow 2.0*, Springer Nature, 2020. https://doi.org/10.1007/978-1-4842-5558-2_5

[30] Stepak A., "Frequency Value Grammar and Information Theory," *Journal Applied Science*, vol. 5, no. 6, pp. 952-964, 2005. DOI:10.3923/jas.2005.952.964

[31] Sun L., Hashimoto K., Yin W., Asai A., Li J., Yu P., and Xiong C., "Adv-BERT: BERT is not Robust on Misspellings! Generating Nature Adversarial Samples on BERT," *arXiv Preprint*, arXiv:2003.04985v1, 2020. https://arxiv.org/pdf/2003.04985.pdf

[32] Vilares J., Alonso M., Doval Y., and Vilares M., "Studying the Effect and Treatment of Misspelled Queries in Cross-Language Information Retrieval," *Information Processing and Management*, vol. 52, no. 4, pp. 646-657, 2016. https://doi.org/10.1016/j.ipm.2015.12.010

[33] Vogel J., Chatbots: Development and Applications, Bachelor's Thesis, HTW Berlin-University of Applied Sciences, 2017. https://jorin.me/chatbots.pdf

[34] Wang W., Yang N., Wei F., Chang B., and Zhou M., "Gated Self-Matching Networks for Reading Comprehension and Question Answering," *in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, pp. 189-198, 2017. https://aclanthology.org/P17-1018

[35] Zang Y., Qi F., Yang C., and Liu Z., "Word-Level Textual Adversarial Attacking as Combinatorial Optimization," *in Proceedings of the 58th Annual Meeting Association Computational Linguistics*, Seattle, pp. 6066-6080, 2020. https://doi.org/10.18653/v1/2020.acl-main.540

[36] Zhang W., Sheng Q., Alhazmi A., and Li C., "Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 3, pp. 1-41, 2020. https://doi.org/10.1145/3374217

[37] Zhao S., Meng R., He D., Saptono A., and Parmanto B., "Integrating Transformer and Paraphrase Rules for Sentence Simplification," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics*, Brussels, pp. 3164-3173, 2018. https://aclanthology.org/D18-1355

Rachid Karra received the Engineer degree in computer science engineering from National School of Mineral Industry (ENIM-2008) of Rabat and has PhD in Natural Language Processing. He participates in several IT-projects. He has several professional certifications in cloud and web technologies (Azure AI Engineer Associate certification; MCSA: Web Applications). His research interests include NLP, Neural Networks, Non-Linear Equations, Optimization and Data Quality.

Abdelali Lasfar is a professor authorized to direct the research. He holds a doctorate in computer science. Dr. Lasfar focused his research on image compression and indexing, text extraction from images, and behavioral studies. Expert on e-Learning Platforms. He was the Head of computer science department. He has been leading a number of research projects and has published articles in SCOPUS indexed journals and international conferences.