Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.026695
Article

Tech Science Press

Impact of Data Quality on Question Answering System Performances

Rachid Karra* **and Abdelali Lasfar**

LASTIMI Laboratory, Mohammadia School of Engineers, Mohammed V University in Rabat, Morocco
*Corresponding Author: Rachid Karra. Email: rachid.karra@est.um5.ac.ma
Received: 02 January 2022; Accepted: 15 February 2022

Abstract: In contrast with the research of new models, little attention has been paid to the impact of low or high-quality data feeding a dialogue system. The present paper makes the first attempt to fill this gap by extending our previous work on question-answering (QA) systems by investigating the effect of misspelling on QA agents and how context changes can enhance the responses. Instead of using large language models trained on huge datasets, we propose a method that enhances the model's score by modifying only the quality and structure of the data feed to the model. It is important to identify the features that modify the agent performance because a high rate of wrong answers can make the students lose their interest in using the QA agent as an additional tool for distant learning. The results demonstrate the accuracy of the proposed context simplification exceeds 85%. These findings shed light on the importance of question data quality and context complexity construct as key dimensions of the QA system. In conclusion, the experimental results on questions and contexts showed that controlling and improving the various aspects of data quality around the QA system can significantly enhance his robustness and performance.

Keywords: DataOps; data quality; QA system; nlp; context simplification

1 Introduction

One of the missions of universities is the dissemination of knowledge. With the COVID-19 pandemic, most of the courses have gone online [1]. Students found themselves alone overnight, without any support or assistance on distance learning. Forums try to help students reduce difficulties and dark spots in a course, typically by creating discussion groups between students and teachers [2]. Dialogue system such as QA agent is an additional brick in the tools made available to students for better courses assimilation [3]. Teachers write their online courses in the most suitable format according to the discipline, target audience, and specific course needs. This freedom generates a multitude of text structures and styles. This format's diversity is not suited for Natural Language Processing (NLP) or understanding (NLU).

There are two approaches to developing a QA agent: Model-Centric and Data-Centric [4]. In a Model-Centric view, we focus on the model (its parameters, iterations, and optimization) and leave the data with some imperfections. The process is to maintain the data in its initial state and iteratively modify hyper-parameters or training algorithms to improve the model. In the Data-Centric view, data is as important as

the configuration and model development cycle [5,6]. The model remains the same, and we do iterations to enhance data quality and modify its structure [7]. Several tools are used to improve their consistency by preparing homogenous and high-quality data [8]. Indeed, the data scientists' community has led an active and fascinating research topic on data quality for decades. Liu et al. [9] propose a data-centric approach to workflow development and the introduction of a standardized data representation to encode heterogeneous results for generative and retrieval models, such as transforming an unstructured text corpus into a knowledge graph. The practical aspect of increasing data quality is through methods such as double-checking, consistency measurement with Cronbach's alpha, or random checking, to name a few. Thus, verifying the consistency of data, in some cases, is more effortless than searching or generating new data. This approach will allow the model to increase accuracy and prediction. Improving data quality can have a better effect than more compromised data [8]. Cleaning up the noise in data can be equally effective as collecting new data.

Generative models trained on large-scale datasets are effective as a QA agent. If used without filter or checker, it runs into ethical issues like misogyny or racism due to the datasets they are trained on. Biases, especially stereotypes of gender, race, or religion, contained in the training data inevitably emerge in the model's behavior [10]. In GPT-3, [11] found male gender dominance with manual and physical occupations and the opposite in nursing and reception occupations. When it comes to training high-capacity models on large-scale datasets, there is a rising concern about data contamination which casts doubt on their actual performance and whether they are overly optimistic about their models' scores [11]. The challenge for generative models is to have normative data sources and an appropriate vocabulary to mitigate these biases. Therefore, adopting a QA system based on a generative schema is not recommended for a teaching environment.

The classic development cycle in a deep learning project begins with the scope definition and data collection (filtered depending on the problem treated). Then, we train-test-validate the model. Finally, we deploy it in production. Model building, deployment, and monitoring are all automated. The model is monitored to have feedback about his behavior in production and if it is working satisfactorily. Having good data and the right model is the start of the overall deploying and operating process. Niu et al. [12] applied different paraphrasing levels and grammar errors to the dialog model's training. The target became more robust to adversarial inputs and responded better to original inputs. DataOps is a data integration of end-to-end collaborative practices and culture across all project phases [5], with an emphasis on a data view. Continuous integration (part of DataOps) helps deliver high-quality software by iterative adjustment in data and parameters. A deep learning project has the particularity of having many stakeholders directly or indirectly involved in a Deep Learning project like developers, security, data scientists, and Data-engineers. Training and tuning the deep learning model is just part of the information system [13]. Therefore, its integration into the information system must not bring functionalities deterioration. Using these organizations' forms reduces life cycle development and allows better collaboration between team members. Rather than focus on enhancing the model only, DataOps cares about end-to-end data improvement. From the point of view of data quality, the log and monitoring data are used to track how the model reacts in real situations and how data changes cause the model performance to degrade [14,15]. The monitoring data is stored in a schema conforming to the needs and organized from a data perspective used to improve the model's training cycle, as shown in Fig. 1. One of the main challenges for DataOps is to provide strategies for quality data production across all the phases of the QA system lifecycle.

The main objective of this study is to identify how the quality of data affects QA systems and what key features will enhance the QA agent responses. The present article is organized into three main parts. In Section 2, we introduce several aspects of QA system data quality. Next, Section 3 describes, in detail, the methodology, the model, and proposed types of change. In Section 4, we present our experimental

analysis and the effects of various features on data quality. Also, we will question if syntactically complex sentences result in more wrong answers contrary to baseline sentences. Intuitively, a sentence that uses turns of phrase and less frequent vocabulary is considered complex construction. A direct sentence with basic vocabulary is considered as a baseline sentence suitable for QA system responses. Finally, Section 5 summarizes our findings and presents prospects.

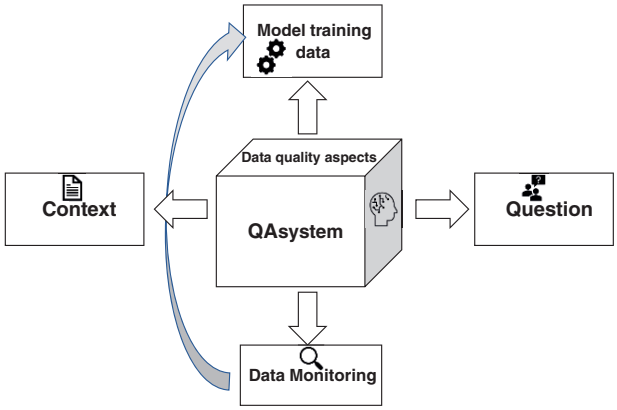

Figure 1: A description of the various aspects of QA system data quality

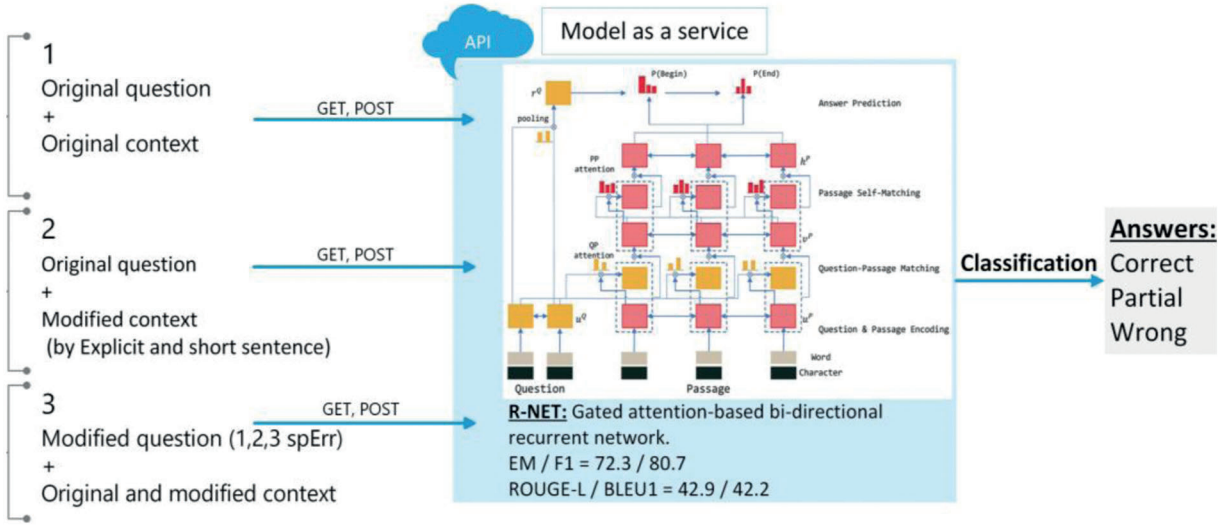

Figure 2: Overview of the experimental workflow

2 QA System's Data Quality

Instead of focusing on models' capacity only with the number of parameters and the size of training text data, we can improve the quality and integrity of the data used in model training. Renggli et al. [8] illustrated the importance of images quality and how noisy labels can alter the model's recognition results. Fig. 1 describes various aspects of data quality for a QA system. White arrows indicate the facets of the data quality of the QA system. The gray arrow presents the monitoring data relayed back to the training phase where the model is continuously changing and adjusting. It is composed of context, questions, monitoring, and training data. Logging data gives an idea about how the QA agent responds to students. We cannot control the quality of the question since it is an external input to the QA system. For the

context, investigating the syntactic structure and its complexity could shed light on some criteria to enhance QA system understanding and response extractions.

2.1 The Model

There are different levers we can use to improve QA agent responses. There are many options for developing an NLP system in general or a closed domain question-answering system. Either we can develop our model from scratch or fine-tune an existing model (transfer learning) to enhance capabilities on specific topics. In the first case, we use several large-scale datasets for training. With the emergence of large datasets, there is a race in their addition for training. To the authors' knowledge, the most efficient models to date are GPT-3, T5, and Wu-DAO 2.0. They are characterized by their large capacities, the number of parameters, and their energy-intensive consumption during training. GPT-3 is a pre-trained language representation, has the same structure as its previous generative model GPT-2, with 175 trillion parameters, and benefits from a set of large training datasets [11]. GTP-3 organized 45TB of data to extract clean 570GB training data. GPT-3 is task-agnostic that must train on task-specific datasets to better succeed in requested functions. The often followed agile approach is to train the model on a dataset. Then improve the dataset and re-train the model on the new dataset. Lastly, compare the results. However, the number of datasets should not gain importance over their relevance. Depending on the model's objective, selected datasets must contain a minimum of relevant features. Feature engineering consists of selecting datasets with relevant features or extracting new data from a combination of existing or new data sources.

The other approach is: we choose the appropriate datasets for our model and we train, test, and validate the model. Datasets that contain errors or inappropriate domain data result in a model that does not perform well [12]. Data scientists spend a lot of their time filtering and cleaning data, which shows the importance of high-quality data. Dataset choices influence the behavior of the model and its responses. Indeed, we cannot train a formal QA system on an informal chat social network dataset. Additionally, a model trained on this kind of dataset can have inappropriate racist or misogynistic utterances.

2.2 Monitoring

QA systems are only as good as the data they interact with, so it is no surprise that data quality is a major focus. Monitoring can be a valuable tool to enhance its faculty to deliver a high-quality service. It ensures that the QA system behaves in ways aligned with the educational sector. From a data Quality-Driven View, fine-grained monitoring tags some criteria like accuracy, cohesiveness, degradation over time, or completeness to improve data quality and discover the effects of low-quality data propagation through the numerous phases. Ré et al. [16] demonstrate how we can improve data quality throughout the system's lifecycle. One of the most promising approaches to improve the resilience of a QA system is to improve the external data (questions and contexts). By using approaches like context cleaning, formatting, and correction pipelines.

As DataOps attempts to understand, measure, and enhance the performances of Machine Learning (ML) models but also the distribution of data throughout the pipelines. The common approach where monitoring can improve the QA system is by continuous improvement [8]. The cycle starts by collecting insights. We study how the QA system responds to students' questions and the rate of correct responses and compare them to foreseen outcomes. Feedback loops provide the ability to adapt the model and deliver rapidly.

2.3 Context

The other way to go is improving the quality of data given to the model. Data quality must improve access to the appropriate information and make data analysis easier. It has a direct impact on the quality of the QA system. Wang et al. [17] defined four dimensions for data quality: accuracy to measure the

accordance with the correct value, relevancy about the field of application, representation to check if data are presented clearly, and accessibility to measure data availability.

In this approach, the model remains the same, and changes are made on the unstructured data from which the QA system draws its answer (see Fig. 2). A good quality structure can bring out patterns in the context, hidden relationships between entities and thus change the score of the chosen response [14]. Low-quality data with symbols and unreadable sequences like equations give inconsistent meaning and mislead the QA system about the right answer. Several errors can be in the course texts and affect the QA agent's response:

- Spelling or grammar errors;
- Using different encodings;
- Duplication or similarity of paragraphs;
- Missing data.

There are strategies to improve poor-quality data, like outliers elimination or replacing missing values by mean or neutral values [6]. For unstructured data like text, we can change ambiguous words by clear words, basic vocabulary, and even reformulate sentences or their orders. For this purpose, the authors adopted the text simplification strategy. The concept of text simplification was inspired by Crossley et al.'s study [18] on simplification of advanced text for L2 students learning a second foreign language. The paper is based on several parameters like words number, cohesion, and syntactic complexity to rate the text. Based on these criteria, we can choose the parameters that seem adapted to the context of our study. Crossley et al. [19] suggested that a vocabulary over-simplification may lead to more confusing text construction. It raises questions about the relationship between the construction of sentences and QA system responses. Is it more difficult to handle a context with complex syntactic sentences? And if it is the case, are there fewer errors in baseline sentences than in complex sentences?

2.4 Question

A priori, the way the question is asked influences the response extracted from the context. Errors in the question change the meaning and information the user is looking for and consequently the selected document or context [12,20]. A proxy applied to the question will prevent inappropriate behavior of the model in a structured and formal environment like education. Therefore, this proxy will necessarily have a positive impact on the QA system [14]. It must be accompanied by filters that detect and correct misspelling and grammar errors.

Correcting single words or correcting words in context are the two approaches to use. The first technique uses the Savary algorithm [21], based on edit distance and elementary repair hypothesis (insertion, deletion, substitution, and transposition) [14]. The second, utilizing Part-Of-Speech (POS)-tagging, inspects the entire context to find errors and correct them [20]. All these solutions can be grafted into the initial solution without requiring a large structure or code change using aspect-oriented programming [22]. Relying on the QA system model is not the only way to have reliable results. One of the aims of this study is to understand the effect of misspelling on QA system responses. Furthermore, we attend to explore how it evolves.

3 Materials and Methods

When choosing a QA agent for our e-learning system, we can opt for a global QA system that handles the questions of all courses or a QA system that points to a particular course. The global QA system requires an additional step; it must find the corresponding course to the student question and then answer it from the course content. The classification task is carried out by choosing the closest course to the context of the question. Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical technique used in many

applications, like search engines, text mining, and document retrieval [23]. It extracts the terms from the question and searches for the highest-scoring document from a list of ranked courses. Some common text retrieval techniques are neural information retrieval [9,24], Bag of Words (BOW), and BM-25 [25]. In our study, the main objective is to identify some levers that can help to enhance the correctness of QA system responses. Consequently, we will focus on a QA system that points to the course's data in its current version. It aims to respond to questions given a passage or context. All the contexts we use for our study are from Wikipedia. For database design, we rely on Open edX learning management system architecture (see Fig. 3) [26]. The QA system is presented as:

$$f_{params} : X_{q,c} \rightarrow Y \tag{1}$$

where f_{params} a nonlinear function, $X_{q,c}\begin{pmatrix} question \\ context \end{pmatrix}$, Y is the answer.

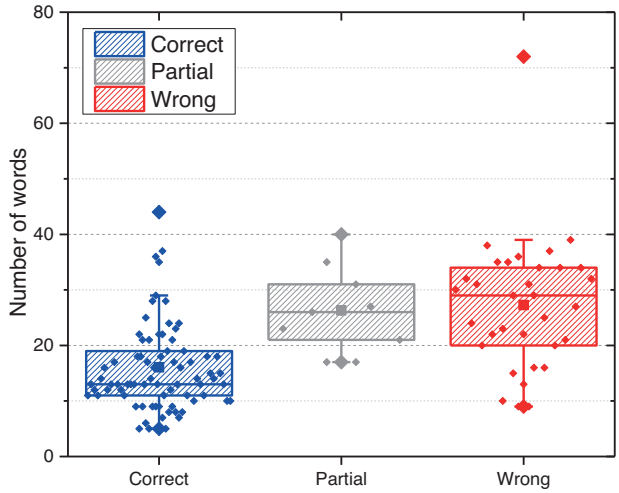

Figure 3: A comparative analysis of the context's sentence length using the three categories {Correct, Partial, Wrong}

3.1 Types of Context's Changes

There are two main approaches to improve the scoring of the QA system: improve the model or improve the quality of the data processed by the model. Extracting information from a text is more difficult for a text with a complex construction than a simple structure. Allen [27] specifies two methods to simplify the context: the structural and intuitive process. The intuitive methodology does not follow any constraints and is based on the writer's intuition and personal judgment [28]. In addition, the intuitive approach is subjective, imprecise, and difficult to quantify. On the other hand, the structural process has rules and a set of words and structures to follow, such as avoiding relative clauses use, relative pronouns, and other advanced grammar constructions like tense, passive voice, or subject-verb agreement. It allows the writers to have similar text construction [29]. These rules objectively guide the teacher to write a simple structured course.

Other context changes can improve the results of the QA system and several constructions (declarative sentence, relative clauses, passive sentence) that can be used and have a positive or negative effect on the QA system's understanding of the context. In our study, we have restricted our changes to two approaches: Explicit and Short sentences because they are less subject to interpretation, easy to apply, and less invasive.

The Explicit method is to replace personal pronouns with their proper names. Hence, the modified context is of lower quality and repetitive. Our goal is to facilitate the QA system's comprehension and not

the literary quality of the text or the written style. In a short sentence, we reduce the number of words in the text that contain the answer. As the length of the text increases, the amount of information it can contain makes it difficult to understand and extract the answer [18]. It is a simple and less invasive technique where each sentence briefly describes an idea. It consists in splitting long texts into units of ideas and separating them into independent phrases. Sentence length can correlate to the number of contained ideas. Short sentences have simple construction without irrelevant details and leave no doubt about the sentence meaning [27]. They are direct sentences with precis words that make the context easier to process. Tab. 1 presents an example of explicit and short sentence changes.

Table 1: Examples of context' changes and their impact on answers

Type of change	Components	Example
a) Explicit	1-Question	*Where was Peter hiding?*
	2-Original context	*...He was hiding in a bush and was so scared he could not move. ...Peter was hiding from his enemies.*
	3-Original answer	*from his enemies*
	4-Modified context	*...Peter was hiding in a bush and was so scared he could not move. ...Peter was hiding from his enemies.*
	5-New answer	*In a bush*
b) Short Sentence	1-Question	*What does DCL stand for?*
	2-Original context	*...may be informally classed as subcategories, commonly: a data query language (DQL), a data definition language (DDL), a data context language (DCL), and a data management layer (DML).*
	3-Original answer	*A data management layer*
	4-Modified context	*...may be informally classed as subcategories, commonly: a data query layer (DQL). A data definition language (DDL). A data context language (DCL). A data management layer (DML).*
	5-New answer	*A data context language*

Initial test (1-2-3) is followed by the adjusted context with the correct answer (1-4-5) from the QA system after applying one of the two types of adjustments.

3.2 Dataset Preparation

We prepared two datasets to test our QA system. The first is for studying the effect of context changes (QA-system-context-changes), while the second is for testing how the QA system reacts to questions spelling errors (QA-system-spelling-errors).

Dataset-1: In the first part, we worked on 82 questions about Python and SQL. The context of these questions is from Wikipedia. Every question includes its context as well as the correct answer. A first iteration is carried out by injecting the question with its context into the R-NET model. We compare the produced answer with the correct answer. Afterward, we classify the obtained answers into three categories: correct (ground truth answer), partial (incomplete answer), and wrong [30].

$$f_{params} : \begin{pmatrix} Question \\ Context_{with\ changes} \end{pmatrix} \rightarrow Y \begin{cases} Correct \\ Partial \\ Wrong \end{cases} \tag{2}$$

Questions with an incorrect or partial answer are subjected to two parallel and separate treatments: explicit and short sentences. Each 6-tuples contains a question, an original context, a first answer, a modified context, and a second answer. The results obtained will be used to compare the two methods. The first method is "Explicit" we replace all personal pronouns with their nouns in the context. The second method is "Short sentences" which entails incorporating short sentences into the context. In both scenarios, we compare the responses obtained from the model with the original correct responses.

Dataset-2: In the second part, we took only the questions and the modified contexts having a correct answer. We tested their behavior against degrading question quality [14,31]. The quality of the question can be affected by several factors, like spelling, grammatical and syntactic errors. Therefore, we restricted our choice to the least subjective: substitution error. Substitution involves misspelling or typing errors like changing a character in a specific word with another (example: "agree" become "afree"). Vilares et al. [32] used a misspelling generator with a range of error rates varying from 0% to 60%, with 10% steps between them. In our case, we tested only {1-2-3}-misspellings. Because in a concrete situation, it is unlikely that a student makes more than three misspelling errors in one question, even in noisy environments.

$$f_{params} : \begin{pmatrix} Question + \eta_{1,2,3} \\ Context \end{pmatrix} \rightarrow Y \begin{cases} Correct \\ Partial \\ Wrong \end{cases} \tag{3}$$

where: $\eta_{1,2,3}$ is the perturbation in questions.

We have divided the spelling errors into three degrees: 1-spelling error, 2-spelling error, and 3-spelling error. The next step is to generate ten questions with an n-spelling error for each question and category. The result of our dataset generation is a set of 670 tuples per n-spelling error for a total of 2010 tuples. The last step is to compare the provided answer with the correct one. During the tests, the context remained the same.

In this study, we rely on 300-dimensional pre-trained word vectors wiki-news-300d-1M.vec. It is a 1-million-word representation trained on the following datasets: Wikipedia 2017, UMBC webbase corpus, and statmt.org news [33]. During our tests, we used a dialogue system based on R-NET as it is [34]. R-NET is trained on large-scale datasets for understanding and question answering, namely SQuAD and MS-MARCO. SQuAD has plus 100,000 questions related to 365 Wikipedia documents and necessary responses to questions based on the context. It restricts responses to the space of all potential spans inside the reference context, as opposed to cloze-style reading comprehension datasets, it limits responses to particular words or entities [30].

The word embedded question and context are pre-processed by a bidirectional recurrent neural network (Bi-RNN) independently [35]. After that, we use an attention-based RNN, getting a perceptive question representation of the context. Then, we apply self-matching attention to assemble indications from the entire context and fine-tune the context representation. Finally, it is fed to the output layer to get the answer prediction through a pointer network. In simple terms, it predicts the answer start and end positions in each passage by a specific question. R-NET obtains ~80 F_1 score and ~72 EM score on the SQuAD-v1.1 dev set.

4 Experimental Results

4.1 Impact of Context Changes on QA System Responses

Now that we have described the experimental setup (DS-1: QA-system-context-changes), we report the performances of each of the proposed experiments. Since this study focuses on the impact of context and questions on the QA system, the subsequent subsections will describe the results of the experiments, analyze them, and discuss the effect of "Explicit" and "Short sentence" changes of the context. It is necessary to rely on clear indices and measures, to scale the impact of data quality on QA system performance [36]. Accuracy for the QA system appears as follow [37]:

$$Accuracy = 1 - \frac{Number\ of\ wrong\ responses}{Number\ of\ all\ responses} \tag{4}$$

In the initial group, we started with a QA system having an accuracy value of 0.58. It is an average score and reflects an unsatisfactory behavior of the QA system. Since the accuracy value of 1 represents an excellent QA system, and near 0 is a mediocre one. If we count the partial answers as wrong answers, accuracy decreases to a value of 0.47.

Changing the context with the "Explicit" approach improves the score by 0.16, resulting in an $Accuracy_{Explicit}$ of 0.74. The "Short sentence" change improves the score by 0.17, resulting in an $Accuracy_{Short}$ of 0.75. By grouping the "Explicit" and "Short Sentence" changes, we improve the accuracy of the QA system to 85% (see Tab. 2). The fact that the two methods intervene on different sentences of the context (intersection 2%) makes these two methods complementary.

Table 2: Impact of context changes on the score of the Q&A agent. Intersection is the percent where the two methods have the same positive effect. Cumulative is the union of the two methods **"**Explicit**"** and **"**Short Sentence**"**

Context changes				
No change	Explicit	Short sentence	Intersection	Cumulative
58%	74%	75%	2%	85%

4.2 The Length of the Sentence in the Context

Fig. 3 indicates the noteworthy influence of sentence length on QA system results. Correct sentences have a lower average words length than partial or wrong answers. There is a good correlation between correct answers and short sentences of the context. According to the results, the category of wrong answers has a higher sentence length than other categories. Wrong answers have a median value of 27 and 25 for the partial category.

The following Fig. 4 shows the distribution of the QA system responses according to the sentence length. Comparatively, the category of correct answers has the lower value for the median 16. While wrong and partial answers do not have a significant difference in the average sentence length, correct answers length is lower than the other two categories.

As seen in Fig. 4, the probability of correct replies decreases as the length of the statement increases. Fig. 4 shows that most incorrect responses go into the range 'W', whereas most correct answers fall into the range 'C'. As a result, about 80% of the QA system's correct answers are in the 'C' area. The distribution of partial responses is principally in range 'W' rather than range 'C'. These results provide no relevant information about partial answers except that the partial responses have an average length

close to wrong answers. For the 'C' area, we obtain an $Accuracy_{C-area}$ of 82%. Whereas, for the 'W' area, the score drops by half to an $Accuracy_{W-area}$ of 42%.

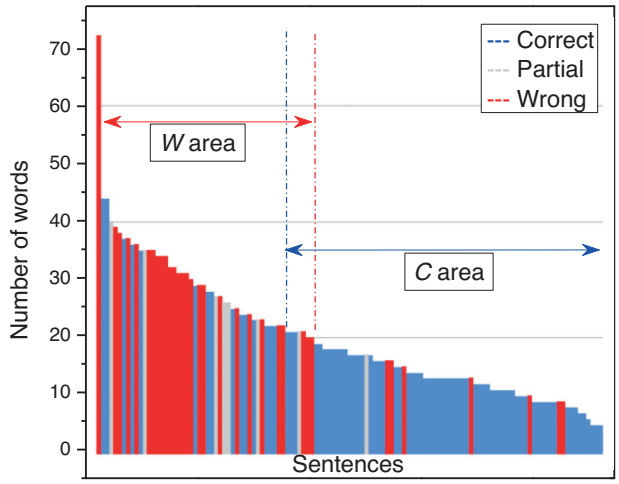

Figure 4: Influence of sentence length on the correctness and repartition of responses

Based on the previous results, the authors could observe that the QA system is more efficient when the context is constructed of short sentences. Likewise, shorter sentences increase QA system performance by almost 20%. Thus, the "Short sentence" can be considered a reliable criterion for QA system data quality.

4.3 Adversarial Attack on Questions

In this section, we will evaluate the robustness of the QA system when exposed to misspelling mistakes in questions (DS-2: QA-system-spelling-errors). The quality of the question takes multiple aspects: spelling, grammar, and syntax errors. Making errors can compromise the understanding of the QA agent and thus the response it will provide.

According to the results, the 1-spelling error test reached 137 wrong answers. With only one misspelling error, we have augmented the errors by 20%. For the 2-spelling error test, 224 answers belong to the category of wrong answers, 14 to the partial category, and 432 to the correct, representing 33%, 2%, and 65%, respectively. The 3-spelling error category reached 295 wrong answers representing approximately 44% of correct answers. These results showed that the category of wrong answers represents a trend to decrease with the increase of misspelling errors. On the other hand, the partial category varies slightly around 11 but does not give a clear trend. It remains negligible compared to the two main categories (Correct, Wrong). Similar to the explicit and short sentence categories, we calculate the accuracy for the n-spelling errors ($Accuracy_{1spError}$ is 79%, $Accuracy_{2spError}$ is 67%, and $Accuracy_{3spError}$ is 56%). The results show more pronounced degradation in the first misspelling error than in the second or third mistake. Detailed distribution of responses by the number of errors and context changes is in Appendix A in the supplementary materials.

The curves of the category of correct answers show that contexts labeled with the "Explicit" method are more "resistant" to spelling errors. Indeed, we see a slower degradation of correct responses than for the "Short sentence" context or the original context. Likewise, the growth in incorrect answers for the original context "Explicit" is slower than for the other categories. Also, the correct answers follow the same trend. Although without reaching statistical significance, the "Explicit" category shows better "resistance" to spelling errors, it plays somewhat an immunization role against adversarial attacks. Thus,

the authors recommend using the "Explicit" method in the first place for context changes before applying "short sentence" change.

The partial category is difficult to automate for an NLP system and usually requires human intervention to classify. In addition, the partial category does not follow a clear trend or pattern. Figs. 5 and 6 show that the partial category is closer to the wrong category and has several shared characteristics with it. To summarize, with "Explicit" and "Short Sentence" changes, the model becomes resistant and shows greater robustness to adversarial examples.

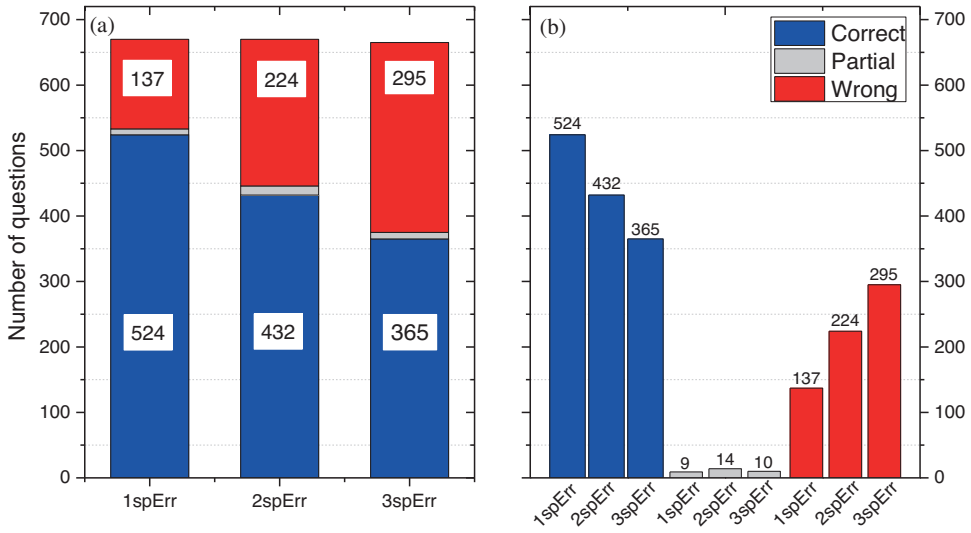

Figure 5: A comparative analysis between {1-2-3}-spelling errors and their impact on categories of responses

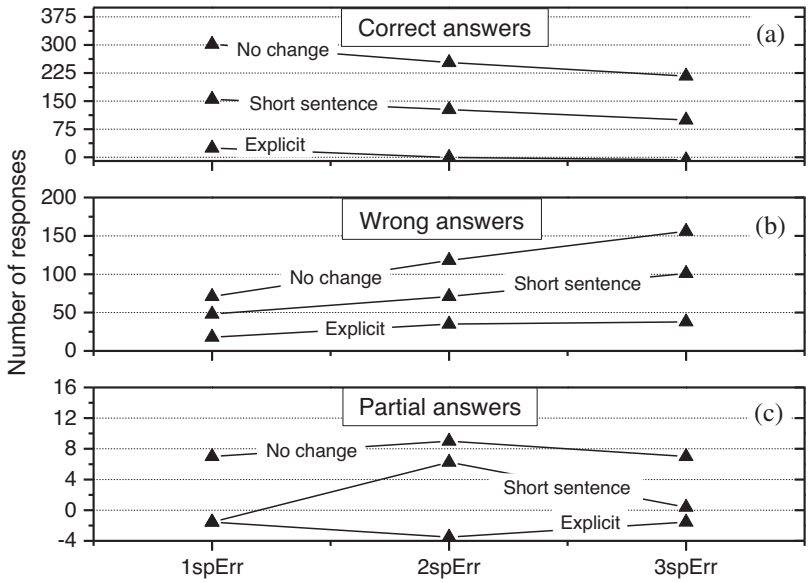

Figure 6: Charts showing the evolution of correct (a), wrong (b) and partial (c) answers with respect to the type of change and the number of errors in the question

5 Discussion

Deciding on the quality of unstructured data is very subjective. For the authors, it depends on the needs and the purpose for which we use the data [37]. Adapting the course to the QA agent is one of the possible ways to improve the responses. According to the preceding sections, there are several ways to process any new course to match the QA system requirements [15]. In some cases, the use of personal pronouns confuses the QA system. Replacing personal pronouns with their explicit values improves QA system responses (see Tab. 1). We have shown that it is more difficult to bring out the correct answers from contexts where the sentences are long. The adoption of short sentences for the context facilitates the work for the QA system. Of course, the efficiency of text simplification involves adopting as much as possible, simple, clear, and non-invasive methods to reduce the syntactic complexity of the context.

As a summary, this study has made the following observations:

- Not every business can use large models due to the size, technical limitations, and infrastructure expenses. Therefore, the authors recommend using more common models like R-NET or BERT and focusing on the model's external actors and their data quality.
- We introduced a simple and original way, based on context's data quality, for enhancing QA system performance.
- Our study focused on “Explicit” personal pronouns and short sentences as two ways to simplify the text and make the QA system robust to adversarial attacks.
- Having as clear and correct questions as possible is a promising strategy to improve the QA system scoring.
- Based on the results, the authors recommend a spelling correction pipeline to avoid student errors and get a better QA system scoring [38]. A vast choice of pipelines is at our hands, we can use yandex. Speller, Damerau Levenshtein [38–40].
- Use a version dedicated to the QA system in e-learning systems implementing the “Explicit” and “Short sentence” criteria.
- Both changes play an immunization role against question misspelling (Appendix A).

Although the two used methods record good results, there are limitations related to the impossibility of applying the two methods to all sentences. Indeed, some incorrect responses following the first use of R-NET were already explicit and short sentences. Also, we should test the proposed method on other models to confirm these findings such as BERT and GPT-2 with good scores in NLP tasks [41]. A second limitation lies in the lack of ability to automate the two methods for the corpus of every course. Finnaly, The study focus on accuracy as the main criteria for the evaluation of the “explicit” and “Short sentence” methods whereas there is other quality dimensions for an NLP model like robustness, precision, and perplexity.

6 Conclusion

The traditional way of writing a course for an e-learning system is not suitable for automatic language processing. We have also seen that the race for new models is not the only voice to follow. In addition to the draft version and the published version, we propose to have a QA agent version. The structure of this version should be based on short sentences and avoid personal pronouns. It will increase the QA system performance and the rate of correct answers.

Thus, some combinations of context simplification or short sentences with error-free questions can enhance the QA system answers. Thereafter, we can choose the features which seem adapted to the context of our experience and the QA system, such as the number of words or the complexity of the vocabulary. According to those experiments, questions misspelling was a source of incorrect responses

and thus caused the QA system's failure to respond to student inquiries. Thereby, it is necessary to have a QA system that can deal with misspelling questions. Having an effective QA system is not just dependent on model intrinsic performance. The structure of the question, as well as the context, must be considered. In future work, we will study context classification based on previous criteria, investigate types of constructions that perform better than others, and propose a global text simplification process for QA systems.

Data Availability: The data used to support the findings of this study are available at https://www.github.com/rachidkarra/ds-questions-contexts

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. L. Bradbury, X. P. Suarez-Sousa, M. Coquyt, T. L. Bockelmann and A. L. Pahl, "Teaching under crisis: Impact and implications of the COVID-19 pandemic on education in Minnesota," *The Interactive Journal of Global Leadership and Learning*, vol. 1, no. 2, pp. 80, 2020.

[2] C. Rapanta, L. Botturi, P. Goodyear, L. Guàrdia and M. Koole, "Online university teaching during and after the covid-19 crisis: Refocusing teacher presence and learning activity," *Postdigital Science and Education*, vol. 2, no. 3, pp. 923–945, 2020.

[3] T. Wambsganss, L. Haas and M. Soellner, "Towards the design of a student-centered question-answering system in educational settings," in *the 29th European Conf. on Information Systems*, Marrakesh, Morocco, pp. 1–12, 2021.

[4] L. Schmarje, Y. -H. Liao and R. Koch, "A Data-centric image classification benchmark," in *The 35th Conf. on Neural Information Processing Systems*, Australia, pp. 7, 2021.

[5] S. Trewin, "Setting up the framework," in *The DataOps Revolution: Delivering the Data-Driven Enterprise*, 1st ed., publisher. Auerbach Publications, Boca Raton, Fla, USA, no. 8, pp. 115–131, 2021.

[6] O. Azeroual, G. Saake and J. Wastl, "Data measurement in research information systems: Metrics for the evaluation of data quality," *Scientometrics*, vol. 115, no. 3, pp. 1271–1290, 2018.

[7] A. Deutsch, L. Sui and V. Vianu, "Specification and verification of data-driven web applications," *Journal of Computer and System Sciences*, vol. 73, pp. 442–474, 2007.

[8] C. Renggli, L. Rimanic, N. M. Gürel, B. Karlaš, W. Wu *et al.,* "A data quality-driven view of mlops, " *IEEE Data Engineering Bulletin*, vol. 44, no. 1, pp. 11–23, 2021.

[9] Z. Liu, G. Ding, A. Bukkittu, M. Gupta, P. Gao *et al.,* "A Data-centric framework for composable NLP workflows," in *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations*, Punta Cana, Dominican Republic, 2020.

[10] P. Costa, "Conversing with personal digital assistants: On gender and artificial intelligence," *Journal of Science and Technology of the Arts*, vol. 10, no. 3, pp. 59–72, 2018.

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan *et al.,* "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[12] T. Niu and M. Bansal, "Adversarial over-sensitivity and over-stability strategies for dialogue models," in *Proc. of the 22nd Conf. on Computational Natural Language Learning*, Brussels, Belgium, pp. 486–496, 2018.

[13] J. Davis, K. Daniels and J. Davis, "Collaboration: Individuals working together," in *Effective DevOps*, 2nd ed., CA, USA: O'Reilly Media, pp. 37–67, 2016.

[14] W. E. Zhang, Q. Z. Sheng, A. Alhazmi and C. Li, "Adversarial attacks on deep learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 24, pp. 1–41, 2020.

[15] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Danemark, pp. 2021–2031, 2017.

[16] C. Ré, F. Niu, P. Gudipati and C. Srisuwananukorn, "Overton: a data system for monitoring and improving machine-learned products," in *Conf. on Innovative Data Systems Research*, Asilomar, CA, USA, 2019.

[17] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[18] S. A. Crossley, D. Allen and D. S. McNamara, "Text simplification and comprehensible input: A case for an intuitive approach," *Language Teaching Research*, vol. 16, no. 1, pp. 89–108, 2012.

[19] S. A. Crossley, M. M. Louwerse, P. M. McCarthy and D. S. McNamara, "A linguistic analysis of simplified and authentic texts," *The Modern Language Journal*, vol. 91, no. 1, pp. 15–30, 2007.

[20] J. Vilares, M. A. Alonso, Y. Doval and M. Vilares, "Studying the effect and treatment of misspelled queries in cross-language information retrieval," *Information Processing & Management*, vol. 52, no. 4, pp. 646–657, 2016.

[21] A. Savary, "Typographical nearest-neighbor search in a finite-state lexicon and its application to spelling correction," in *Int. Conf. on Implementation and Application of Automata, CIAA 2001*, Pretoria, South Africa, vol. 2494, pp. 251–260, 2001.

[22] D. Robinson, "Creating pluggable code," in *Aspect-Oriented Programming with the e Verification Language*, Elsevier, Amsterdam, 2007.

[23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[24] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai *et al.,* "A deep look into neural ranking models for information retrieval," *Information Processing & Management*, vol. 57, no. 6, pp. 102067, 2020.

[25] A. I. Obasa, N. Salim and A. Khan, "Hybridization of bag-of-words and forum metadata for web forum question post detection," *Indian Journal of Science and Technology*, vol. 8, no. 32, pp. 1–12, 2016.

[26] R. Karra and A. Lasfar, "Enhancing education system with a Q&A chatbot: A case based on open edX platform," in *Int. Conf. on Digital Technologies and Applications, ICDTA 21*, Fez, Morocco, pp. 655–662, 2021.

[27] D. Allen, "A study of the role of relative clauses in the simplification of news texts for learners of English," *System*, vol. 37, no. 4, pp. 585–599, 2009.

[28] D. N. Young, "Linguistic simplification of SL reading material: Effective instructional practice?," *The Modern Language Journal*, vol. 83, no. 3, pp. 350–366, 1999.

[29] T. Jin and X. Lu, "A Data-driven approach to text adaptation in teaching material preparation: Design, implementation, and teacher professional development," *Tesol Quarterly*, vol. 52, no. 2, pp. 457–467, 2018.

[30] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing*, Austin, Texas, pp. 2383–2392, 2016.

[31] Z. Zhao, D. Dua and S. Singh, "Generating natural adversarial examples," in *Int. Conf. on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, vol. 6, 2018.

[32] J. Vilares, M. Vilares and J. Otero, "Managing misspelled queries in IR applications," *Information Processing & Management*, vol. 47, no. 2, pp. 263–286, 2011.

[33] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation, LREC 2018*, Miyazaki, Japan, vol. 11, 2018.

[34] B. Wang, T. Yao, Q. Zhang, J. Xu, Z. Tian *et al.,* "Document gated reader for open-domain question answering," in *Proc. of the 42nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Paris, France, pp. 85–94, 2019.

[35] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky and S. Khudanpur, "Recurrent neural network based language model," *Interspeech*, vol. 2, no. 3, pp. 4, 2010.

[36] A. Osman, N. Salim and F. Saeed, "Quality dimensions features for identifying high-quality user replies in text forum threads using classification methods," *PLoS One*, vol. 14, no. 5, pp. e0215516, 2019.

[37] Y. W. Lee, L. L. Pipino, J. D. Funk and R. Y. Wang, "Assessing data quality, part 2," in *Journey to Data Quality*, Cambridge, Mass: MIT Press, no. 4, pp. 53–67, 2006.

[38] E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction," in *Proc. of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong, pp. 286–293, 2000.

[39] Z. Z. Wint, T. Ducros and M. Aritsugi, "Spell corrector to social media datasets in message filtering systems," in *The Twelfth Int. Conf. on Digital Information Management, ICDIM 2017*, Fukuoka, Japan, pp. 209–215, 2017.

[40] Y. Chaabi and F. Ataa Allah, "Amazigh spell checker using damerau-levenshtein algorithm and n-gram," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, pp. S1319157821001828, 2021.

[41] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, vol. 1, pp. 4171–4186, 2019.

Appendix A

Distribution of QA system's responses by context changes {Original, Explicit and Short sentence} and number of errors {1sp-1sp-3sp Errors}.

	Original context			Explicit change			Short sentence change		
	Correct	Partial	Wrong	Correct	Partial	Wrong	Correct	Partial	Wrong
1*spError*	302	7	71	61	1	18	161	1	48
2*spError*	253	9	118	45	0	35	134	5	71
3*spError*	217	7	156	41	1	38	107	2	101
$\Sigma - spError$	772	23	345	147	2	91	402	8	220