# To Answer or Not to Answer? Filtering Questions for QA Systems

Paulo Pirozelli[1,4](✉), Anarosa A. F. Brandão[2,4], Sarajane M. Peres[3,4], and Fabio G. Cozman[2,4]

[1] Instituto de Estudos Avançados, São Paulo, Brazil
paulo.pirozelli.silva@usp.br
[2] Escola Politécnica, São Paulo, Brazil
{anarosa.brandao,sarajane,fgcozman}@usp.br
[3] Escola de Artes, Ciências e Humanidades, São Paulo, Brazil
[4] Center for Artificial Intelligence (C4AI), São Paulo, Brazil

**Abstract.** Question answering (QA) systems are usually structured as strict conditional generators, which return an answer for every input question. Sometimes, however, the policy of always responding to questions may prove itself harmful, given the possibility of giving inaccurate answers, particularly for ambiguous or sensitive questions; instead, it may be better for a QA system to decide which questions should be answered or not. In this paper, we explore dual system architectures that filter unanswerable or meaningless questions, thus answering only a subset of the questions raised. Two experiments are performed in order to evaluate this modular approach: a classification on SQuAD 2.0 for unanswerable questions, and a regression on Pirá for question meaningfulness. Despite the difficulties involved in the tasks, we show that filtering questions may contribute to improve the quality of the answers generated by QA systems. By using classification and regression models to filter questions, we can get better control over the accuracy of the answers produced by the answerer systems.

**Keywords:** Question answering · Answer triggering · Dual system · Question filtering

## 1 Introduction

A question answering (QA) system may be thought as conditional generator that predicts an answer given a question, often with the support of some context. In other words, for every input question, the system returns an answer. In practice, however, no QA system is able to answer any possible question, even over a

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/ 07665-4) and by the IBM Corporation. Fabio G. Cozman thanks the support of the National Council for Scientific and Technological Development of Brazil (CNPq grant #312180/2018-7).

J. C. Xavier-Junior and R. A. Rios (Eds.): BRACIS 2022, LNAI 13654, pp. 464–478, 2022.
https://doi.org/10.1007/978-3-031-21689-3_33

restricted domain, except when questions are restricted to a strict format. The reasons for this limitation are varied: from the lack of necessary information, to malformed questions, ambiguity, and tacit assumptions.

The policy of always answering questions may be a detrimental one in some contexts. This is particularly true for high-risk AI systems, where there are "significant risks to the health and safety or fundamental rights of persons" [6]. In such domains, it is often better for a QA system to confine itself to questions it is strongly certain of and avoid answering sensitive or dubious requests. Deciding which questions should move through the QA system is important to guarantee safety and factual grounding [23].

If it is the case that some questions should not be answered at all, then a QA system has to understand which questions should receive an answer and which should not. Two main approaches can be taken for that purpose. First, a QA system may be trained directly on datasets containing answerable and unanswerable questions, simultaneously learning *when* and *what* to respond. Second, a QA system may have a dual composition, in which a first component filters inadequate questions and the other module answers the selected questions.

In this paper, we investigate QA systems of the latter type; in particular, we analyze the connection, in those dual systems, of question filtering and QA quality. Our aim is to understand: i) how accurate can a classifier/regressor be on question answerability and meaningfulness; and ii) how do filtering systems affect the quality of the answering system. Although end-to-end systems have produced state-of-the-art results for answer triggering datasets [5,11,15], by identifying answerable questions as part of the training process, and have been under recent scrutiny as regards their self-evaluation awareness [9], our goal here is to quantify the relation between the level of answerability and meaningfulness of a question and the quality of the generated responses. By using a dual system architecture, in which a model previously identifies inappropriate questions, it is possible to modularize a critical step of a QA system, assuring a better control of what is being answered.

This paper is organized as follows. Section 2 provides an overview of the technical literature and describes the data used in our experiments. Two main tasks are conducted by us: first, we consider a traditional (discrete) answer triggering approach, in which the problem is to figure out which questions have an answer (Sect. 3); then, we explore a regression system in which we measure a question's degree of meaningfulness (Sect. 4). In both cases, we explore the effectiveness of filtering systems and how they affect the outcome of answerer models. In Sect. 5 we discuss the results of our investigation as well as limitations of the current approaches and future directions for research. Finally, we conclude with a few remarks on dual QA systems (Sect. 6).[1]

[1] In order to assure reproducibility, codes, dataset partitions, and trained models are made available at the project's GitHub repository: https://github.com/C4AI/Pira/tree/main/Triggering.

## 2 Background

Answer triggering is the task of deciding whether a question should be answered or not. By allowing some questions not to have an answer, models are required to learn *when* they should answer a given question. In Sect. 2.1, we briefly present the main architectures for QA systems, such as rule-based, end-to-end, and modular approaches. Section 2.2 reviews the datasets available for answering triggering and related tasks. Section 2.3 describes the two datasets used in our experiments, SQuAD 2.0 and Pirá.

### 2.1 Question Answering Systems

QA systems can have many different architectures. Before the popularization of deep learning, dialogue systems usually employed a mixture of rule-based approaches and feature-based classifiers, often intertwined in complex architectures, as in the Watson DeepQA system [7].

More recently, neural approaches came to dominate the field, at least for research purposes. Popular among these are end-to-end systems, such as T5 [17], and decoder (e.g., BERT [5]) and encoder-based models (e.g., GPT [3,16]), adapted for question answering tasks. There are also QA systems that combine different specialized mechanisms within the same training process. It is the case, for example, of RAG (Retrieval-Augmented Generation) [13], which uses a neural retriever, DPR [10], and a language generator, BART [12].

Finally, modular systems combine independent models that execute specific functions in the QA pipeline, without end-to-end training. DrQA [4], for instance, has a document retriever, based on bigram hashing and TF-IDF matching, and a document reader, which uses an achitecture of bidirectional LSTMs.

In this paper, we study a modular architecture for QA systems with two components, which we refer to as dual system. It consists of a filtering model that assess which questions should be answered and an answerer model that responds only to the selected questions.

### 2.2 Answer Triggering

Answer triggering was first defined by Yang et al. [26], together with a purposefully-developed dataset, WikiQA. Questions in WikiQA were based on Bing query logs, and the summary of the associated Wikipedia pages were used to determine if questions had an answer or not. SeqAL [8] was also based on Wikipedia, but using a larger number of questions from more domains and using the full entry pages. Answer triggering datasets grew out in popularity with SQuAD 2.0 [18], which added a large number of unanswerable questions to the original reading compreenhsion SQuAD dataset [19]. More importantly, unanswerable questions in SQuAD 2.0 were deliberately produced to have putative candidate answers. Although not a strict answer triggering dataset, Pirá [2] brings a number of human assessment on question meaningfulness that has a similar shape.

It is not always clear whether a question can be answered or not. Questions can be ambiguous or poorly structured, and whether or not a question has an answer may lay on a gray area, depending on contextual information that is not readily available. Some reading comprehension datasets try to overcome this difficulty by including a third outcome signaling uncertainty. ReCO [24] is a dataset of opinion-based queries in Chinese which uses three candidate answers for annotation: a positive one like `Yes`, a negative one like `No`, and an undefined one in case the question cannot be answered with the given documents. QuAIL [21] is a multi-choice dataset developed to include three degrees of certainty: answerable questions (given a context), unanswerable questions (even with context and world knowledge), and partially certain questions (when a good guess can be made).

### 2.3 Datasets

Two datasets are used in our experiments, SQuAD 2.0 and Pirá. Table 1 depicts their main statistics.[2]

**Table 1.** Number of QA sets for different splits of SQuAD 2.0 and Pirá.

| Model | # QA instance (%) | | | |
|---|---|---|---|---|
| | Train | Validation | Test | Total |
| SQuAD 2.0 | 130319 (91.65%) | 5936 (4.17%) | 5937 (4.17%) | 142192 (100%) |
| Pirá | 1755 (80.28%) | 215 (9.83%) | 216 (9.88%) | 2186 (100%) |

***SQuAD 2.0.*** SQuAD 2.0 is an extractive reading comprehension dataset. It combines the original SQuAD dataset [19], a reading comprehension resource with 100K+ questions, with approximately 53K unanswerable questions (marked as empty strings). To produce the unanswerable questions, annotators were asked to create questions over paragraphs that could not be correctly answered from these texts only. To avoid that questions unrelated to the context were created, annotators were instructed to produce questions that were relevant to the context and which contained plausible answers in the paragraph (such as the name of a person for a Who-type question). These plausible answers serve as effective distractors to questions, making it harder to realize what questions are in fact unanswerable. In total, SQuAD 2.0 contains around 151K questions, of which approximately 35.5% do not have answers. Another aspect of the dataset is that answerable questions present multiple answers, made by different annotators; for our experiments, we use only the first answer for each question as the ground

[2] In Pirá, only QA sets with meaningful evaluations were used. For the original dataset, the numbers would be: train: 1896 (79.98%), validation: 225 (9.96%), test: 227 (10%), total: 2258 (100%).

truth answer. The test set for this dataset is not publicly available; we thus break the original validation set into two equally-sized partitions to get our validation and test sets (which explains the smaller number of QA sets described in Table 1 as compared to the total number of QA sets of SQuAD 2.0).

***Pirá.*** Pirá [2] is a bilingual question answering dataset on the ocean, the Brazilian coast, and climate change. QA sets in Pirá are based on two corpora of supporting texts: one composed of scientific abstracts on the Brazilian Coast, and the other of small excerpts of two United Nations reports on the ocean. The dataset generation process comprised an assessment phase in which QA sets were manually evaluated in a number of aspects. Among these evaluations, participants were asked as to whether the QA sets were meaningful, based on a Likert scale (1 - Strongly disagree, 2 - Disagree, 3 - Neither agree nor disagree, 4 - Agree, 5 - Strongly agree). Figure 1 displays the distribution of QA instances in the test set by level of question meaningfulness. We also use the human validation answers produced in the assessments phase for our second experiment. Pirá contains 2248 QA sets in total. To conduct the experiments described in this paper, the dataset was splitted into three random partitions: training (80%), validation (10%), and test (10%) sets.

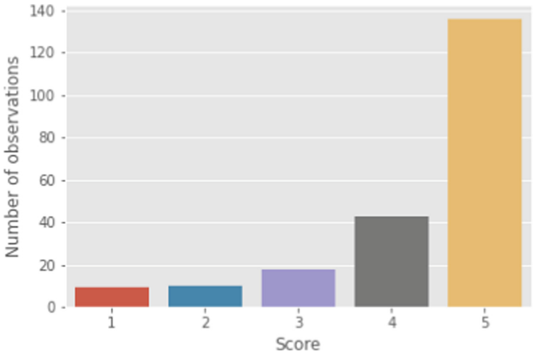


**Fig. 1.** Number of QA sets in the test set of Pirá by question meaningfulness level.

As regards the experiments conducted in this paper, SQuAD 2.0 and Pirá differ in important aspects. First, SQuAD 2.0 is an answering triggering dataset, meaning that some questions in it may have an answer while others may not, whereas in Pirá all questions are associated with answers. In addition, the former uses a qualitative ordinal (binary) variable, while the latter uses a numerical discrete variable (1–5). Finally, SQuAD 2.0 is concerned with answerability whereas Pirá is concerned with question meaningfulness. Those differences will be relevant to the experiments described above.

## 3 Answer Triggering

As a first contribution, we are interested in testing the usefulness of filtering questions based on answerability. As our answerer system, we use a DistilBERT-base model [22], fine-tuned on SQuAD 1.1 (a dataset where all questions have answers); context and question are concatenated and used as input. Three different scenarios are compared to measure the effects of question filtering on answer quality:

- questions from SQuAD 2.0 are passed indistinctly to the DistilBERT answerer system;
- questions are first grouped by answerability labels based on ground truth classifications and then passed to the answerer system; and
- questions are first grouped by answerability labels based on a classification model and then passed to the answerer system.

These three possibilities are illustrated in Fig. 2.

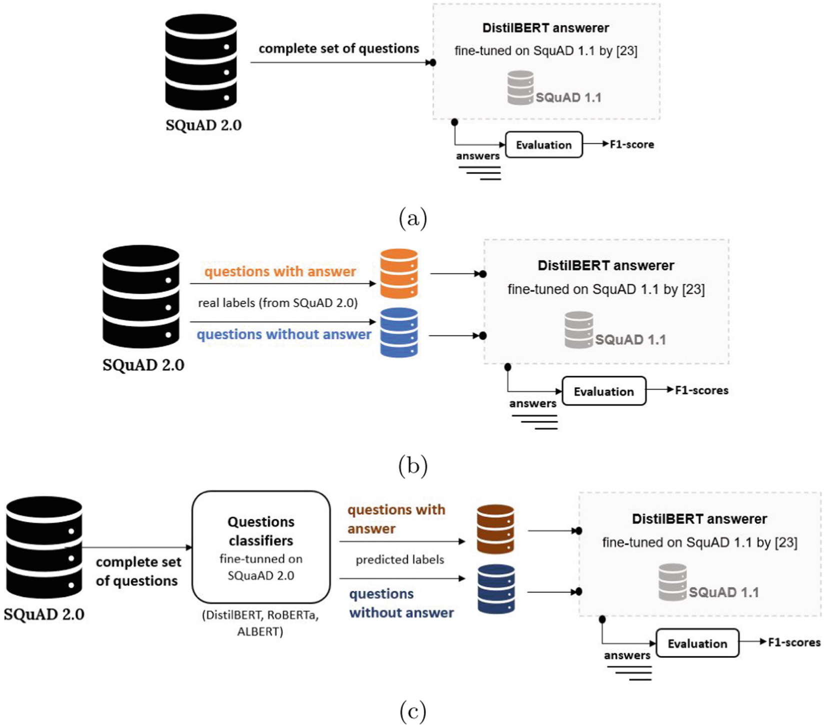
When generated answers are compared to the actual, manually-created ones, the QA system achieves a F1-score of 38.70 in the full test set (Fig. 2a).[3] Table 2 brings the results for this and the following tests. Furthermore, when only answerable questions are selected (based on real labels), there is a considerable increase in the quality of answers, with the F1-score going up to 78.25, a gain of 102.19% (Fig. 2b). This difference is due to the attempt, in the first case, to answer questions that have no answer whatsoever; when restricted solely to questions that do have answers, the average quality of answers improves.

**Table 2.** F1-score for a DistilBERT answerer model fine-tuned on SQuAD 1.1, when applied to the test set of SQuAD 2.0; the F1-score (0–100) is obtained by comparing the predicted answers to the (first) ground truth answers. Real labels are obtained from the annotated SQuAD 2.0 dataset (where unanswerable questions are presented as empty strings); predicted labels are obtained from three classification models: DistilBERT, RoBERTa, and ALBERT, all fine-tuned on SQuAD 2.0. Results for real and predicted labels are divided by answerable (Answ.) and unanswerable (Unansw.) questions. The F1-score of an empty answer (i.e., an unanswerable question) is by definition 0. In bold, the best result for answerable questions based on the predicted labels.

| Model | Total | Real label | | Predicted label | |
|---|---|---|---|---|---|
| | | Answ. | Unansw. | Answ. | Unansw. |
| DistilBERT | 38.70 | 78.25 | 0 | 39.42 | 18.53 |
| RoBERTa | | | | 46.08 | 33.13 |
| **ALBERT** | | | | **63.54** | **6.81** |

Hence, a filtering procedure that was able to filter answerable questions could in theory improve the quality of the answers generated by the answerer module.

[3] F1-score is implemented with the official SQuAD script. Available at: https://rajpurkar.github.io/SQuAD-explorer/.



(a)

(b)

(c)

**Fig. 2.** Different tests performed for assessing the relation between answerability and answer quality. From top to bottom: (a) questions are passed indistinctly to the answerer model; (b) questions are first grouped by answerability labels based on ground truth labels before going through the answerer model; and (c) questions are first grouped by answerability labels obtained from a classifier before going through the answerer model. The answers generated with the answerer system are then compared to the actual answers from SQuAD 2.0 using F1-score as an agreement metric.

In order to test this hypothesis, we perform a classification task on SQuAD 2.0's test set for question answerability. Three transformer models are used for the task: DistilBERT [22], RoBERTa [15], and ALBERT [11]; all of them fine-tuned on SQuAD 2.0 (since we want a model that already knows some information regarding answerability). A concatenation of context and question is used as input. All models are trained for 8 epochs in the train set, and the best log is chosen based on the validation set. Accuracy and F1-score for the three classifiers, as well as the number of predicted labels, are shown in Table 3.

After training the classifiers, we predict answerability labels for the questions in the test set of SQuAD 2.0 based on each of the models. Next, we calculate the F1-score for answerable or unanswerable separately, according to the predicted labels (Fig. 2c). As can be seen from Table 2, there is a considerable difference

**Table 3. Left**: Distribution of predicted labels for DistilBERT, RoBERTa, ALBERT, and ground truth in the test set (Answ. = Answerable, Unansw. = Unanswerable). **Right**: Accuracy and F1-score for DistilBERT, RoBERTa, ALBERT in the answer triggering task. Best result in bold.

| Model | # Predicted labels | | Results | |
|---|---|---|---|---|
| | Answ. | Unansw. | Accuracy | F1-score |
| DistilBERT | 5731 (96.53%) | 206 (3.47%) | 51.03 | 37.90 |
| RoBERTa | 2553 (43.00%) | 3384 (57.00%) | 57.84 | 57.59 |
| **ALBERT** | **3337 (56.21%)** | **2600 (43.79%)** | **84.08** | **84.03** |
| Test set | 3001 (50.55%) | 2936 (49.45%) | | |

in the quality of answers when answerability is taken into account.[4] When only questions that are predicted to have an answer are considered, the F1-score goes up to 39.42 (+1.87%) with DistilBERT, to 46.08 (19.07+%) with RoBERTa, and to 63.54 with ALBERT (64.19+%); as regards to questions predicted as having no answer, F1-score goes down to 18.53 (−52.12%), 33.13 (−14.39%), and 6.18 (−82.38%), respectively. The unintuitive fact that RoBERTa achieves better results than DistilBERT in both answerable and unanswerable questions is explained by a base rate problem: the DistilBERT model classifies the majority of questions as answerable (96.53%), whereas the other two models classify only 43.00% and 56.21% of the questions as answerable, respectively; a statistics closer to actual percentage of answerable questions in the test set (50.55%).

## 4 Continuous Answering Triggering

Traditionally, answering triggering is understood as in our previous experiment: a binary classification task on whether a question can be answered or not. In reality, though, the answerability of a question is often a complex affair. Although many questions can be undoubtedly categorized as having an answer or not, others lay on the middle part of the spectrum.

Therefore, as a second contribution, we explore whether more fine-grained information on questions can help to achieve better QA systems. For this task, we work with the English part of Pirá, a reading comprehension dataset on the ocean and the Brazilian coast (cf. Sect. 2.3). Contrary to answer triggering datasets *tout court* (cf. Sect. 2.2), Pirá does not possess explicit features indicating unswerability or certainty degrees. Instead, we use its manual evaluations on question meaningfulness as proxies for answerability: questions with a low value for meaningfulness are treated as having a lower degree of answerability.

[4] Both the classifiers described in this section and the regressors trained in the next use random initializations that may resul in slightly different predictions. To ensure the consistency of results, we repeated the same experiment 10 times each. The results described here are, therefore, representative of the trained models.

The Likert scale (1–5) scores used in the assessments provide a detailed level of analysis of answerability, more so than a binary or three-point alternative; even better, it permits us to treat question meaningfulness as an ordinal variable. Based on that, we reframe our answer triggering problem as a regression task, in which we aim to predict the degree of meaningfulness of a question in a 1 to 5 scale.

To see whether our meaningfulness regressor is indeed useful, we pair it again with a QA system. In theory, a QA system that only answers high-quality questions should give superior answers overall. Thus, we conduct a number of experiments to measure the quality of the QA system when questions are filtered by their degree of meaningfulness. First, we fine-tune a DistilBERT, ALBERT, and RoBERTa models on Pirá, using a concatenation of question and context as input. Models were trained for a total of 8 epochs in the training set, and evaluations were performed on the validation test. Similarly to the previous experiment, the ALBERT-based approach achieved the smallest loss and Root Mean Square Error in the validation set; for this reason, it was chosen as our regressor system. For the QA systems, we use two DistilBERT models fine-tuned for question answering on SQuAD 1.1 and SQuAD 2.0, respectively.

After training the regressor models, we predict the degree of meaningfulness for the questions in the test set. Questions are then grouped in 10 progressively smaller partitions, based on the predictions made with the regressor. For our first test, we measure the quality of the answers given by the QA system for these different partitions. As smaller partitions select questions evaluated better according to our regressor, we expect that the answers generated for them to be comparatively better. To measure the quality of answers, we rely again on F1-score; the original answer in Pirá serves as the ground truth.

Figure 3(a) shows the results for both QA models. The two graphs exhibit a similar trend. In both cases, F1-score goes up when smaller partitions of the test set feed the QA system. In this as in other tests, results for both QA models are similar—something expected, given that the F1-score between the answers generated with these two models is 70.03. Furthermore, as in the other tests discussed above, the F1-score for SQuAD 1.1 is usually higher than for SQuAD 2.0. Finally, the high spike in the last partitions may not be as significant as it appears; rather, it is likely due to randomness, given its small size, with only 21 observations.

F1-score is a metric based on the presence of the same tokens in the ground and predicted answers. It is well known, however, that automatic metrics may not correlate well with human evaluations [14]. There is also a more straightforward shortcoming of F1-score: this metric is not able to identify subtler similarity phenomena between answers, involving paraphrasing. For this reason, we decided to check the similarity between predicted and ground truth answers based on vector representations. Embeddings for both the original and the generated answer were produced with Sentence-BERT [20], and cosine dissimilarity (1 - cosine distance) served as a measure of semantic similarity (rather than word similarity, as in the case of F1-score).

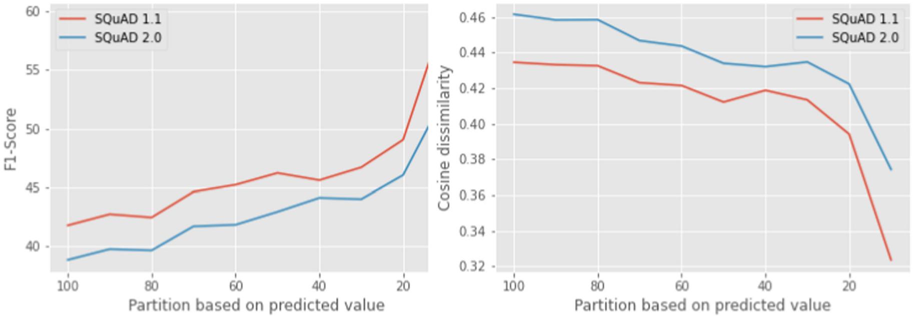


**Fig. 3. Left:** F1-score for different partitions of the test set (based on the predicted values of a RoBERTa regressor), comparing the ground truth answers in Pirá with answers generated by two DiltilBERT models fine-tuned on SQuAD 1.1 and SQuAD 2.0, respectively. **Right:** osine dissimilarity for different partitions of the test set (based on the predicted values of a RoBERTa regressor), comparing the ground truth answers in Pirá with answers generated by two DiltilBERT models fine-tuned on SQuAD 1.1 and SQuAD 2.0, respectively.

Figure 3(b) shows the results for both QA models. Answer dissimilarity, as measured by the cosine distance of sentence embeddings, tends to fall for smaller partitions; in other words, answers get better when more meaningful questions are selected, similarly to what was observed for F1-score. As was the case for F1-score, SQuAD 1.1 performs better than SQuAD 2.0.

The results obtained here are considerably worse than the results of the original SQuAD 1.1 and SQuAD 2.0 datasets. As reported in the original papers, a F1-score of 51 can be achieved with a strong logistic regression model for SQuAD 1.1 [19], and neural models can achieve a F1-score of 86 on SQuAD 1.1 and 66 on SQuAD 2.0 [18]. What explains the worse results for Pirá 2.0? Part of the gap may be explained by the non-extractive nature of its answers and the technical nature of the supporting texts. Another hypothesis is that our regressor is unable to detect levels of answerability. In order to test this possibility, we used the ground truth evaluations of question meaningfulness from Pirá to partition the dataset, instead of the predictions based on the regressor. Although this information is never present in real applications, this analysis may point to shortcomings derived from our prediction process. Figure 4(a) shows the results in F1-score for both QA models, for each level of meaningfulness.

It seems counter-intuitive that lower quality questions (according to the annotators' evaluations) achieve higher F1-scores. One reason for that may be a problem with human evaluations, perhaps due to a flawed instruction process. Nonetheless, we must also consider the small number of examples of low meaningful questions (cf. Fig. 1). One of the reasons our regressor may perform badly is because our regression model did not have access to many questions with a low degree of meaningfulness. For the larger group of questions evaluated from 3 to 5, F1-score seems to work more or less as expected.

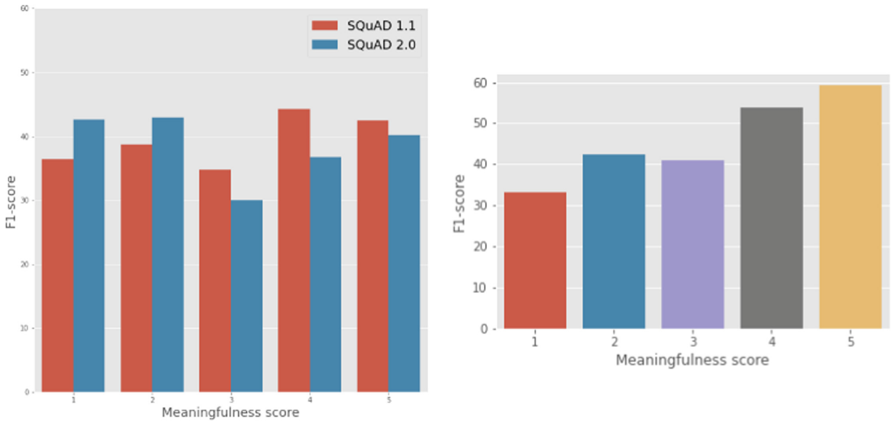


**Fig. 4. Left:** Average F1-score for each level of question meaningfulness (based on ground truth evaluations), comparing the original answers from Pirá and answers generated by two DistilBERT answerer models fine-tuned on SQuAD 1.1 and SQuAD 2.0, respectively. **Right:** Average F1-score for each level of question meaningfulness (based on ground truth evaluations), comparing the original and validation answers from Pirá.

As a final test, we evaluated whether part of the explanation for the low F1-score we got was not also caused by limitations in answerer models. In order to assess that, we tested the agreement, measured by F1-score, between answers given in the QA creation phase and those given in the evaluation phase. Again, we used real assessments of question meaningfulness as the criterion for partitioning the test set. Figure 4(b) presents the results for this test. A few facts can be observed. First, F1-score is considerably higher here than in previous tests, in which the generated answers were taken into account. Second, except for a slight oscillation for intermediary values, the F1-score demonstrates a consistent behavior, achieving larger values as the level of meaningfulness grows.

## 5 Discussion

The two experiments run in this work have shown that filtering out inadequate questions leads to better QA systems. Developing models that decide which questions to answer can help to achieve greater control over the quality of the answers. Results, however, were less than ideal, as the two trained models—the classifier and the regressor—were only relatively successful in selecting which questions should be passed to the answerer system. Therefore, if a dual system is to be implemented, better filtering models are highly needed.

For our first experiment on classification, the difficulty seems to derive from the nature of SQuAD 2.0, which has purposely-generated plausible unanswered questions. Unswerability appears to involve subtle elements that are not always captured by language models restricted to word correlations. More generally, common causes that affect the performance of answer triggering systems are related to syntactic bias (paying more attention to the structure of an answer than to its content); the presence of irrelevant information within the target answer; and lexical ambiguities [1].

As for the second experiment with Pirá, the limitations seem to be caused by a non-systematic annotation process. That feeling is strengthened by qualitative analysis of the dataset. In particular, when evaluating a question's meaningfulness, a number of aspects were often conflated by annotators: grammaticality, answerability, and context.

Our filtering system has been focused on finding low-quality questions, understood as those questions that cannot be correctly answered or that are ill-formulated. This filtering process, however, could be extended to a number of other situations. In particular, developing question filtering models is a real necessity for high-risk AI systems, since sensitive contexts demand that QA models avoid answering some questions. More importantly, these systems require a strict control over answer certainty. A separate module for filtering questions may, thus, provide a modular and inspectable tool.

Finally, as the small number of datasets available show, more answer triggering resources are needed. Particularly valuable would be datasets containing fine-grained annotations beyond simple answerability (binary) labels, such as annotations focused on the sources of low-quality in questions—e.g., grammar issues, lack of contextual information, ambiguity. Furthermore, it is important to develop filtering systems that can classify questions with respect to features other than accuracy, such as language toxicity [25], or based on conversational attributes, such as sensibleness, specificity, and interestingness [23].

## 6 Conclusion

In this paper, we have explored whether filtering models can contribute to control the quality of answers generated by answerer models. Two experiments were conducted: a classification task, with the aim of finding answerable questions; and a regression task, in which we predicted the degree of meaningfulness of questions. For both tasks, results showed a correlation between the ability to filter some questions and the quality of the answers generated.

The analysis conducted in this paper is a first step in the attempt to investigate modular approaches to QA systems and, in particular, on detached models that can select questions and control answer quality. In the future, we wish to expand this investigation in a number of directions. First, we want to develop better QA filtering systems, perhaps with the employment of specifically-developed architectures. Second, we want to train these models in multiple datasets, hoping that information from multiple sources can assist the task of discovering more general clues on answerability. Finally, we intend to explore filtering system where the task is to avoid answering sensitive questions and where high confidence in responses is needed.

## References

1. Acheampong, K.N., Tian, W., Sifah, E.B., Opuni-Boachie, K.O.-A.: The emergence, advancement and future of textual answer triggering. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) SAI 2020. AISC, vol. 1229, pp. 674–693. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52246-9_50
2. Paschoal, A.F.A., et al.: Pirá: a bilingual portuguese-english dataset for question-answering about the ocean. In: 30th ACM International Conference on Information and Knowledge Management (CIKM 2021) (2021). https://doi.org/10.1145/3459637.3482012
3. Brown, T.B., et al.: Language models are few-shot learners. CoRR abs/2005.14165 (2020). https://arxiv.org/abs/2005.14165
4. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading Wikipedia to answer open-domain questions. In: Barzilay, R., Kan, M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, 30 July–4 August, Volume 1: Long Papers, pp. 1870–1879. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1171
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
6. European-Commission: Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN#footnote8
7. Ferrucci, D.A.: Introduction to "this is watson". IBM J. Res. Dev. **56**(3), 1 (2012). https://doi.org/10.1147/JRD.2012.2184356
8. Jurczyk, T., Zhai, M., Choi, J.D.: SelQA: a new benchmark for selection-based question answering. In: 28th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2016, San Jose, CA, USA, 6–8 November 2016, pp. 820–827. IEEE Computer Society (2016). https://doi.org/10.1109/ICTAI.2016.0128
9. Kadavath, S., et al.: Language models (mostly) know what they know (2022). https://arxiv.org/abs/2207.05221
10. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 16–20 November 2020, pp. 6769–6781. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.550
11. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020. OpenReview.net (2020). https://openreview.net/forum?id=H1eA7AEtvS
12. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020, pp.

7871–7880. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.703

13. Lewis, P.S.H., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6–12 December 2020, virtual (2020)
14. Liu, C., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., Pineau, J.: How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: Su, J., Carreras, X., Duh, K. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, 1–4 November 2016, pp. 2122–2132. The Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/d16-1230
15. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692
16. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
17. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020). http://jmlr.org/papers/v21/20-074.html
18. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: unanswerable questions for squad. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018, Volume 2: Short Papers, pp. 784–789. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/P18-2124
19. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100, 000+ questions for machine comprehension of text. In: Su, J., Carreras, X., Duh, K. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP, Austin, Texas, USA, 1–4 November 2016, pp. 2383–2392. The Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/d16-1264
20. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, 3–7 November 2019, pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1410
21. Rogers, A., Kovaleva, O., Downey, M., Rumshisky, A.: Getting closer to AI complete question answering: a set of prerequisite real tasks. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020, pp. 8722–8731. AAAI Press (2020). https://ojs.aaai.org/index.php/AAAI/article/view/6398
22. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108 (2019). http://arxiv.org/abs/1910.01108
23. Thoppilan, R., et al.: LaMDA: language models for dialog applications (2022)
24. Wang, B., Yao, T., Zhang, Q., Xu, J., Wang, X.: ReCO: a large scale Chinese reading comprehension dataset on opinion. CoRR abs/2006.12146 (2020). https://arxiv.org/abs/2006.12146

25. Welbl, J., et al.: Challenges in detoxifying language models. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2447–2469. Association for Computational Linguistics, November 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.210
26. Yang, Y., Yih, W.T., Meek, C.: WikiQA: a challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2013–2018. Association for Computational Linguistics, Lisbon, September 2015. https://doi.org/10.18653/v1/D15-1237