Proceedings of the OSACT 2022 Workshop @LREC2022, pages 136–145
Marseille, 20 June 2022

SMASH at Qur’an QA 2022: Creating Better Faithful Data Splits for Low-resourced Question Answering Scenarios

Amr Keleg and Walid Magdy
School of Informatics, University of Edinburgh
Edinburgh, UK
a.keleg@sms.ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

The Qur’an QA 2022 shared task aims at assessing the possibility of building systems that can extract answers to religious questions given relevant passages from the Holy Qur’an. This paper describes SMASH’s system that was used to participate in this shared task. Our experiments reveal a data leakage issue among the different splits of the dataset. This leakage problem hinders the reliability of using the models’ performance on the development dataset as a proxy for the ability of the models to generalize to new unseen samples. After creating better faithful splits from the original dataset, the basic strategy of fine-tuning a language model pretrained on classical Arabic text yielded the best performance on the new evaluation split. The results achieved by the model suggests that the small scale dataset is not enough to fine-tune large transformer-based language models in a way that generalizes well. Conversely, we believe that further attention could be paid to the type of questions that are being used to train the models given the sensitivity of the data.

Keywords: Question Answering, Reading Comprehension Question Answering, Arabic NLP

1. Introduction

Automatic Question Answering (QA) task is gaining increased attention in recent years. The task aims at building models that can provide answers to various user-generated questions by utilizing a large set of curated documents. The type of understanding and reasoning required to answer these questions automatically is challenging. Open-domain QA aims at extracting answers using knowledge graphs and information retrieval systems, or generating answers using large pre-trained transformer-based architectures (Chen and Yih, 2020). Conversely, Reading Comprehension QA (RCQA) aims at extracting a span from a specified passage as the answer to a question. Training models for RCQA generally depends on building large scale datasets of question (Q), answer (A), passage (P) triples in which the answer (A) is a span of contiguous text extracted from the passage (P). For example, the SQUAD dataset was built by crowdsourcing more than 100k triples of Question/Answer/Passage where human annotators were asked to pose questions, and extract their answers from 536 English Wikipedia articles (Rajpurkar et al., 2016). The availability of such large datasets allows researchers to train models that can better generalize to unseen questions, thus advancing the field of Question Answering.

The Qur’an QA 2022 shared-task is another example of the RCQA tasks (Malhas et al., 2022). The Qur’an QA task provides 1,337 triples of questions, passages, and their answers (Malhas et al., 2022). In addition to the small size of the dataset, having questions written in Modern Standard Arabic (MSA), and passages written in Classical Arabic (CA) makes it more challenging. For instance, questions such as ما هي الإشارات للدماغ أو لأجزاء من الدّماغ في القرآن؟, and هل أشار القرآن الى نقص الأكسجين في المرتفعات؟ contain lexical terms that are not part of CA.

In this paper, we provide the system description of the SMASH[1] team of the University of Edinburgh in the task. We built a system for answering questions given passages from the holy Qur’an and make our code available for public[2]. In addition, we provide our general thoughts on the task and potential suggestions for improvements. Our team achieved a pRR score of 0.4004 in the official submitted run to the task.

The remaining sections of the paper are organized as follows: §2 shows the steps we took to create better faithful splits of the dataset, §3 describes the model’s architecture, §4 reports the results achieved by the different model variants that were tested, and finally §5 gives some directions that might help with building better models.

2. Data Preparation

Rogers et al. (2020) argues that dataset creators need to provide an additional annotation for each triple to indicate the type of reasoning needed in order to answer the question. Providing such taxonomy allows for analyzing the performance of models trained using the dataset. Depending on a single aggregated metric for evaluating a model will neglect the fact that the model might be poorly performing on some question types, given that the datasets might be skewed in a way such that some question types are over-represented, while other types are under-represented.

Automatic categorization of question types: starting from the aforementioned recommendation, we classified the questions in the Qur’an QA dataset according to the interrogative article that appears in the question as a proxy for the type of reasoning needed to answer these

[1] `https://smash.inf.ed.ac.uk/`

[2] We release the code through: `https://github.com/AMR-KELEG/SMASH-QuranQA`

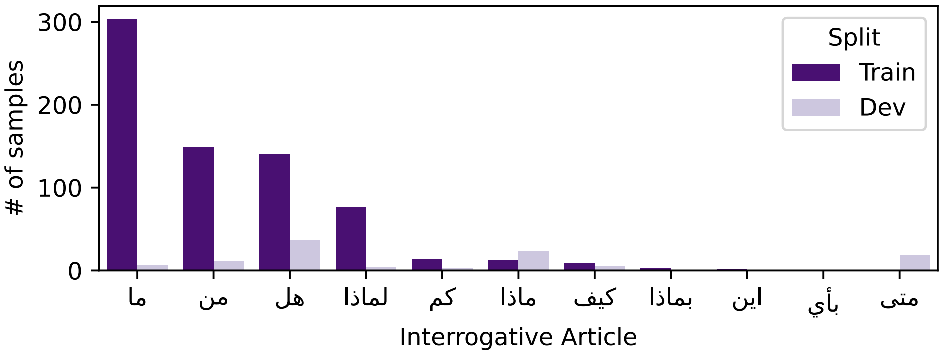

Figure 1: The distribution of question types among the train and development splits.

Shared Answer	Shared Passage	Question (Dev)	Question (Train)
من شاء فليؤمن ومن شاء فليكفر	وقل الحق من ربكم فمن شاء فليؤمن ومن شاء فليكفر إنا أعتدنا للظالمين نارا أحاط بهم سرادقها وإن يستغيثوا يغاثوا بماء كالمهل يشوي الوجوه بئس الشراب وساءت مرتفقا. إن الذين آمنوا وعملوا الصالحات إنا لا نضيع أجر من أحسن عملا. أولئك لهم جنات عدن تجري من تحتهم الأنهار يحلون فيها من أساور من ذهب ويلبسون ثيابا خضرا من سندس وإستبرق متكئين فيها على الأرائك نعم الثواب وحسنت مرتفقا.	هل سمح الإسلام بحرية الاعتقاد بالدخول إلى الإسلام؟	اتهم القرآن بأنه السبب في الدكتاتورية الإسلامية لكونه أباح التكفير وقتال الكفار حتى يسلموا، كيف نرد على ذلك؟
أطيعوا الرسول لعلكم ترحمون	وعد الله الذين آمنوا منكم وعملوا الصالحات ليستخلفنهم في الأرض كما استخلف الذين من قبلهم وليمكنن لهم دينهم الذي ارتضى لهم وليبدلنهم من بعد خوفهم أمنا يعبدونني لا يشركون بي شيئا ومن كفر بعد ذلك فأولئك هم الفاسقون. وأقيموا الصلاة وآتوا الزكاة وأطيعوا الرسول لعلكم ترحمون. لا تحسبن الذين كفروا معجزين في الأرض ومأواهم النار ولبئس المصير.	هل هناك إسلام بدون الحديث الشريف؟	لماذا لا يكتفي المسلمون بالقرآن الكريم ويلجأون للسنة أيضاً؟
ما كان قولهم إلا أن قالوا ربنا اغفر لنا ذنوبنا وإسرافنا في أمرنا وثبت أقدامنا وانصرنا على القوم الكافرين	وما كان لنفس أن تموت إلا بإذن الله كتابا مؤجلا ومن يرد ثواب الدنيا نؤته منها ومن يرد ثواب الآخرة نؤته منها وسنجزي الشاكرين. وكأين من نبي قاتل معه ربيون كثير فما وهنوا لما أصابهم في سبيل الله وما ضعفوا وما استكانوا والله يحب الصابرين. وما كان قولهم إلا أن قالوا ربنا اغفر لنا ذنوبنا وإسرافنا في أمرنا وثبت أقدامنا وانصرنا على القوم الكافرين. فآتاهم الله ثواب الدنيا وحسن ثواب الآخرة والله يحب المحسنين.	ماذا يشمل الإحسان؟	من هم المحسنون؟
قل يا عباد الذين آمنوا اتقوا ربكم	أمن هو قانت آناء الليل ساجدا وقائما يحذر الآخرة ويرجو رحمة ربه قل هل يستوي الذين يعلمون والذين لا يعلمون إنما يتذكر أولو الألباب. قل يا عباد الذين آمنوا اتقوا ربكم للذين أحسنوا في هذه الدنيا حسنة وأرض الله واسعة إنما يوفى الصابرون أجرهم بغير حساب.	ماذا يشمل الإحسان؟	من هم المحسنون؟

Table 1: Examples of data leakage in the original training and development splits. Leakage is sometimes caused by having paraphrased questions in the two splits that refer to the same passages, and have the same answer span.

questions.First, we manually compiled a list of twelve interrogative articles by investigating their occurrences in the training and development dataset splits. These articles are لماذا، كيف، بأي، بماذا، من، كم، كيف، ما، ماذا، هل، اين، متى. Since the word من can be either a preposition or an interrogative article, it is discarded in case the question contains another interrogative article. In case the question contains more than one interrogative article, the one occurring first is selected as the question type. An "NA" article is used in case a question contains none of the interrogative articles listed above. Figure 1 shows how the distribution of questions types is different between the train and development splits. It is also clear that polar interrogative questions (هل), what questions (ماذا), and when questions (متي) are frequent in the development split.

While fine-tuning an Arabic BERT model to extract the answer span (as described in §3), we noticed that the model's performance on what (ماذا) and polar interrogative questions (هل) is much better than the other question types. Our initial interpretation was that the model is able to reason about these types of questions in a way that generalizes to new unseen questions in the development dataset split. However, manually investigating the data showed that most of the questions of these types in the development dataset are paraphrases of questions in the training dataset that have exactly the same answer spans (A) from the same passages (P). Table 1 provides some examples demonstrating the issue of having questions in the development split that are mere paraphrases of other questions in the training split. These questions refer to the same passage, are paraphrases of each others, and thus have the exact same answer span. Questions belonging to these two question types represent 60% of the development dataset. A model that overfits the training data can achieve high scores by just generating the same answer span for a passage that was seen in the training data without doing any kind of reasoning. Consequently, a high performance on the provided development set might be misleading which does not help in reaching a robust optimized model for the task. On the other hand, we noticed that 17.5% of the development dataset belongs to the question type 'when متى' that is not part of the training dataset. One can argue that it is nearly impossible to expect the model to generalize to question types that were not encountered in the training data. A similar issue was flagged by Lewis et al. (2021) as they showed that there exists an overlap between the training and testing splits of multiple open-domain QA datasets, which in turn affects the ability of these datasets to be used as benchmarks for the various QA models.

Building faithful splits: based on limitations discussed above in the original training and the development datasets, we decided to generate new training and development splits by concatenating both the original training and development splits then dividing the dataset into four mutually exclusive datasets: (1) context in domain with leakage (i.e. training and development set share questions that have the same passage and answer) $D_{(1)\ in+leakage}$, (2) context in domain without leakage $D_{(2)\ in+no\ leakage}$, (3) hard out of domain $D_{(3)\ ood+hard}$, and (4) easy out of domain $D_{(4)\ ood+easy}$. First, $D_{(1)\ in+leakage}$ is formed by selecting all samples having repeated (passage, answer) pairs or (question, answer) pairs[3]. This type represents samples that the model can memorize, and thus are not useful for evaluating the generalization abilities of the model. $D_{(2)\ in+no\ leakage}$ contains samples of repeated passages that are not part of $D_{(1)\ in+leakage}$. All the remaining samples have unique passages (i.e.: passages that occur only once in the concatenated datasets).

Split name	Train	Dev	Total
$D_{(1)in+leakage}$	119	181	300
$D_{(2)in+no\ leakage}$	209	32	241
$D_{(3)ood+hard}$	0	60	60
$D_{(4)ood+easy}$	189	29	218
Total	517	302	819

Table 2: Number of triples in the new faithful splits of the dataset.

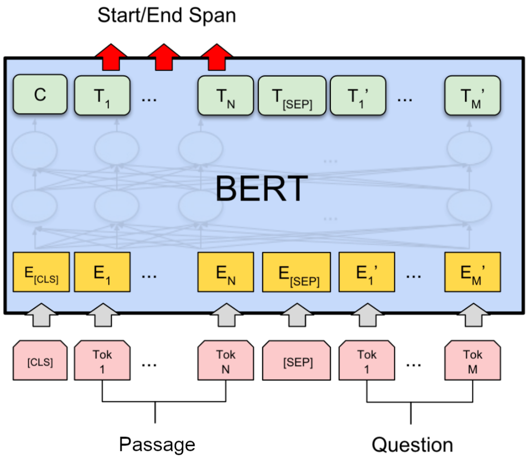

Figure 2: Using BERT for Reading Comprehension Question Answering (Devlin et al., 2019). In our implementation, the passage (P) followed by the question (Q) is fed to the BERT model.

These samples are split by assigning the ones having their question appearing three times or less to $D_{(3)\ ood+hard}$. These questions are rare, and thus the model should find it tricky to overfit them. Lastly, samples of unique passages which have their question appearing four or more times are assigned to $D_{(4)\ ood+easy}$.

After categorizing the samples, the four datasets were split into training, and development splits as follows. For $D_{(2)\ in+no\ leakage}$, and $D_{(4)\ ood+easy}$, each dataset was randomly shuffled and split into training and development datasets using 86.7/13.3 splitting percentages which are the same ratios used in the original dataset[4]. For $D_{(1)\ in+leakage}$, only one question is kept for each repeated (question, answer) and (passage, answer) pairs in the training dataset. The remaining samples of $D_{(1)\ in+leakage}$ are added to the development split. On the other hand, the whole $D_{(3)\ ood+hard}$ is used as a development split in order to measure the ability of the model to reason about the different question types without relying on memorizing answers that are frequent for specific questions or specific passages. Table 2 shows the number of the training and development samples

[3] We use full-match to consider two strings to be matching, which implies that for instance the following answers ناقة, الناقة are not considered to be repeated.

[4] The number of samples in the development dataset of each type is rounded to the nearest integer. The remaining samples are used as the training dataset

within the new faithful splits[5].

Detecting overfitting using the new faithful splits: We hypothesize that splitting the original dataset into four splits provides a proxy for predicting whether a model is still learning, or is just overfitting the training examples. Knowing that models need large number of samples to operate effectively, training samples from the four splits are compiled, and used as a whole to train/fine-tune models. Irrespective of the model being deployed, we think that a model which is overfitting the samples of the compiled training data would have the following performance trend on the development samples of the different splits (sorted in a descending order):

- $\mathbf{D}_{\mathbf{(1)in+leakage}}$: Samples within this split either share the same passage-answer pair or the same question-answer with one sample of the compiled training dataset. An overfitting model might be tempted to yield high probabilities for the answer's tokens, irrespective of the passage or the question.
- $\mathbf{D}_{\mathbf{(4)ood+easy}}$: While passages within this split are unique and are not encountered in the training data, the fact that the questions are previously encountered in the dataset, would in some cases imply that the answers needed to be extracted from this unique passage have lexical overlap with the answers of the same questions within the training data.
- $\mathbf{D}_{\mathbf{(3)ood+hard}}$: The fact that both the questions and passages are unique implies that the model would need to achieve good generalization in order to be able to properly answer these questions.
- $\mathbf{D}_{\mathbf{(2)in+no\ leakage}}$: This split is the most challenging one, since the model has encountered the passages within this split in the training dataset, however the new questions imply that the model should be able to understand the question in order to generate the right answer instead of just recalling the answer of the question in the training dataset.

3. Model Architecture

3.1. Model used in our submission

Given the success of BERT models with the RCQA task (Devlin et al., 2019), we fine-tuned the CAMELBERT-CA (Inoue et al., 2021) to predict the span of the answer given both the passage and the questions as an input to the model separated by the special $[SEP]$ token as shown in Figure 2. While there is a large number of avaialbe Arabic BERT models that showed their quality peformance on several tasks (Farha and Magdy, 2021), we decided to use the CAMELBERT-CA model for our task since it is pretrained on the OpenITI dataset, which is a large curated corpus of books written in Classical Arabic, including the Holy Quran (Nigst et al., 2020). The fact that the model's pretraining corpus contains text written in Classical Arabic makes it more suitable to the Qur'an QA task.

After tokenizing the input into subwords, the model $Vanilla + CA$ is fine-tuned to independently predict the probability that the answer span starts at each subword and the probability that the answer span ends at each subword. In inference time, a simple greedy decoding method is used to predict the right answer span. More specifically, the subwords at which the answer span begins/ends are these subwords having the highest start/end probabilities, respectively. In case the answer span is invalid (i.e.: the subword at which the span ends precedes the subword at which it starts), then the answer span is considered to start at the start index and end at the last subword of the passage (i.e.: the subword just before the $[SEP]$ token). The model was fine-tuned for 16 epochs, while saving checkpoints of the weights at the end of each epoch. An Adam optimizer was used with a learning rate set to 10^{-5}, beta values set to 0.9, and 0.98, and batches of 32 passage|question sequences having a maximum length of 512 subwords[6]. A single NVIDIA Quadro-RTX-8000 GPU of 48 GB VRAM was used to fine-tune the models. In order to simplify the fine-tuning process, no hyper-parameter tuning was performed. Since the task allows providing five different answers to each sample and being motivated by keeping our solution simple, the whole passage was used as the second ranked trivial answer.

3.2. Abandoned experiments

As an attempt to improve the performance of the fine-tuned BERT-based model, we tested different independent tricks that did not manage to improve the performance of the basic fine-tuned CAMELBERT-CA model ($Vanilla + CA$).

Using a pretrained MSA model ($Vanilla + MSA$) Since the questions are written in MSA and given that CAMELBERT-MSA is pretrained on larger data than CAMELBERT-CA, then fine-tuning CAMELBERT-MSA might give the model a better understanding of the questions, and thus better reasoning.

Embedding Named Entities (NER) Motivated by the fact that some questions are about the prophets, angels and other religious named entities, the intuition was that giving the model an extra signal might help in extracting spans that are related to these entities. First, a list of named-entities was compiled from multiple Wikipedia pages related to the Qur'an[7]. Then, a learnable entity embedding

[5] Examples of triples from the four splits are listed in the Appendix.

[6] Padding was used in case a sequence is shorter. Only two sequences in the training/development data had longer lengths than 512 subwords, and these sequences were truncated to the maximum length of 512 subwords.

[7] https://ar.wikipedia.org/wiki/قائمة_أشخاص_ذكروا_في_الكتاب_المقدس_والقرآن,

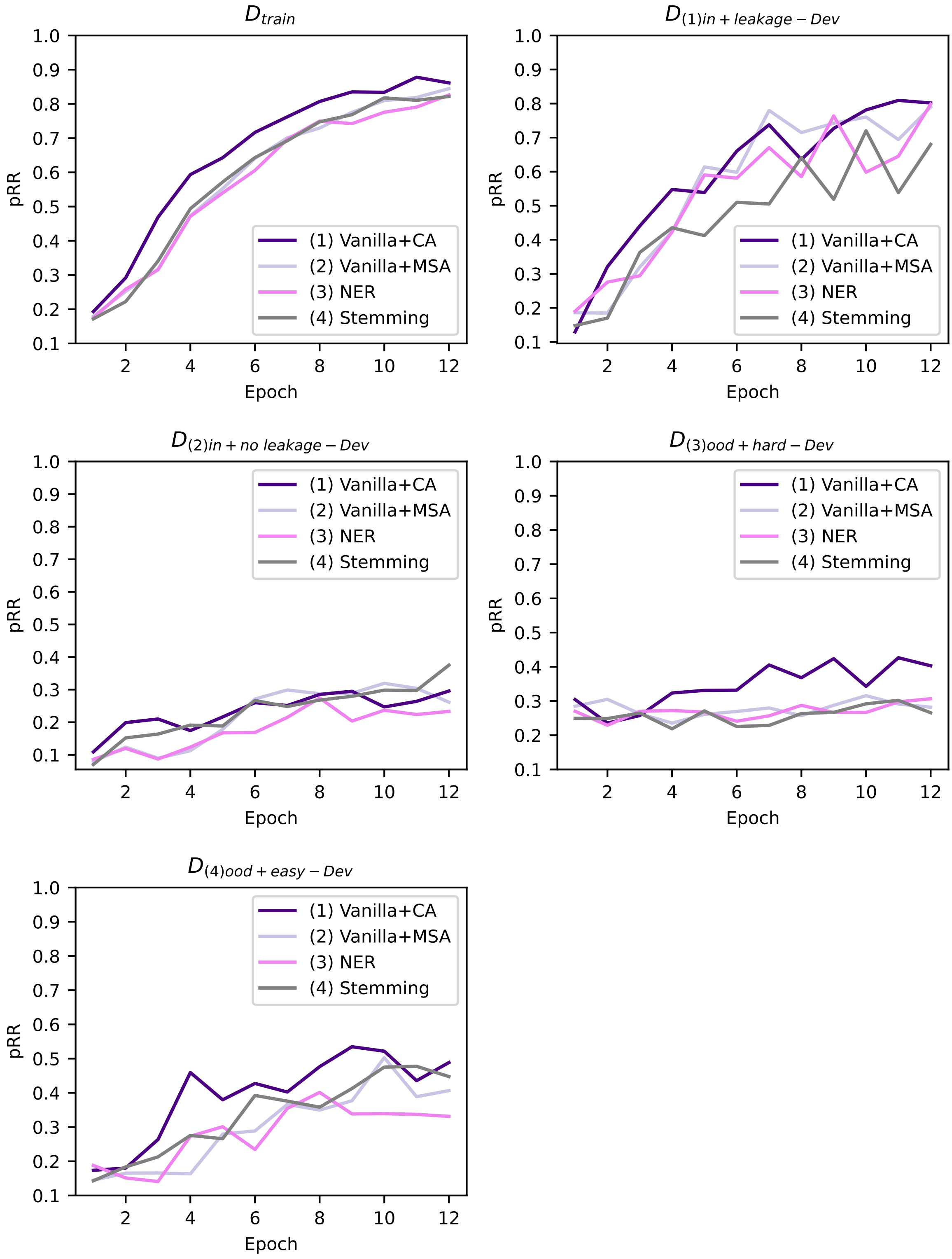

Figure 3: pRR scores computed at the end of each training epoch on the different data splits. D_{train} refers to the compiled training sets from all the data splits. This compiled dataset is used to fine-tune the models. On the other hand, models are evaluated on the development samples of each split independently.

was added to the input embeddings of the tokens that are among the list of compiled named-entities. The entity embedding is a parameter similar to positional embeddings used to encode the position of subtokens in the input sentence.

https://ar.wikipedia.org/wiki/حيوانات_ذكرت_في_القرآن, https://ar.wikipedia.org/wiki/شخصيات_ذكرت_في_القرآن

Stemming the text (*$Stemming$*) As a way to reduce the morphological diversity of Arabic tokens that might hinder the model's ability to find connections between the question and the passage, Farasa (Abdelali et al., 2016) was used to stem the text before feeding it into the CAMELBERT-CA. In order to be able to extract a span from the raw passage, mapping needed to be done between the span of tokens before and after stemming. It is worth mentioning that stemming does not affect the number of tokens.

More complex answer span decoding methods We tested with decoding multiple answer spans based on the start/end probabilities of the passage subtokens. Our empirical results showed that just depending on the start/end probabilities is not enough to rank the decoded spans. Therefore, we opted to using the greedy decoding method mentioned at the beginning of this section.

4. Results

Using the new faithful splits of the dataset described in §2, the four variants of the model (namely $(1)Vanilla + CA$, $(2)Vanilla + MSA$, $(3)NER$, and $(4)Stemming$) were fine-tuned on the compiled training samples of all the four splits. Figure 3 shows pRR scores computed at the end of each fine-tuning epoch. The model is typically evaluated on the compiled set of training samples in addition to the development samples of each of the four splits.

Analyzing the models' performance: Looking at the pRR scores, we observe that the four variants of the model perform better on samples that are similar to the ones found in the training dataset. More specifically, the scores on the $D_{(1)in+leakage}$ split are nearly on par with these achieved on the compiled set of training samples D_{train}. While the scores on the other three splits are lower than these achieved on the $D_{(1)in+leakage}$ split, the models are show higher performance on the $D_{(4)ood+easy}$ split in compared to the harder $D_{(3)ood+hard}$, and $D_{(2)in+no\ leakage}$ splits. Given that the number of evaluation samples in the $D_{(3)ood+hard}$ split is nearly double these in the $D_{(2)in+no\ leakage}$ split, the $D_{(3)ood+hard}$ split was used to guide the selection of the best performing variant. Surprisingly, the vanilla CAMELBERT-CA model ($Vanilla + CA$) fine-tuned on the training samples outperform all the other variants by an absolute difference of nearly 0.1. This can be attributed to the fact that the training set is too small for the models to generalize to new samples. Consequently, using more complex models might be not beneficial since it is hard to tune the parameters using such small dataset.

Error analysis on the official testing dataset: The pRR scores achieved by the model indicate that it is unable to reason about all the questions in the testing dataset. Manual inspection of the answers extracted from passages within the testing data reveals that the model sometimes tend to ignore the question, and extract the same answer span for specific passages that occurred in the training dataset. For example, the model predicts the same answer span of the following question in the training datasets كيف اهلك الله قوم عاد؟ when it is fed the same passage with a different question كيف اهلك الله قوم ثمود؟.

Assessing the stability of the models: Reimers and Gurevych (2017) showed how deep learning models are susceptible to achieve unstable performances when spurious parameters such as the random seed are changed. Therefore, we are reporting the distribution of the pRR scores achieved by the four model variants when evaluated on the $D_{(1)in+leakage}$, and $D_{(3)ood+hard}$ splits in Figure 4. Focusing our attention first on the performance of the models on the $D_{(3)ood+hard}$ split reveals the superior performance achieved by the $Vanilla + CA$ model in compared to all the other model variants. Moreover, the box plots show a non-negligible variance in the pRR achieved by the models as an effect of just changing the random seed. Conversely, the pRR scores achieved on the $D_{(1)in+leakage}$ split yields different conclusions, with having the $Vanilla + MSA$ model outperforming the $Vanilla + CA$ one. We think that these contradicting observations demonstrate the harmful impact that data leakage might have in case design decisions are blindly taken based on the performance on the development split without investigating the samples represented within the split, and comparing them to the ones in the training split.

Results of our submissions: Following these observations, we made two submissions based on the $Vanilla + CA$ model in which the model was fine-tuned using 1, and 3 as the random seeds respectively. The pRR scores reported in Table 3 indicate that the official results on the hidden test set, and these achieved on the $D_{(3)ood+hard}$ split are similar to a great extent, which in turn might mean that the $D_{(3)ood+hard}$ split can be reliably used to evaluate the ability of the model to generalize to unseen data.

5. Going Further

Given how models are sensitive to the samples within the training dataset, we believe that researchers interested in extending the dataset should be mindful of the types of questions and the passages of these new samples. A potential way of preventing the models from overfitting would be to have multiple questions of different types referring to different spans from the same passage. Doing so might prevent the models from extracting the same answer span for a passage while ignoring the question. This strategy is employed in large RCQA datasets such as SQUAD, where more than 100,000 questions were generated from 23,215 paragraphs extracted from only 536 Wikipedia articles (i.e.: The average number of questions for each paragraph is 4.31) (Rajpurkar et al., 2016). On the other hand, the training split of the QRCD dataset

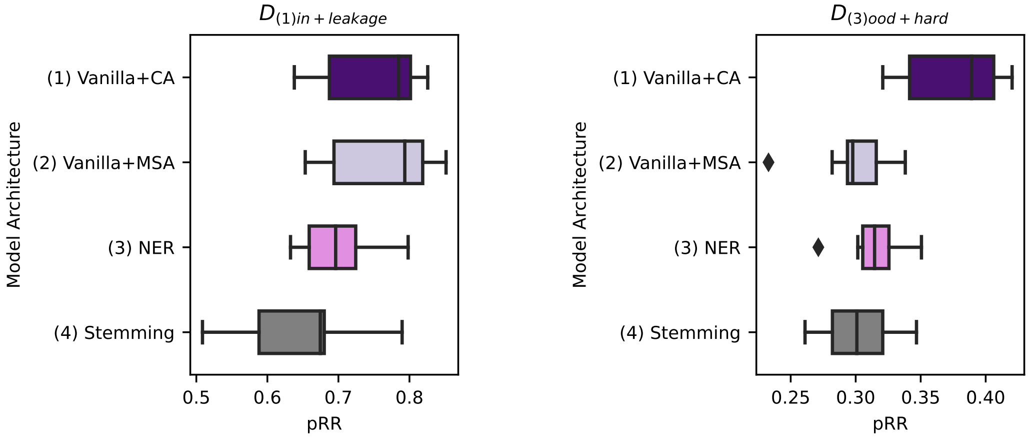

Figure 4: Distribution of pRR scores on the development samples of the $D_{(1)in+leakage}$, and $D_{(3)ood+hard}$ splits for 10 different random seeds of each architecture.

Model name	Official pRR score on the hidden test set	pRR score on the $\mathbf{D_{(3)ood+hard}}$ Dev split
$Vanilla + CA_{seed=1}$	0.3801	0.4073
$Vanilla + CA_{seed=3}$	0.4004	0.4083

Table 3: pRR scores of the submitted systems on the hidden test set, and the $D_{(3)ood+hard}$ dataset

has 468 unique passages, and 710 questions (i.e.: The average number of questions for each passage is 1.51). It is worth mentioning that for SQUAD, the paragraphs were first compiled, and then annotators were asked to pose questions that are answered by spans within these paragraphs. Conversely, the AyaTEC dataset, from which the QRCD dataset is created, was built by first compiling a list of questions, and then searching for Quranic verses that answer such questions (Malhas and Elsayed, 2020). Moreover, providing annotations for the reasoning type required to answer the questions would be beneficial for analyzing the performance of the model on different types of questions and interpreting the behavior of these models.

Finally, and given the sensitivity of this task, it might be better to avoid questions that can have multiple interpretations, and would create unnecessary controversy. For example, "متى يحل الإسلام دم الشخص؟" is a question that appears 19 times in the original development dataset, where some answer spans are specific to some contexts, and can not be provided as an answer to this question in general. For instance, the following answer span "فإذا انسلخ الأشهر الحرم فاقتلوا المشركين حيث وجدتموهم وخذوهم واحصروهم واقعدوا لهم كل مرصد" that Muslims should defend themselves against a group that attack them. This does not in turn imply by any means that Islam urges Muslims to kill each and every person who participated in this war. Consequently, extracting the following answer span for the specified extreme question would be misleading. Given the diversity of topics in the Quran and the small size of the dataset, it might be better to train models to answer factoid questions. Answers to such questions do not depend on the context, and therefore will not cause any unnecessary controversy. This is also motivated by the fact that the answers extracted by models are not generally interpretable, and thus one will not be able to reason about why a model is behaving in a specific way.

6. Conclusion

Despite the advancements researchers have achieved in solving a diverse set of Natural Language Processing (NLP) tasks using the (pretrain then fine-tune) paradigm, our experiments show that using a relatively small-sized Reading Comprehension Question Answering (RCQA) dataset for fine-tuning a large pretrained language model is challenging, especially if we are aiming at having models that can generalize to different types of questions that require complex reasoning. Moreover, we indicate that the dataset used for fine-tuning the models might have a data leakage problem between the training and developments splits. This problem hinders the possibility of using the model's performance on the development set as a reliable proxy for the model's generalization abilities to new unseen samples.

7. Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

8. Bibliographical References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for

Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California, June. Association for Computational Linguistics.

Chen, D. and Yih, W.-t. (2020). Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Farha, I. A. and Magdy, W. (2021). Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 21–31.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics.

Lewis, P., Stenetorp, P., and Riedel, S. (2021). Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online, April. Association for Computational Linguistics.

Malhas, R. and Elsayed, T. (2020). AyaTEC. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19:1 – 21.

Malhas, R., Mansour, W., and Elsayed, T. (2022). Qur'an QA 2022: Overview of the first shared task on question answering over the holy Qur'an. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

Nigst, L., Romanov, M., Savant, S. B., Seydi, M., and Verkinderen, P. (2020). OpenITI: a Machine-Readable Corpus of Islamicate Texts, June.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September. Association for Computational Linguistics.

Rogers, A., Kovaleva, O., Downey, M., and Rumshisky, A. (2020). Getting closer to AI complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731, Apr.

Appendix

Tables 4, and 5 showcase some samples from the new faithful splits, along with the original splits of these examples. The first set of examples demonstrates how questions that share the same passages and answer spans are grouped together in the $D_{(1)in+leakage}$ split. The second set shows the difficulty of the $D_{(2)\ in+no\ leakage}$ split, since the model needs to extract an answer span from a passage that is used to answer another unrelated question. This is particularly hard in case the model has overfitted the training data in a sense that it generates the same answer span for the same passage irrespective of the question being asked. Examples of questions in $D_{(3)ood\ +\ hard}$ in Table 5 showcase the complexity of these rare questions that are not part of the training dataset. The model will need to have superior generalization in order to be able to have proper reasoning, and consequently answer these questions. The last example shows two questions referring to different passages yet having some lexical overlap. We think that large models such as BERT have the ability to consider وعمل صالحاً and وعملوا الصالحات are inflections of the same lexical items, and consequently will be to some extent able to extract the correct answer span for both cases even if one of them is not part of the training dataset.

Answer	Passage	Question	Original Split	New Split
إبراهيم وإسماعيل	وإذ ابتلى إبراهيم ربه بكلمات فأتمهن قال إني جاعلك للناس إماما قال ومن ذريتي قال لا ينال عهدي الظالمين. وإذ جعلنا البيت مثابة للناس وأمنا واتخذوا من مقام إبراهيم مصلى وعهدنا إلى إبراهيم وإسماعيل أن طهرا بيتي للطائفين والعاكفين والركع السجود. وإذ قال إبراهيم رب اجعل هذا بلدا آمنا وارزق أهله من الثمرات من آمن منهم بالله واليوم الآخر قال ومن كفر فأمتعه قليلا ثم أضطره إلى عذاب النار وبئس المصير. وإذ يرفع إبراهيم القواعد من البيت وإسماعيل ربنا تقبل منا إنك أنت السميع العليم. ربنا واجعلنا مسلمين لك ومن ذريتنا أمة مسلمة لك وأرنا مناسكنا وتب علينا إنك أنت التواب الرحيم. ربنا وابعث فيهم رسولا منهم يتلو عليهم آياتك ويعلمهم الكتاب والحكمة ويزكيهم إنك أنت العزيز الحكيم.	من هم الأنبياء الذين ذكروا في القرآن على أنهم مسلمون؟	Train	$D_{(1)in+leakage-Train}$
		من بنى الكعبة؟	Dev	$D_{(1)in+leakage-Dev}$
من اهتدى فلنفسه ومن ضل فإنما يضل عليها	ولئن سألتهم من خلق السماوات والأرض ليقولن الله قل أفرأيتم ما تدعون من دون الله إن أرادني الله بضر هل هن كاشفات ضره أو أرادني برحمة هل هن ممسكات رحمته قل حسبي الله عليه يتوكل المتوكلون. قل يا قوم اعملوا على مكانتكم إني عامل فسوف تعلمون. من يأتيه عذاب يخزيه ويحل عليه عذاب مقيم. إنا أنزلنا عليك الكتاب للناس بالحق فمن اهتدى فلنفسه ومن ضل فإنما يضل عليها وما أنت عليهم بوكيل.	اتهم القرآن بأنه السبب في الدكتاتورية الإسلامية لكونه أباح التكفير وقتال الكفار حتى يسلموا، كيف نرد على ذلك؟	Train	$D_{(1)in+leakage-Train}$
		ما هي الدلائل التي تشير بأن الانسان مخير؟	Dev	$D_{(1)in+leakage-Dev}$
		إن كان الله قدّر علي أفعالي فلماذا يحاسبني؟	Dev	$D_{(1)in+leakage-Dev}$
		لماذا سيُحاسب ويُعذب الضال يوم القيامة ان كان """من يضلل الله فما له من هاد"""" كما ورد من قوله تعالى في آية 32 و آية 63 من سورة الزمر؟	Dev	$D_{(1)in+leakage-Dev}$
		هل سمح الإسلام بحرية الاعتقاد بالدخول إلى الإسلام؟	Dev	$D_{(1)in+leakage-Dev}$
مدين	إذ أوحينا إلى أمك ما يوحى. أن اقذفيه في التابوت فاقذفيه في اليم فليلقه اليم بالساحل يأخذه عدو لي وعدو له وألقيت عليك محبة مني ولتصنع على عيني. إذ تمشي أختك فتقول هل أدلكم على من يكفله فرجعناك إلى أمك كي تقر عينها ولا تحزن وقتلت نفسا فنجيناك من الغم وفتناك فتونا فلبثت سنين في أهل مدين ثم جئت على قدر يا موسى. واصطنعتك لنفسي. اذهب أنت وأخوك بآياتي ولا تنيا في ذكري. اذهبا إلى فرعون إنه طغى. فقولا له قولا لينا لعله يتذكر أو يخشى. قالا ربنا إننا نخاف أن يفرط علينا أو أن يطغى. قال لا تخافا إنني معكما أسمع وأرى. فأتياه فقولا إنا رسولا ربك فأرسل معنا بني إسرائيل ولا تعذبهم قد جئناك بآية من ربك والسلام على من اتبع الهدى. إنا قد أوحي إلينا أن العذاب على من كذب وتولى.	ما هي اسماء المدن المذكورة في القرآن؟	Train	$D_{(1)in+leakage-Train}$
لعله يتذكر أو يخشى		ما هو أثر الكلام الطيب؟	Train	$D_{(2)\ in+no\ leakage-Train}$

Table 4: Examples of samples showing the characteristics of the new four faithful splits.

Answer	Passage	Question	Original Split	New Split
من يرد الله أن يهديه يشرح صدره للإسلام ومن يرد أن يضله يجعل صدره ضيقا حرجا كأنما يصعد في السماء	فمن يرد الله أن يهديه يشرح صدره للإسلام ومن يرد أن يضله يجعل صدره ضيقا حرجا كأنما يصعد في السماء كذلك يجعل الله الرجس على الذين لا يؤمنون. وهذا صراط ربك مستقيما قد فصلنا الآيات لقوم يذكرون. لهم دار السلام عند ربهم وهو وليهم بما كانوا يعملون.	هل أشار القرآن الى نقص الأكسجين في المرتفعات؟	Train	$D_{(3)ood\,+\,hard-Dev}$
ناصيتها	وإلى عاد أخاهم هودا قال يا قوم اعبدوا الله ما لكم من إله غيره إن أنتم إلا مفترون. يا قوم لا أسألكم عليه أجرا إن أجري إلا على الذي فطرني أفلا تعقلون. ويا قوم استغفروا ربكم ثم توبوا إليه يرسل السماء عليكم مدرارا ويزدكم قوة إلى قوتكم ولا تتولوا مجرمين. قالوا يا هود ما جئتنا ببينة وما نحن بتاركي آلهتنا عن قولك وما نحن لك بمؤمنين. إن نقول إلا اعتراك بعض آلهتنا بسوء قال إني أشهد الله واشهدوا أني بريء مما تشركون. من دونه فكيدوني جميعا ثم لا تنظرون. إني توكلت على الله ربي وربكم ما من دابة إلا هو آخذ بناصيتها إن ربي على صراط مستقيم. فإن تولوا فقد أبلغتكم ما أرسلت به إليكم ويستخلف ربي قوما غيركم ولا تضرونه شيئا إن ربي على كل شيء حفيظ. ولما جاء أمرنا نجينا هودا والذين آمنوا معه برحمة منا ونجيناهم من عذاب غليظ. وتلك عاد جحدوا بآيات ربهم وعصوا رسله واتبعوا أمر كل جبار عنيد. وأتبعوا في هذه الدنيا لعنة ويوم القيامة ألا إن عادا كفروا ربهم ألا بعدا لعاد قوم هود.	ما هي الإشارات للدماغ أو لأجزاء من الدّماغ في القرآن؟	Train	$D_{(3)ood\,+\,hard-Dev}$
من دعا إلى الله وعمل صالحا وقال إنني من المسلمين	ومن أحسن قولا ممن دعا إلى الله وعمل صالحا وقال إنني من المسلمين. ولا تستوي الحسنة ولا السيئة ادفع بالتي هي أحسن فإذا الذي بينك وبينه عداوة كأنه ولي حميم. وما يلقاها إلا الذين صبروا وما يلقاها إلا ذو حظ عظيم. وإما ينزغنك من الشيطان نزغ فاستعذ بالله إنه هو السميع العليم.	من هم المحسنون؟	Train	$D_{(1)in+leakage-Dev}$
الذين آمنوا وعملوا الصالحات	وقل الحق من ربكم فمن شاء فليؤمن ومن شاء فليكفر إنا أعتدنا للظالمين نارا أحاط بهم سرادقها وإن يستغيثوا يغاثوا بماء كالمهل يشوي الوجوه بئس الشراب وساءت مرتفقا. إن الذين آمنوا وعملوا الصالحات إنا لا نضيع أجر من أحسن عملا. أولئك لهم جنات عدن تجري من تحتهم الأنهار يحلون فيها من أساور من ذهب ويلبسون ثيابا خضرا من سندس وإستبرق متكئين فيها على الأرائك نعم الثواب وحسنت مرتفقا.	من هم المحسنون؟	Train	$D_{(4)in+leakage-Train}$

Table 5: Examples of samples showing the characteristics of the new four faithful splits.