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

Amr Keleg, Walid Magdy

a.keleg@sms.ed.ac.uk

Institute for Language, Cognition and Computation
University of Edinburgh

What makes QRCD special?

Dataset	Total number of questions	Number of unique paragraphs(passages)	Average no. questions per passage
SQUAD[1]	100,000	23,215	4.31

[1] SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al., EMNLP 2016)

What makes QRCD special?

Dataset	**Total number of questions**	**Number of unique paragraphs(passages)**	**Average no. questions per passage**
SQUAD[1]	100,000	23,215	4.31
QRCD (training split)	710	468	1.51

[1] SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al., EMNLP 2016)

What makes QRCD special?

Dataset	Total number of questions	Number of unique paragraphs(passages)	Average no. questions per passage
SQUAD[1]	100,000	23,215	4.31
QRCD (training split)	710	468	1.51

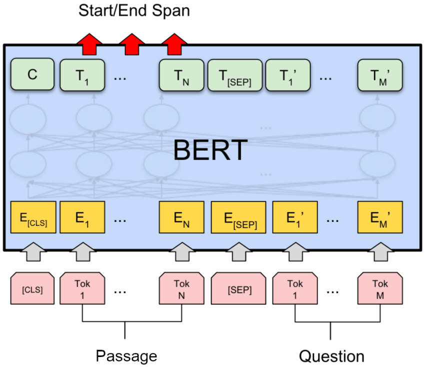
QRCD:

- Smaller size
- Less **number of questions per passage**

[1] SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al., EMNLP 2016)

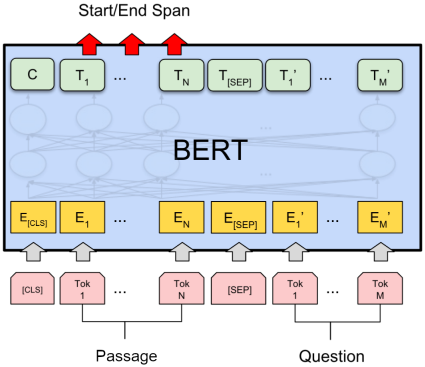
Typical model fine-tuning

Main architecture used:



Typical model fine-tuning

Main architecture used:



Models within our experiments

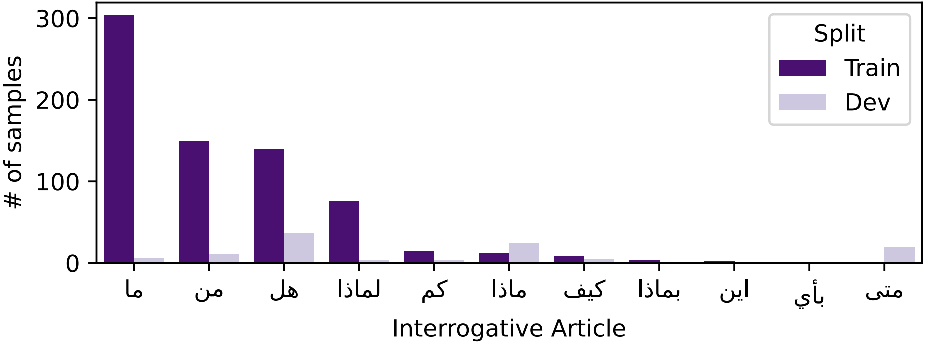
- **(1)Vanilla + CA**: Fine-tuning CAMELBERT-CA
- **(2)Vanilla + MSA**: Fine-tuning CAMELBERT-MSA
- **(3)NER**: Fine-tuning CAMELBERT-CA + augmenting the model with information about Quranic Named Entities (NEs)
- **(4)Stemming**: Fine-tuning CAMELBERT-CA over stemmed text using Farasa

Error Analysis using question types

- **Interrogative article** لماذا، كيف، بأي، بماذا، من، كم، كيف، ما، ماذا، هل، اين، متى as a proxy for **reasoning needed to be done**.

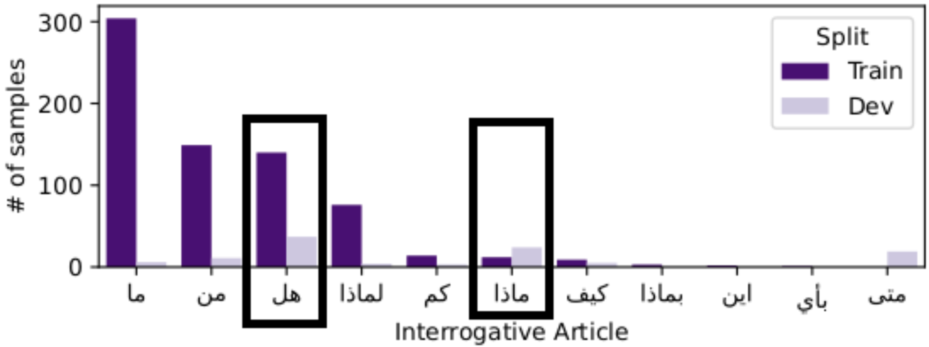
Error Analysis using question types

- **Interrogative article** لماذا، كيف، بأي، بماذا، من، كم، كيف، ما، ماذا، هل، اين، متى as a proxy for **reasoning needed to be done**.



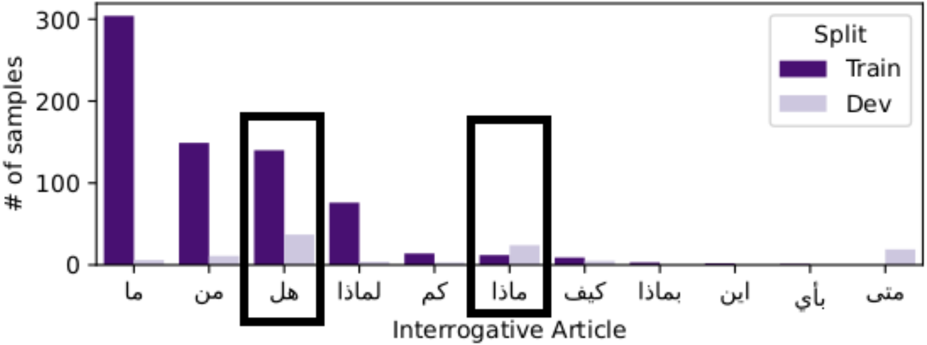
Error Analysis using question types

- **Interrogative article** لماذا، كيف، بأي، بماذا، من، كم، كيف، ما، ماذا، هل، اين، متى as a proxy for **reasoning needed to be done**.



Error Analysis using question types

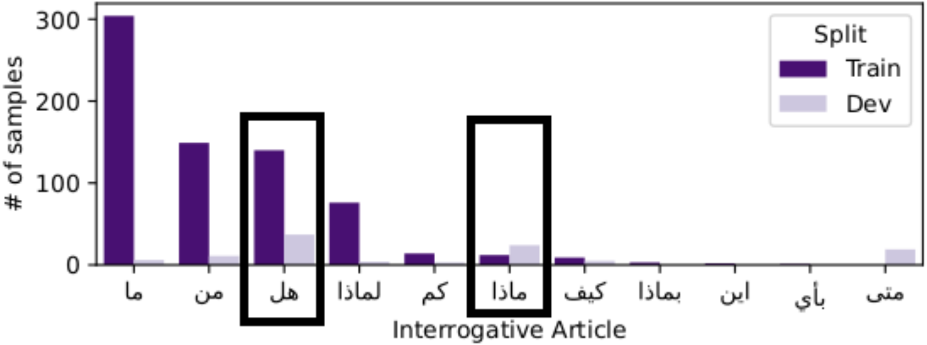
- **Interrogative article** لماذا، كيف، بأي، بماذا، من، كم، كيف، ما، ماذا، هل، اين، متى as a proxy for **reasoning needed to be done**.



- Noticed **high PRR scores** for questions having the article **هل - ماذا**
.Good generalization?

Error Analysis using question types

- **Interrogative article** لماذا، كيف، بأي، بماذا، من، كم، كيف، ما، ماذا، هل، اين، متى as a proxy for **reasoning needed to be done**.



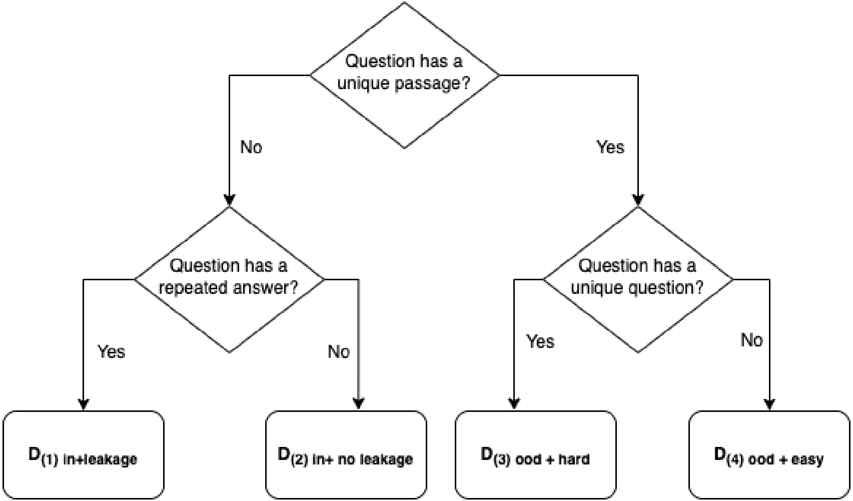
- Noticed **high PRR scores** for questions having the article **هل - ماذا** .Good generalization? Not really!

Examples of leakage between train/dev splits

Shared Answer	Shared Passage	Question (Dev)	Question (Train)
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؟

Resplitting the training/development splits

After concatenating the training and development splits:



Difficulty of generalization for different datasets

Dataset	Characteristics	Generalization
$\mathbf{D_{(1)in+leakage}}$	Repeated passage and repeated answer	Memorize the answer for this passage

Difficulty of generalization for different datasets

Dataset	**Characteristics**	**Generalization**
$\mathbf{D}_{\mathbf{(1)in+leakage}}$	Repeated passage and repeated answer	Memorize the answer for this passage
$\mathbf{D}_{\mathbf{(4)ood+easy}}$	Unique passage and repeated question	Possibly answer is similar to these of the repeated question (possibility of memorization)

Difficulty of generalization for different datasets

Dataset	Characteristics	Generalization
$\mathbf{D_{(1)in+leakage}}$	Repeated passage and repeated answer	Memorize the answer for this passage
$\mathbf{D_{(4)ood+easy}}$	Unique passage and repeated question	Possibly answer is similar to these of the repeated question (possibility of memorization)
$\mathbf{D_{(3)ood+hard}}$	Unique passage and unique question	Unseen questions that can test the true generalization of the model

Difficulty of generalization for different datasets

Dataset	Characteristics	Generalization
$\mathbf{D}_{\mathbf{(1)in+leakage}}$	Repeated passage and repeated answer	Memorize the answer for this passage
$\mathbf{D}_{\mathbf{(4)ood+easy}}$	Unique passage and repeated question	Possibly answer is similar to these of the repeated question (possibility of memorization)
$\mathbf{D}_{\mathbf{(3)ood+hard}}$	Unique passage and unique question	Unseen questions that can test the true generalization of the model
$\mathbf{D}_{\mathbf{(2)in+no\ leakage}}$	Repeated passage and unique question	If the model memorizes an answer for the passage then it will fail badly for such questions

New data splits

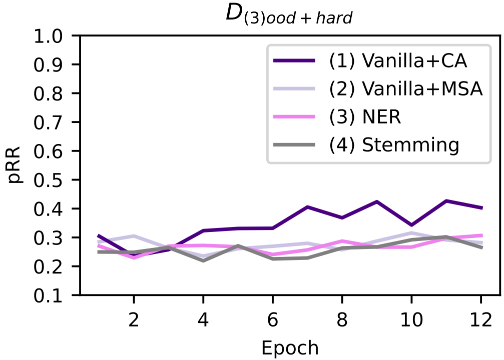
- Random split for $D_{(2)\textit{in+no leakage}}$, $D_{(4)\textit{ood+easy}}$.
- Keep only one instance from triples having repeated passage-answer within $D_{(1)\textit{in+leakage}}$.
- Use $D_{(3)\textit{ood+hard}}$ as development set only.

Results of fine-tuning models

Dataset	pRR scores on the development splits
$\mathbf{D}_{\mathbf{(1)in+leakage}}$	≈ 0.8
$\mathbf{D}_{\mathbf{(4)ood+easy}}$	≈ 0.5
$\mathbf{D}_{\mathbf{(3)ood+hard}}$	≈ 0.4
$\mathbf{D}_{\mathbf{(2)in+no\ leakage}}$	$\approx [0.3,0.4]$

Results of fine-tuning models

Dataset	pRR scores on the development splits
$\mathbf{D}_{\mathbf{(1)in+leakage}}$	≈ 0.8
$\mathbf{D}_{\mathbf{(4)ood+easy}}$	≈ 0.5
$\mathbf{D}_{\mathbf{(3)ood+hard}}$	≈ 0.4
$\mathbf{D}_{\mathbf{(2)in+no\ leakage}}$	$\approx [0.3,0.4]$



Comparison of results to the official ones

Stability of the models under different random seeds:

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

Reflections and going further

- Trying to minimize leakage between training and development splits might be better for fairly comparing models.

Reflections and going further

- Trying to minimize leakage between training and development splits might be better for fairly comparing models.
- Using models pretrained on MSA is crucial for questions such as هل أشار القرآن الى نقص الأكسجين في المرتفعات؟.

Reflections and going further

- Trying to minimize leakage between training and development splits might be better for fairly comparing models.
- Using models pretrained on MSA is crucial for questions such as هل أشار القرآن الى نقص الأكسجين في المرتفعات؟.
- If the way further is creating a **larger dataset** that covers more topics and/or question types:
    - Try to have more than one non-trivial question for each unique passage
      (easier said than done).

Reflections and going further

- Trying to minimize leakage between training and development splits might be better for fairly comparing models.
- Using models pretrained on MSA is crucial for questions such as هل أشار القرآن الى نقص الأكسجين في المرتفعات؟.
- If the way further is creating a **larger dataset** that covers more topics and/or question types:
  - Try to have more than one non-trivial question for each unique passage (easier said than done).
- Avoid questions that can have multiple interpretations (e.g.:“متى يحل الإسلام دم الشخص؟”).