Extending the Scope of Out-of-Domain: Examining QA models in multiple subdomains

Chenyang Lyu[1], Jennifer Foster[1], Yvette Graham[2]
[1]School of Computing, Dublin City University, Ireland
[2]School of Computer Science & Statistics, Trinity College Dublin, Ireland

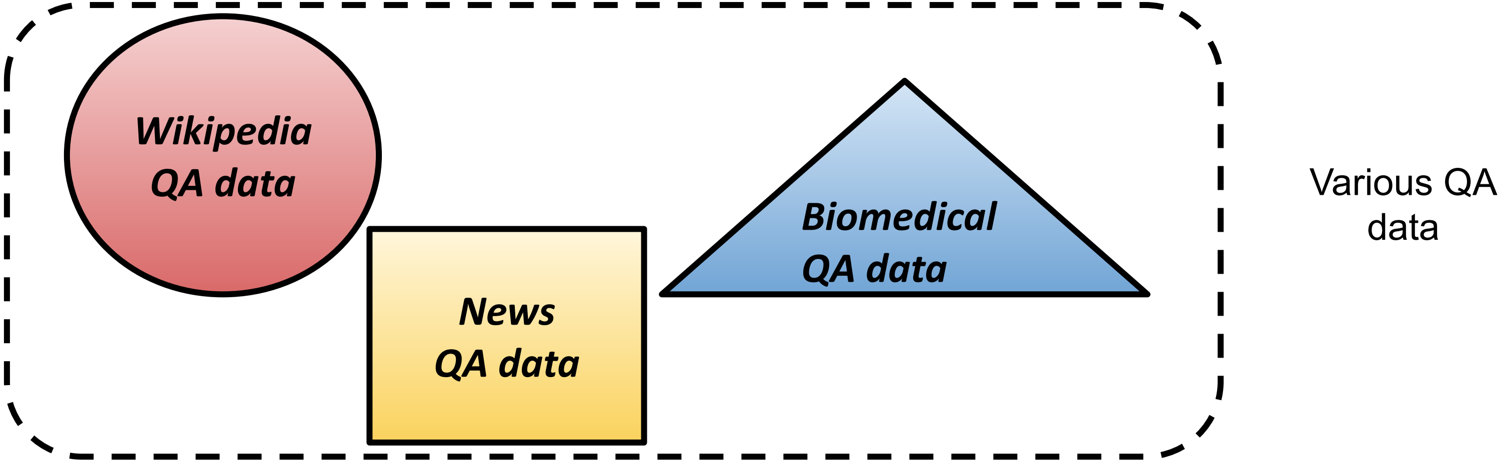
Workshop on Insights from Negative Results in NLP
ACL 2022

Outline

- Introduction
- Out-of-subdomain generalizability
- Experimental results
- Conclusion

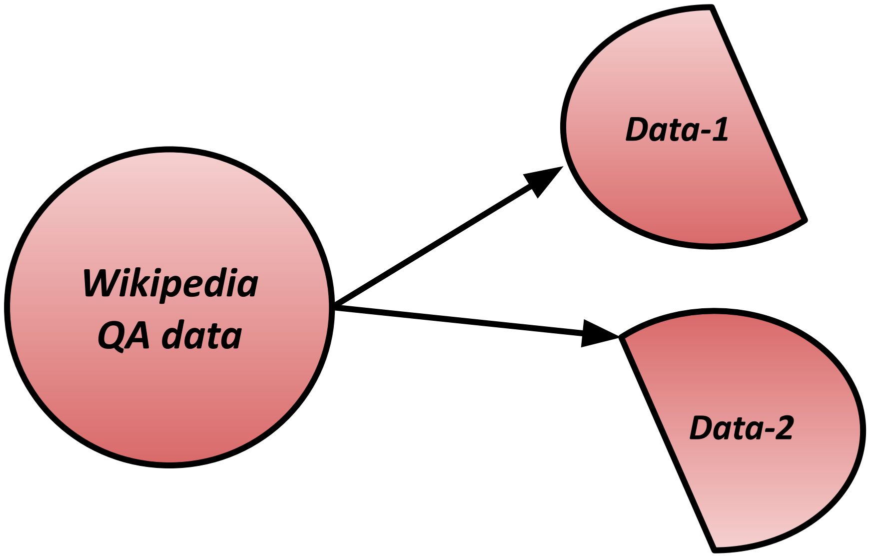
Introduction

- Out-of-domain performance of QA systems is strongly connected to their generalizability and robustness.
- Previous studies mostly focus on coarse-grained general domains (e.g. news domain, wikipedia domain)



Introduction

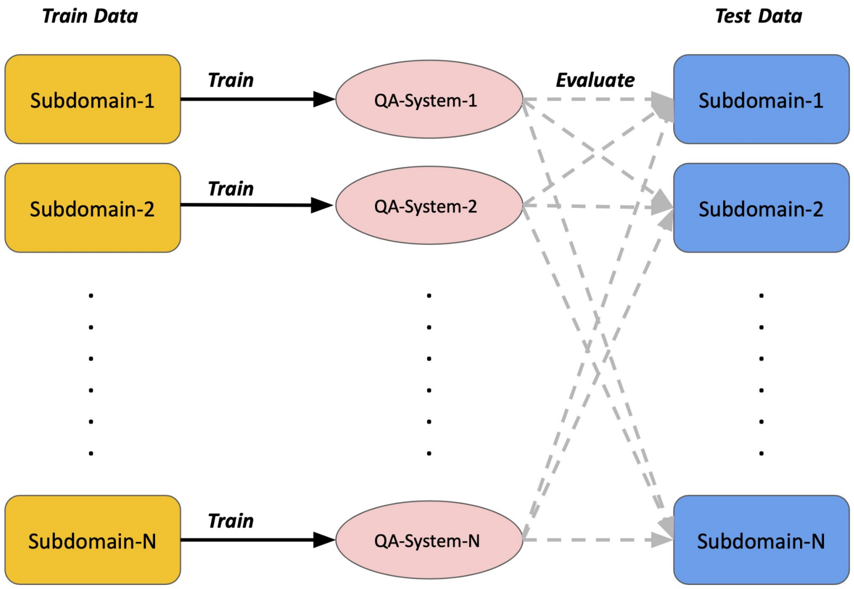
- But we cannot ignore the effect of the subdomains defined by the internal characteristics of QA datasets.



Subdomains could potentially introduce bias to QA systems, such as those defined by context-question overlap (Sen 2020).

Out-of-subdomain generalizability

- Can QA systems trained on a subdomain perform well on the other subdomains?



Out-of-subdomain generalizability

- We investigate three subdomains defined by the internal characteristics of QA datasets:
  - Question type
  - Text length (context, question, answer)
  - Answer position (character-level, word-level, sentence-level)

Experiments

- English Datasets:
  - QC (Li and Roth, 2002), a question classification dataset
  - SQuAD1.1(Rajpurkar et al., 2016), a wikipedia-based extractive QA dataset
  - NewsQA (Trischler et al., 2017), a news-based extractive QA dataset
- QA systems and question classification model:
  - BERT-base-uncased

Experiments

- Experiment methodology:
    - We split the dataset into subdomains, then train QA systems on each subdomain and evaluate them on all subdomains.
    - In the training process on each subdomain, we train QA systems using increasingly large subsets sampled from subdomain data.

If an internal characteristic is not a source of bias, then the performance of all QA systems trained on each subdomain should be the same or very close.

Experiment 1 - question type results

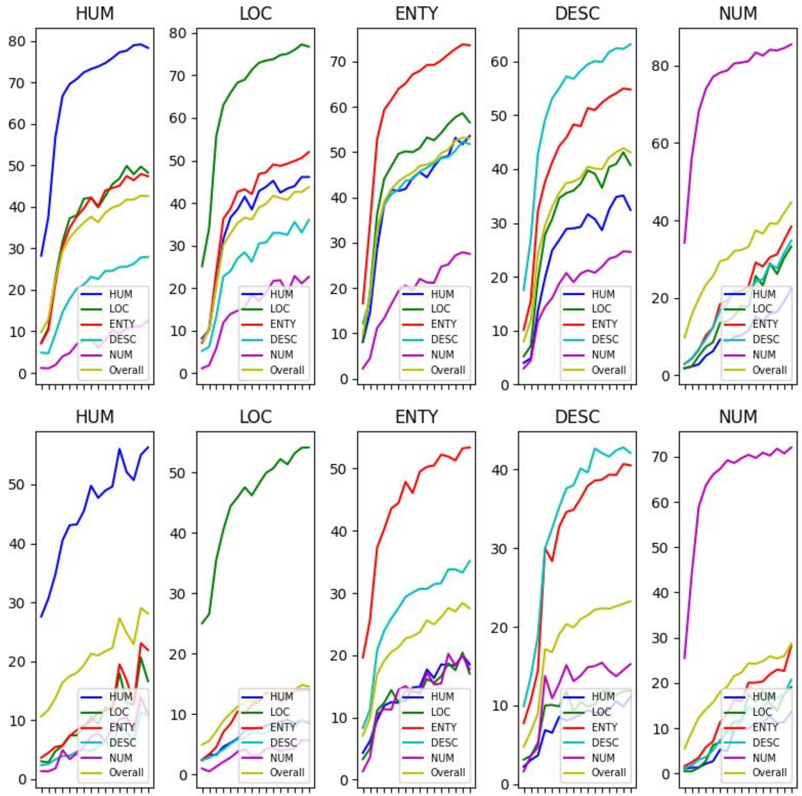
- We categorize all QA examples according to their question type, which has five classes including *HUM, LOC, ENTY, DESC, NUM*.

		LOC	ENTY	HUM	NUM	DESC
SQuAD1.1	Train set	11.4	27.6	20.7	24.5	15.5
	Dev set	10.5	27.6	21.0	23.0	17.4
NewsQA	Train set	11.4	16.9	30.0	18.8	22.6
	Dev set	12.3	16.9	32.2	17.8	20.5

The proportion (%) of each question type in SQuAD1.1 and NewsQA.

Experiment 1 - question type results

- The curve of F-1 scores of QA systems trained on each question type subdomain with increasing sample size on SQuAD (top) and NewsQA (bottom).
- A QA system learns to answer a certain type of question mainly from the examples of the same question type, especially for *NUM, LOC* and *HUM* questions.



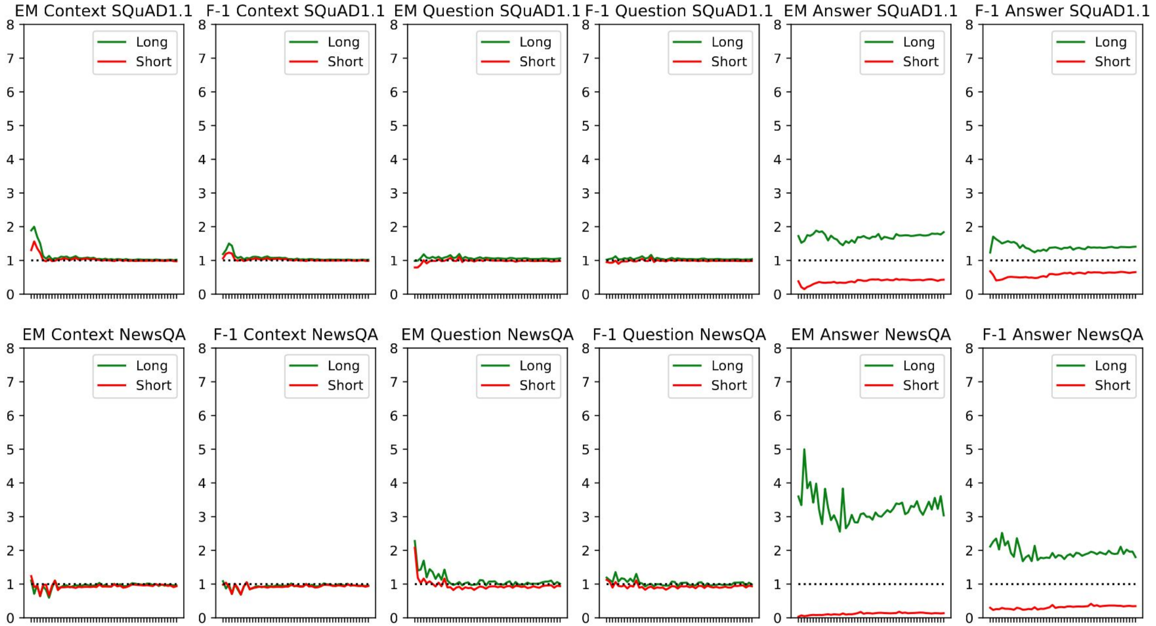
Experiment 2 - text length results

- We split the QA examples into ***long*** and ***short*** groups according to the median of context length, question length and answer length.

	Context	Question	Answer
SQuAD1.1	110	10	2
NewsQA	534	6	2

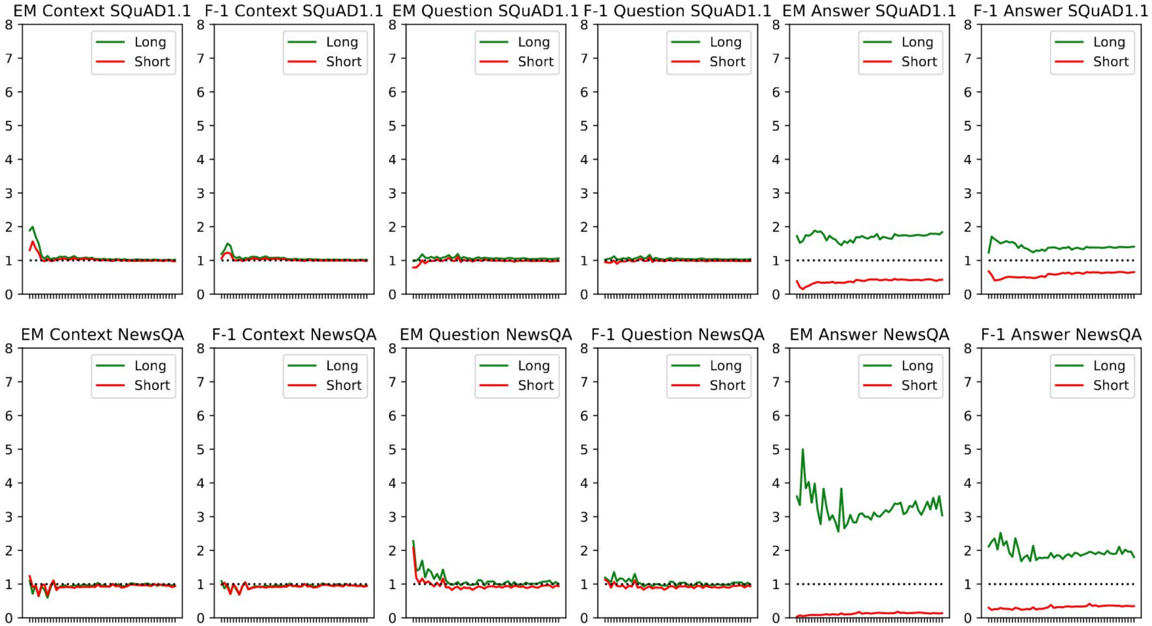
The median we used to partition the ***long*** and ***short*** groups

Experiment 2 - text length results



The ratio ($\frac{QA_L}{QA_S}$) of EM and F-1 score on ***Long*** and ***Short*** groups on SQuAD1.1 (top) and NewsQA (bottom)

Experiment 2 - text length results



- *If $\frac{QA_L}{QA_S}$ is close to 1, that means they don't have obvious difference in terms of performance on text length subdomains.*
- ***Context** and **question length** do **NOT** affect the performance of QA systems on Long and Short groups.*
- ***Answer length IS** a source of bias to QA systems.*

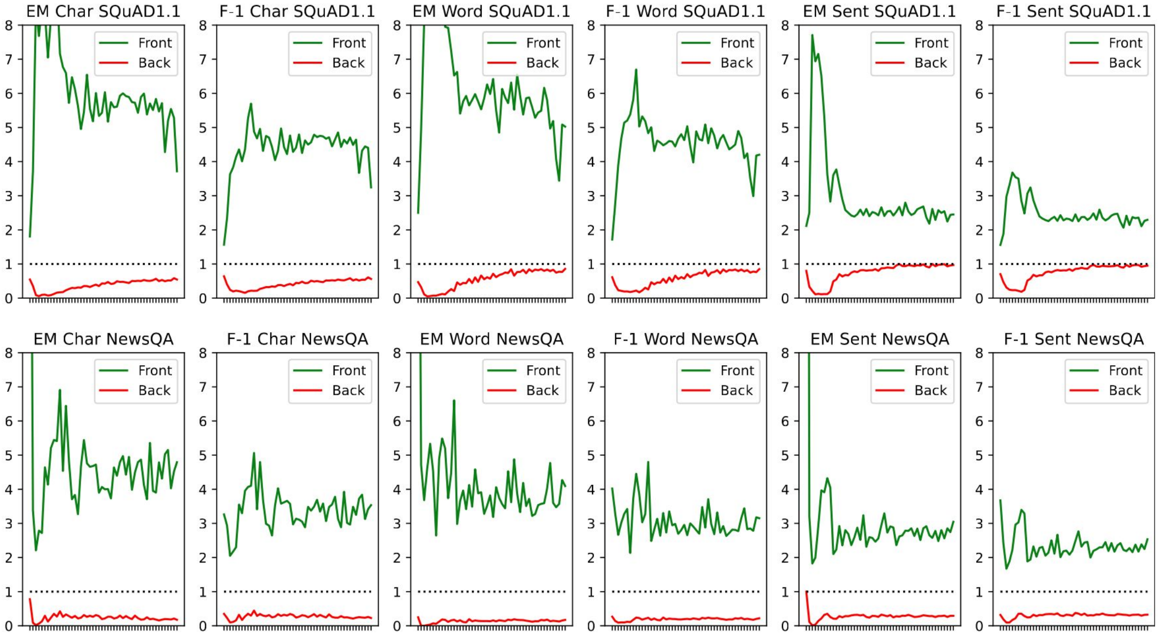
Experiment 3 - answer position results

- We split QA examples into *front* and *back* groups according to their answer positions at the character, word and sentence level.

	Char	Word	Sent
SQuAD1.1	262	46	1
NewsQA	358	67	2

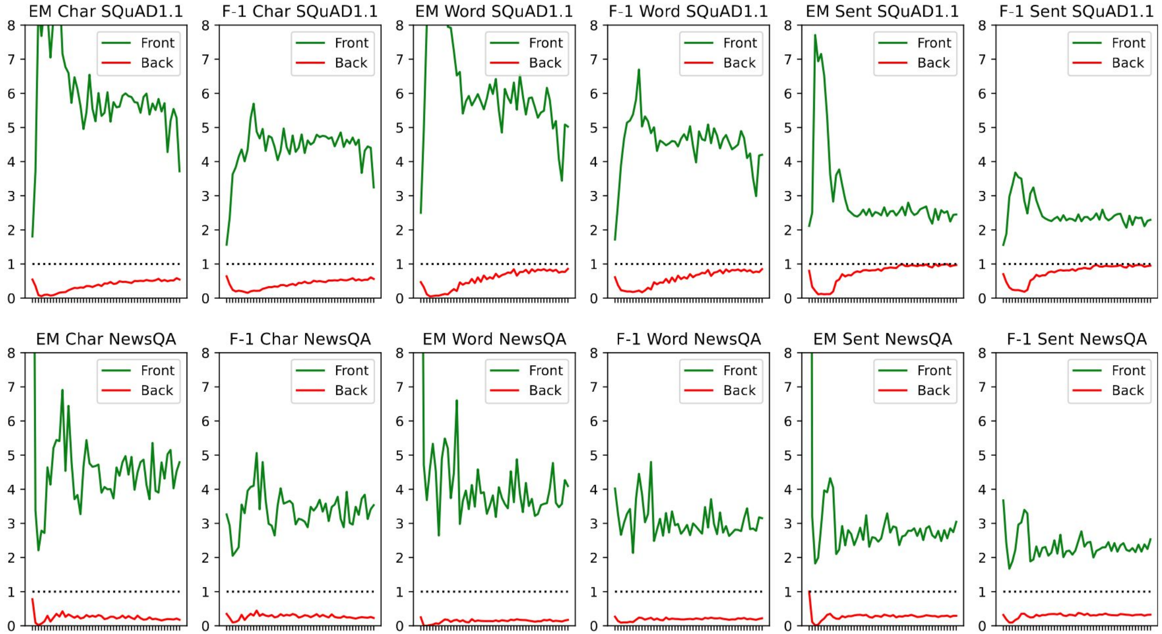
The median we used to partition the ***front*** and ***back*** groups

Experiment 3 - answer position results



The ratio ($\frac{QA_F}{QA_B}$) of EM and F-1 score on ***Front*** and ***Back*** groups on SQuAD1.1 (top) and NewsQA (bottom)

Experiment 3 - answer position results



- *If $\frac{QA_F}{QA_B}$ is close to 1, that means they don't have an obvious difference in terms of performance on text length subdomains.*
- ***Answer position at all three levels** is a source of bias for QA systems.*

Conclusion

- **Question type**, **answer length** and **answer position** are a source of bias to QA systems.
- We should control the distribution of these subdomains when constructing QA datasets and training QA systems.
- While context length poses minor effects on the generalizability of QA systems, we can speed up the training process and reduce the training cost for QA systems by shortening the contexts.

Thanks for listening!