Unify and Conquer
Towards a *Unified* View of Machine Comprehension

Daniel Khashabi
Allen Institute for AI, Seattle
Moving towards NLU, via QA
Moving towards NLU, via QA

- **Natural Language Understanding:**
  - Interpret a given text similar to humans.

- **Measuring the progress by answering questions.**
  - A system that is better in understanding language, should have a higher chance of answering these questions.

- This has been used in the field for many years
  - Question Answering,
  - Reading Comprehension,
  - Machine Comprehension, etc.
Moving towards NLU, via QA

• Natural Language Understanding:
  • Interpret a given text similar to humans.

• Measuring the progress by answering questions.
  • A system that is better in understanding language, should have a higher chance of answering these questions.

• This has been used in the field for many years
  • Question Answering,
  • Reading Comprehension,
  • Machine Comprehension, etc.

[Winograd, 1972; McCarthy 1976; Lehnert, 1977b; others]
Moving towards NLU, via QA

• Natural Language Understanding:
  • Interpret a given text similar to humans.

• Measuring the progress by answering questions.
  • A system that is better in understanding language, should have a higher chance of answering these questions.

• This has been used in the field for many years
  • Question Answering,
  • Reading Comprehension,
  • Machine Comprehension, etc.

[Winograd, 1972; McCarthy 1976; Lehnert, 1977b; others]
QA; a broad definition

- **Task**: Question Answering (QA)
QA; a broad definition

- **Task**: Question Answering (QA)

“What does photosynthesis produce that helps plants grow?”
QA; a broad definition

• **Task:** Question Answering (QA)

“What does photosynthesis produce that helps plants grow?”

**Input:** A question, along with additional information (hints, docs, images, etc.)
• **Task:** Question Answering (QA)

“How does photosynthesis produce that helps plants grow?”

**Input:** A question, along with additional information (hints, docs, images, etc.)
**Task:** Question Answering (QA)

“*What does photosynthesis produce that helps plants grow?*”

**Input:** A question, along with additional information (hints, docs, images, etc.)
• Task: Question Answering (QA)

“What does photosynthesis produce that helps plants grow?”

Input: A question, along with additional information (hints, docs, images, etc.)

Output: a string that addresses the input question.
QA datasets

- TREC-8
- TREC-9
- MCTest
- RACE
- ARC
- SQuAD 1
- SQuAD 2
- WinoGrande
- OBQA
- BoolQ
- NarQA
- DROP
- ComQA

Timeline:
- 2000
- 2005
- 2010
- 2015
- 2020

…
• Motivations for publishing new datasets:
  • Unexplored reasoning challenges
  • Alternate (better?) evaluation protocol (expand)
Motivations for publishing new datasets:
- Unexplored reasoning challenges
- Alternate (better?) evaluation protocol (expand)
• Motivations for publishing new datasets:
  • Unexplored reasoning challenges
  • Alternate (better?) evaluation protocol (expand)
QA datasets

TREC-8, TREC-9, TREC-2001-2005, MCTest, RACE, SQuAD 1, SQuAD 2, WinoGrande, NarQA, DROP, ComQA, OBQA, BoolQ, [Rajpurkar et al, 2016]
Question: “At what speed did the turbine operate?”

[Rajpurkar et al, 2016]
Question: “At what speed did the turbine operate?”

Candidates: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Question: “At what speed did the turbine operate?”

Candidates: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ... 

[Rajpurkar et al, 2016]
Question: “At what speed did the turbine operate?”

Candidates: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

“16,000 rpm”
QA datasets


MCTest

RACE ARC

SQuAD 1

SQuAD 2

DROP

ComQA

WinoGrande

OBQA

BoolQ

NarQA

[Clark et al, 2018]
Question: “What does photosynthesis produce that helps plants grow?”

[Clark et al, 2018]
Question: “What does photosynthesis produce that helps plants grow?”

Candidates: (A) water  
(B) oxygen  
(C) protein  
(D) sugar  

[Clark et al, 2018]
Question: “What does photosynthesis produce that helps plants grow? ”

Candidates:  
(A) water  
(B) oxygen  
(C) protein  
(D) sugar

[Clark et al, 2018]
Question: "What does photosynthesis produce that helps plants grow?"

Candidates:
- (A) water
- (B) oxygen
- (C) protein
- (D) sugar

"The big kid"

[Clark et al, 2018]
QA Terminology

• “Task”: well-formed response for a well-formed question.

  Input: well-formed question  
  Output: a well-formed response

• “Format”: QA with particular assumptions about input/output.
  • Defined by datasets.
  • A necessity for automatic evaluation.
  • Depends on the reasoning problem, too.
**QA Terminology**

- **“Task”:** well-formed response for a well-formed question.
  - Input: well-formed question
  - Output: a well-formed response

- **“Format”:** QA with particular **assumptions** about input/output.
  - Defined by datasets.
  - A necessity for automatic evaluation.
  - Depends on the reasoning problem, too.
QA Terminology

• “Task”: well-formed response for a well-formed question.
  - Input: well-formed question
  - Output: a well-formed response

• “Format”: QA with particular assumptions about input/output.
  - Defined by datasets.
  - A necessity for automatic evaluation.
  - Depends on the reasoning problem, too.

<table>
<thead>
<tr>
<th>Format</th>
<th>Example dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-choice</td>
<td>CommonsenseQA [Talmor et al’19]</td>
</tr>
<tr>
<td>YesNo</td>
<td>BoolQ [Clark et al’19]</td>
</tr>
<tr>
<td>extractive</td>
<td>SQuAD [Rajpurkar et al’16]</td>
</tr>
<tr>
<td>abstractive</td>
<td>NarrativeQA [Kociský et al’18]</td>
</tr>
</tbody>
</table>
QA Terminology

• “Task”: well-formed response for a well-formed question.
  
  Input: well-formed question

  ![Robot](image)

  Output: a well-formed response

• “Format”: QA with particular assumptions about input/output.
  
  - Defined by datasets.
  - A necessity for automatic evaluation.
  - Depends on the reasoning problem, too.

<table>
<thead>
<tr>
<th>Format</th>
<th>Example dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-choice</td>
<td>CommonsenseQA [Talmor et al’19]</td>
</tr>
<tr>
<td>YesNo</td>
<td>BoolQ [Clark et al’19]</td>
</tr>
<tr>
<td>extractive</td>
<td>SQuAD [Rajpurkar et al’16]</td>
</tr>
<tr>
<td>abstractive</td>
<td>NarrativeQA [Kociský et al’18]</td>
</tr>
</tbody>
</table>
Our progress in QA: the good

• More general language representations.
Our progress in QA: the good

• More general language representations.

[Harabagiu et al, 2000; others]
Our progress in QA: the good

- More general language representations.
Our progress in QA: the bad
Our progress in QA: the bad

- Task-specific assumptions
Our progress in QA: the bad

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td></td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td></td>
</tr>
<tr>
<td>Extractive QA</td>
<td></td>
</tr>
<tr>
<td>Abstractive QA</td>
<td></td>
</tr>
</tbody>
</table>
Our progress in QA: the bad

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td>binary output</td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td></td>
</tr>
<tr>
<td>Extractive QA</td>
<td></td>
</tr>
<tr>
<td>Abstractive QA</td>
<td></td>
</tr>
</tbody>
</table>
Our progress in QA: the bad

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td><em>binary output</em></td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td><em>exactly one of the candidate answers is correct.</em></td>
</tr>
<tr>
<td>Extractive QA</td>
<td></td>
</tr>
<tr>
<td>Abstractive QA</td>
<td></td>
</tr>
</tbody>
</table>
Our progress in QA: the bad

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td>binary output</td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td>exactly one of the candidate answers is correct.</td>
</tr>
<tr>
<td>Extractive QA</td>
<td>answer is a subset of a given paragraph</td>
</tr>
<tr>
<td>Abstractive QA</td>
<td></td>
</tr>
</tbody>
</table>
Our progress in QA: the bad

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td>binary output</td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td>exactly one of the candidate answers is correct.</td>
</tr>
<tr>
<td>Extractive QA</td>
<td>answer is a subset of a given paragraph</td>
</tr>
<tr>
<td>Abstractive QA</td>
<td>answer is a mixture of what is given and items not given.</td>
</tr>
</tbody>
</table>
Our progress in QA: the bad

Consequences of format-specific design:
- Prevent generalization across formats
- Don’t benefit from labeled data of other formats.

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td>binary output</td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td>exactly one of the candidate answers is correct.</td>
</tr>
<tr>
<td>Extractive QA</td>
<td>answer is a subset of a given paragraph</td>
</tr>
<tr>
<td>Abstractive QA</td>
<td>answer is a mixture of what is given and items not given.</td>
</tr>
</tbody>
</table>
Our progress in QA: the bad

Consequences of format-specific design:
• Prevent generalization across formats
• Don’t benefit from labeled data of other formats.

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td>binary output</td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td>exactly one of the candidate answers is correct.</td>
</tr>
<tr>
<td>Extractive QA</td>
<td>answer is a subset of a given paragraph</td>
</tr>
<tr>
<td>Abstractive QA</td>
<td>answer is a mixture of what is given and items not given.</td>
</tr>
</tbody>
</table>
Our progress in QA: the bad

Consequences of format-specific design:
- Prevent generalization across formats
- Don’t benefit from labeled data of other formats.

<table>
<thead>
<tr>
<th>format</th>
<th>assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No QA</td>
<td>binary output</td>
</tr>
<tr>
<td>Multiple-choice QA</td>
<td>exactly one of the candidate answers is correct.</td>
</tr>
<tr>
<td>Extractive QA</td>
<td>answer is a subset of a given paragraph</td>
</tr>
<tr>
<td>Abstractive QA</td>
<td>answer is a mixture of what is given and items not given.</td>
</tr>
</tbody>
</table>
formats-specialized models

ExtractiveQA

MultipleChoiceQA
Question: “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

“16,000 rpm”
Excerpts from the document:

**ExtractiveQA**

**Question:** “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) **16,000 rpm** bladeless turbine. ...

**MultipleChoiceQA**

**Question:** “What does photosynthesis produce that helps plants grow?”

- (A) water
- (B) oxygen
- (C) protein
- (D) sugar

**Answer:** “sugar”
ExtractiveQA

**Question:** “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

“16,000 rpm”

MultipleChoiceQA

**Question:** “What does photosynthesis produce that helps plants grow?”

(A) water  
(B) oxygen  
(C) protein  
(D) sugar

“sugar”
Question: “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

“16,000 rpm”

Question: “What does photosynthesis produce that helps plants grow?”

(A) water
(B) oxygen
(C) protein
(D) sugar

“sugar”
Question: “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

“16,000 rpm”

Question: “What does photosynthesis produce that helps plants grow?”

(A) water
(B) oxygen
(C) protein
(D) sugar

“sugar”
Question: “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

“16,000 rpm”

Question: “What does photosynthesis produce that helps plants grow?”

(A) water
(B) oxygen
(C) protein
(D) sugar

“sugar”
**Beyond formats-specialized models**

**ExtractiveQA**

**Question:** “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) **16,000 rpm** bladeless turbine. ...

“16,000 rpm”

**MultipleChoiceQA**

**Question:** “What does photosynthesis produce that helps plants grow?”

(A) water  
(B) oxygen  
(C) protein  
(D) sugar

“sugar”
Beyond formats-specialized models

ExtractiveQA

Question: “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

MultipleChoiceQA

Question: “What does photosynthesis produce that helps plants grow?”

(A) water
(B) oxygen
(C) protein
(D) sugar

“16,000 rpm”

“sugar”
Beyond formats-specialized models

ExtractiveQA

**Question:** “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) **16,000 rpm** bladeless turbine. ...

“16,000 rpm”

MultipleChoiceQA

**Question:** “What does photosynthesis produce that helps plants grow?”

(A) water  
(B) oxygen  
(C) protein  
(D) sugar

“sugar”
Beyond formats-specialized models

ExtractiveQA

Question: “At what speed did the turbine operate?”

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

MultipleChoiceQA

Question: “What does photosynthesis produce that helps plants grow?”

(A) water  
(B) oxygen  
(C) protein  
(D) sugar

“16,000 rpm”

“sugar”
Talk Summary & Statement

• Creating **format-specific QA** models distance us from broad QA.

• There is **overlap** between underlying reasoning abilities of formats.
  • One can **benefit** from **mixing** QA formats.

• **UnifiedQA**: a single QA system working across four common QA formats.
  • Fine-tuning models pre-trained on UnifiedQA yields **SOTA** results.
Creating **format-specific QA** models **distance** us from broad QA.

There is **overlap** between underlying reasoning abilities of formats.
  • One can **benefit** from **mixing** QA formats.

**UnifiedQA**: a single QA system working across four common QA formats.
  • Fine-tuning models pre-trained on UnifiedQA yields **SOTA** results.
• Creating **format-specific QA** models **distance** us from broad QA.

• There is **overlap** between underlying reasoning abilities of formats.
  • One can **benefit** from **mixing** QA formats.

• UnifiedQA: a single QA system working across four common QA formats.
  • Fine-tuning models pre-trained on UnifiedQA yields **SOTA** results.
Earlier work on multi-task learning
Earlier work on multi-task learning

- In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

- The choice of tasks is also important.
  - Earlier works select too broad of tasks.
    - E.g., Raffel et al’19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

- We narrow the scope of tasks to stay within the boundaries of QA.
  - No task/format specific encoding.
Earlier work on multi-task learning

• In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

• The choice of tasks is also important.
  • Earlier works select too broad of tasks.
    • E.g., Raffel et al’19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

• We narrow the scope of tasks to stay within the boundaries of QA.
  • No task/format specific encoding.

 Didn’t work before; why would it work now? 😐
Earlier work on multi-task learning

• In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

• The choice of tasks is also important.
  • Earlier works select too broad of tasks.
    • E.g., Raffel et al'19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

• We narrow the scope of tasks to stay within the boundaries of QA.
  • No task/format specific encoding.
Earlier work on multi-task learning

• In the same spirit as multi-task learning. [Caruana'97; McCann et al’18]

• The choice of tasks is also important.
  • Earlier works select too broad of tasks.
    • E.g., Raffel et al’19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

• We narrow the scope of tasks to stay within the boundaries of QA.
  • No task/format specific encoding.

Didn’t work before; why would it work now? 🤔
Earlier work on multi-task learning

• In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

• The choice of tasks is also important.
  • Earlier works select too broad of tasks.
    • E.g., Raffel et al’19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

• We narrow the scope of tasks to stay within the boundaries of QA.
  • No task/format specific encoding.
Earlier work on multi-task learning

• In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

• The choice of tasks is also important.
  • Earlier works select too broad of tasks.
    • E.g., Raffel et al’19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

• We narrow the scope of tasks to stay within the boundaries of QA.
  • No task/format specific encoding.
Roadmap

1. Generalization across formats
2. UnifiedQA + Empirical Intuitions
3. Discussion and next steps
Roadmap

1. Generalization across formats
2. UnifiedQA + Empirical Intuitions
3. Discussion and next steps
UnifiedQA: a high-level definition
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.
2. The input should be natural.
   • Minimal-enough for a human solver to infer the format.
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.
2. The input should be natural.
   • Minimal-enough for a human solver to infer the format.
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.
2. The input should be natural.
   • Minimal-enough for a human solver to infer the format.

“What causes sound?
(A) sunlight (B) vibrations (C) x-rays (D) pitch”
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.

2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

   “What causes sound?

   (A) sunlight (B) vibrations (C) x-rays (D) pitch”

   “vibrations”
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of **QA formats**.

2. The input should be **natural**.
   - Minimal-Enough for a human solver to infer the format.

```
“Is Jamaica considered part of the United States?

(Jamaica) Jamaica (/dʒəˈmeɪkə/ (listen)) is an island country situated in the Caribbean Sea. Spanning 10,990 square kilometres (4,240 sq mi) in area, it is the third-largest island of the Greater Antilles and the fourth-largest island country in the Caribbean.”
```
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.
2. The input should be natural.
   • Minimal-enough for a human solver to infer the format.

“Is Jamaica considered part of the United States?

(Jamaica) Jamaica (/dʒəˈmeɪkə/ ( listen)) is an island country situated in the Caribbean Sea. Spanning 10,990 square kilometres (4,240 sq mi) in area, it is the third-largest island of the Greater Antilles and the fourth-largest island country in the Caribbean.”

“no”
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.

2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

   “What type of musical instruments did the Yuan bring to China? (Yuan dynasty) Western musical instruments were introduced to enrich Chinese performing arts. From this period dates the conversion to Islam, by Muslims of Central Asia, of growing numbers of Chinese in the northwest and southwest. …”
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.

2. The input should be natural.
   - Minimal-enough for a human solver to infer the format.

“What type of musical instruments did the Yuan bring to China?

(Yuan_dynasty) Western musical instruments were introduced to enrich Chinese performing arts. From this period dates the conversion to Islam, by Muslims of Central Asia, of growing numbers of Chinese in the northwest and southwest. …”

“Western musical instruments”
UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of QA formats.
2. The input should be natural.
   - Minimal-enough for a human solver to infer the format.

> What type of musical instruments did the Yuan bring to China?

(Yuan_dynasty) Western musical instruments were introduced to enrich Chinese performing arts. From this period dates the conversion to Islam, by Muslims of Central Asia, of growing numbers of Chinese in the northwest and southwest. ...

> "Western musical instruments"
UnifiedQA: towards an implementation
UnifiedQA: towards an implementation

• Use text-to-text architectures
  • T5 [Raffal et al, 2020], BART [Lewis et al, 2019], etc.

• Train simultaneously on all datasets jointly together.
  • Batches contains the same number of instances from each training set.
UnifiedQA: towards an implementation

- Use text-to-text architectures
  - T5 [Raffal et al, 2020], BART [Lewis et al, 2019], etc.

- Train simultaneously on all datasets jointly together.
  - Batches contains the same number of instances from each training set.
UnifiedQA: towards an implementation

• Use text-to-text architectures
  • T5 [Raffel et al, 2020], BART [Lewis et al, 2019], etc.

• Train simultaneously on all datasets jointly together.
  • Batches contains the same number of instances from each training set.
Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?
Mixing pairs of formats: experiment (1)

• Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice)

w/ datasets of different formats.
<table>
<thead>
<tr>
<th>Mixing pairs of formats: experiment (1)</th>
</tr>
</thead>
</table>

- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice) w/ datasets of different formats.

<table>
<thead>
<tr>
<th>Trained on RACE</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

- | RACE | MCTest |
- | [Lai et al. 17] | [Richardson et al. 15] |
- | 50 | }
• Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice) w/ datasets of different formats.

Trained on RACE

<table>
<thead>
<tr>
<th>Format</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACE</td>
<td>55.8</td>
</tr>
<tr>
<td>MCTest</td>
<td>62.5</td>
</tr>
</tbody>
</table>

[Richardson et al. 15]

[Richardson et al. 15]
Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice) w/ datasets of different formats.

- Trained on RACE
- Trained on RACE + SQuAD 1

*Comparison of performance:*

<table>
<thead>
<tr>
<th>Format</th>
<th>RACE 55.8</th>
<th>MCTest 62.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained on RACE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trained on RACE + SQuAD 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Richardson et al. 15]

[Richardson et al. 17]
Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice) w/ datasets of different formats.

- Trained on RACE
- Trained on RACE + SQuAD 1

<table>
<thead>
<tr>
<th>Format</th>
<th>Dataset</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACE</td>
<td>[Lai et al. 17]</td>
<td>59.1</td>
</tr>
<tr>
<td>RACE</td>
<td>Trained on RACE</td>
<td>55.8</td>
</tr>
<tr>
<td>MCTest</td>
<td>[Richardson et al. 15]</td>
<td>62.5</td>
</tr>
</tbody>
</table>
• Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice) w/ datasets of different formats.

- Trained on RACE
- Trained on RACE + SQuAD 1

<table>
<thead>
<tr>
<th>Test</th>
<th>Trained on RACE</th>
<th>Trained on RACE + SQuAD 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACE</td>
<td>55.8</td>
<td>59.1</td>
</tr>
<tr>
<td>MCTest</td>
<td>62.5</td>
<td>69.4</td>
</tr>
</tbody>
</table>

[Richardson et al. 15]

[88]
Is there any value in out-of-format training?

Mixing BoolQ (YesNo)

w/ datasets of different formats.

- Trained on BoolQ
- Trained on BoolQ + X

<table>
<thead>
<tr>
<th></th>
<th>BoolQ</th>
<th>BoolQ-CS</th>
<th>MultiRC (YN subset)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Clark et al. 19]</td>
<td>61.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Gardner et al. 20]</td>
<td>76.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[K et al. 18]</td>
<td>53.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Is there any value in out-of-format training?

Mixing BoolQ (YesNo) w/ datasets of different formats.

Trained on BoolQ

Trained on BoolQ + X

Mixing pairs of formats: experiment (2)

<table>
<thead>
<tr>
<th>Format</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoolQ</td>
<td>76.4</td>
</tr>
<tr>
<td>BoolQ-CS</td>
<td></td>
</tr>
<tr>
<td>MultiRC (YN subset)</td>
<td></td>
</tr>
</tbody>
</table>

[Clark et al. 19] [Gardner et al. 20] [K et al. 18]
• Is there any value in out-of-format training?

Mixing BoolQ (YesNo) w/ datasets of different formats.

<table>
<thead>
<tr>
<th></th>
<th>BoolQ</th>
<th>BoolQ-CS</th>
<th>MultiRC (YN subset)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained on BoolQ</td>
<td>76.4</td>
<td>53.4</td>
<td></td>
</tr>
<tr>
<td>Trained on BoolQ + X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Clark et al. 19]  
[Gardner et al. 20]  
[K et al. 18]
Is there any value in out-of-format training?

Mixing BoolQ (YesNo) w/ datasets of different formats.

- Trained on BoolQ
- Trained on BoolQ + X

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Score</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoolQ</td>
<td>76.4</td>
<td>[Clark et al. 19]</td>
</tr>
<tr>
<td>BoolQ-CS</td>
<td>53.4</td>
<td>[Gardner et al. 20]</td>
</tr>
<tr>
<td>MultiRC (YN subset)</td>
<td>64.1</td>
<td>[K et al. 18]</td>
</tr>
</tbody>
</table>
Mixing pairs of formats: experiment (2)

- Is there any value in out-of-format training?

Mixing BoolQ (YesNo) w/ datasets of different formats.

Trained on BoolQ

Trained on BoolQ + X

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoolQ</td>
<td>76.4</td>
</tr>
<tr>
<td>BoolQ-CS</td>
<td>53.4</td>
</tr>
<tr>
<td>MultiRC (YN subset)</td>
<td>64.1</td>
</tr>
</tbody>
</table>

X = SQuAD 1 (Extractive)

[Clark et al. 19]  [Gardner et al. 20]  [K et al. 18]
Is there any value in out-of-format training?

Mixing BoolQ (YesNo) w/ datasets of different formats.

- Trained on BoolQ
- Trained on BoolQ + X

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Output Format</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoolQ</td>
<td>Extractive</td>
<td>76.4</td>
</tr>
<tr>
<td>BoolQ-CS</td>
<td>Abstractive</td>
<td>61.0</td>
</tr>
<tr>
<td>MultiRC (YN subset)</td>
<td></td>
<td>64.1</td>
</tr>
</tbody>
</table>

- X=SQuAD 1 (Extractive)
- X=NarQA (Abstractive)

References:
- [Clark et al. 19]
- [Gardner et al. 20]
- [K et al. 18]
Is there any value in out-of-format training?

Mixing BoolQ (YesNo) w/ datasets of different formats.

- Trained on BoolQ
- Trained on BoolQ + X

Mixing pairs of formats: experiment (2)

- BoolQ: 76.4
- BoolQ-CS: 53.4
- MultiRC (YN subset): 64.1

X=SQuAD 1 (Extractive)
X=NarQA (Abstractive)
X=SQuAD 1 (Extractive)

[Clark et al. 19]  [Gardner et al. 20]  [K et al. 18]
Roadmap

1. Generalization across formats
2. UnifiedQA + Empirical Intuitions
3. Discussion and next steps
1. Generalization across formats

2. UnifiedQA + Empirical Intuitions

3. Discussion and next steps
UnifiedQA-v1
UnifiedQA-v1

• Trained on the union of different formats:
  • Extractive: SQuAD 1.1, SQuAD 2.0
  • Abstractive: NarrativeQA
  • Multiple-choice: RACE, ARC, OBQA, MCTest
  • YesNo: BoolQ

• Architectures:
  • T5 (11B, 3B, ...)
  • BART (large)
UnifiedQA-v1

• Trained on the union of different formats:
  • Extractive: SQuAD 1.1, SQuAD 2.0
  • Abstractive: NarrativeQA
  • Multiple-choice: RACE, ARC, OBQA, MCTest
  • YesNo: BoolQ

• Architectures:
  • T5 (11B, 3B, ...)
  • BART (large)

https://github.com/allenai/unifiedqa
Intuition #1: Comparison w/ Dedicated Models
Intuition #1: Comparison w/ Dedicated Models

The chart compares the performance of Dedicated Models and UnifiedQA across various datasets. The datasets include: SQuAD1.1, SQuAD2, RACE, OBOA, ARC-Easy, ARC-Chal, MCTest, BoolQ, NarQA, and the average (Avg.). The dedicated models show higher scores in most categories, indicating superior performance in these tasks.
Intuition #1: Comparison w/ Dedicated Models

- Is UnifiedQA as good as systems dedicated to individual datasets?

- UnifiedQA performs almost as good as individual T5 models targeted to each dataset.
Intuition #1: Comparison w/ Dedicated Models

- Is UnifiedQA as good as systems dedicated to individual datasets?

UnifiedQA performs almost as good as individual T5 models targeted to each dataset.

<table>
<thead>
<tr>
<th></th>
<th>SQuAD2</th>
<th>RACE</th>
<th>BoolQ</th>
<th>NarQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5 (SQuAD 2)</td>
<td>91</td>
<td>33</td>
<td>12</td>
<td>51</td>
</tr>
<tr>
<td>T5 (RACE)</td>
<td>43</td>
<td>87</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>T5 (BoolQ)</td>
<td>4</td>
<td>22</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>T5 (NarQA)</td>
<td>45</td>
<td>48</td>
<td>47</td>
<td>65</td>
</tr>
<tr>
<td>UnifiedQA</td>
<td>90</td>
<td>87</td>
<td>90</td>
<td>65</td>
</tr>
</tbody>
</table>

![Evaluation Sets Graph](image-url)
Intuition #2: UnseenDatasets
Intuition #2: Unseen Datasets

<table>
<thead>
<tr>
<th>Evaluation Sets</th>
<th>NewsQA</th>
<th>Quoref</th>
<th>DROP</th>
<th>DROP-CS</th>
<th>QASC</th>
<th>CommonsenseQA</th>
<th>NP-BoolQ</th>
<th>BoolQ-CS</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>UnifiedQA [EX]</td>
<td>59</td>
<td>65</td>
<td>25</td>
<td>24</td>
<td>55</td>
<td>63</td>
<td>21</td>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>UnifiedQA [AB]</td>
<td>58</td>
<td>68</td>
<td>31</td>
<td>37</td>
<td>54</td>
<td>59</td>
<td>27</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>UnifiedQA [MC]</td>
<td>48</td>
<td>68</td>
<td>29</td>
<td>37</td>
<td>68</td>
<td>76</td>
<td>3</td>
<td>6</td>
<td>44</td>
</tr>
<tr>
<td>UnifiedQA [YN]</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>21</td>
<td>79</td>
<td>79</td>
<td>22</td>
</tr>
<tr>
<td>UnifiedQA</td>
<td>59</td>
<td>63</td>
<td>33</td>
<td>40</td>
<td>68</td>
<td>76</td>
<td>81</td>
<td>80</td>
<td>62</td>
</tr>
</tbody>
</table>
Intuition #2: Unseen Datasets

• Does UnifiedQA generalize well to unseen datasets?

<table>
<thead>
<tr>
<th>Evaluation Sets</th>
<th>NewsQA</th>
<th>Quoref</th>
<th>DROP</th>
<th>DROP-CS</th>
<th>QASC</th>
<th>Commonse nseQA</th>
<th>NP-BoolQ</th>
<th>BoolQ-CS</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>UnifiedQA [EX]</td>
<td>59</td>
<td>65</td>
<td>25</td>
<td>24</td>
<td>55</td>
<td>63</td>
<td>21</td>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>UnifiedQA [AB]</td>
<td>58</td>
<td>68</td>
<td>31</td>
<td>37</td>
<td>54</td>
<td>59</td>
<td>27</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>UnifiedQA [MC]</td>
<td>48</td>
<td>68</td>
<td>29</td>
<td>37</td>
<td>68</td>
<td>76</td>
<td>3</td>
<td>6</td>
<td>44</td>
</tr>
<tr>
<td>UnifiedQA [YN]</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>21</td>
<td>79</td>
<td>79</td>
<td>22</td>
</tr>
<tr>
<td>UnifiedQA</td>
<td>59</td>
<td>63</td>
<td>33</td>
<td>40</td>
<td>68</td>
<td>76</td>
<td>81</td>
<td>80</td>
<td>62</td>
</tr>
</tbody>
</table>

- UnifiedQA shows much stronger generalization across a wide range of datasets.
Fine-tuning on UnifiedQA

- Is there a value in using UnifiedQA as a starting point for fine-tuning?
  - Show SOTA on 10 datasets (OBQA, QASC, RACE, WinoGrande, PIQA, SIQA, ROPES)
  - Similar trends for BART

![Graph showing performance metrics](image)

- Fine-tuned on T5
- Fine-tuned UnifiedQA (based on T5)

Performance metrics:
- ARC-chall
  - [Clark et al. 18]
- CommonsenseQA
  - [Talmor et al. 19]
- OBQA
  - [Khot et al. 19]
Fine-tuning on UnifiedQA

- Is there a value in using UnifiedQA as a starting point for fine-tuning?
- Show SOTA on 10 datasets (OBQA, QASC, RACE, WinoGrande, PIQA, SIQA, ROPES)
- Similar trends for BART

![Graph showing fine-tuning results for ARC-chall, CommonsenseQA, and OBQA datasets.](chart.png)
• Is there a value in using UnifiedQA as a starting point for fine-tuning?
  • Show SOTA on 10 datasets (OBQA, QASC, RACE, WinoGrande, PIQA, SIQA, ROPES)
  • Similar trends for BART

```
87
84
79
78
78
75
70
65
65
70
75
85

Fine-tuned on T5
Fine-tuned UnifiedQA (based on T5)
```

![Graph showing comparison between Fine-tuned on T5 and Fine-tuned UnifiedQA (based on T5)]
Demo

https://unifiedqa.apps.allenai.org
Roadmap

1. Generalization across formats
2. UnifiedQA + Empirical Intuitions
3. Discussion and next steps
1. **Generalization across formats**

2. **UnifiedQA + Empirical Intuitions**

3. **Discussion and next steps**
Methodological Issue: Data Leakage
Methodological Issue: Data Leakage

• “have you done some studies on overlap across datasets?”

• Easy answer:
  • not much surface-form overlap between the datasets.

• Nuanced/ difficult answer:
  • more data (especially during pre-training) increases the chances of (indirect) leakage.
Methodological Issue: Data Leakage

• “have you done some studies on overlap across datasets?”

• Easy answer:
  • not much surface-form overlap between the datasets.

• Nuanced/ difficult answer:
  • more data (especially during pre-training) increases the chances of (indirect) leakage.
Where do we go from here?
Where do we go from here?

• More formats
  • Can we incorporate other “natural” variations of QA in the study?

• Smaller models:
  • Can we build small and accurate models to make it more available?

• Beyond QA/Text:
  • Can you take these ideas and apply it to some other problems?
Where do we go from here?

• More formats
  • Can we incorporate other “natural” variations of QA in the study?

• Smaller models:
  • Can we build small and accurate models to make it more available?

• Beyond QA/Text:
  • Can you take these ideas and apply it to some other problems?
Where do we go from here?

• More formats
  • Can we incorporate other “natural” variations of QA in the study?

• Smaller models:
  • Can we build small and accurate models to make it more available?

• Beyond QA/Text:
  • Can you take these ideas and apply it to some other problems?
Take-home points

• The field relies *excessively* format-specific assumptions for system design.
  • Instead, we should move towards more general QA architectures.

• **Incentive:** there is value in mixing QA datasets of different formats.

• UnifiedQA, a single pre-trained QA system seeking to bring unification across common QA formats.

https://github.com/allenai/unifiedqa