SituatedQA:
Incorporating Extra-Linguistic Contexts into QA
Michael J.Q. Zhang and Eunsol Choi
Motivation

Where were the last Winter Olympic Games held?

How many gold medals did we win in the Winter Olympics 2018?

List of Winter Olympic Games

<table>
<thead>
<tr>
<th>No.</th>
<th>Year</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXII</td>
<td>2014</td>
<td>Sochi</td>
</tr>
<tr>
<td>XXIII</td>
<td>2018</td>
<td>Pyeongchang</td>
</tr>
<tr>
<td>XXIV</td>
<td>2022</td>
<td>Beijing</td>
</tr>
</tbody>
</table>

2018 Winter Olympics Medal Table

<table>
<thead>
<tr>
<th>Rank</th>
<th>NOC</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Australia (AUS)</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Austria (AUT)</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>Belarus (BLR)</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>Belgium (BEL)</td>
<td>0</td>
</tr>
</tbody>
</table>
**Motivation**

Where were the last Winter Olympic Games held?

Sochi, Russia

How many gold medals did we win in the Winter Olympics 2018?

**List of Winter Olympic Games**

<table>
<thead>
<tr>
<th>No.</th>
<th>Year</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXII</td>
<td>2014</td>
<td>Sochi</td>
</tr>
<tr>
<td>XXIII</td>
<td>2018</td>
<td>Pyeongchang</td>
</tr>
<tr>
<td>XXIV</td>
<td>2022</td>
<td>Beijing</td>
</tr>
</tbody>
</table>

**2018 Winter Olympics Medal Table**

<table>
<thead>
<tr>
<th>Rank</th>
<th>NOC</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Australia (AUS)</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Austria (AUT)</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>Belarus (BLR)</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>Belgium (BEL)</td>
<td>0</td>
</tr>
</tbody>
</table>
Motivation

Where were the last Winter Olympic Games held?
Sochi, Russia

How many gold medals did we win in the Winter Olympics 2018?
5

List of Winter Olympic Games

<table>
<thead>
<tr>
<th>No.</th>
<th>Year</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXII</td>
<td>2014</td>
<td>Sochi</td>
</tr>
<tr>
<td>XXIII</td>
<td>2018</td>
<td>Pyeongchang</td>
</tr>
<tr>
<td>XXIV</td>
<td>2022</td>
<td>Beijing</td>
</tr>
</tbody>
</table>

2018 Winter Olympics Medal Table

<table>
<thead>
<tr>
<th>Rank</th>
<th>NOC</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Australia (AUS)</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Austria (AUT)</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>Belarus (BLR)</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>Belgium (BEL)</td>
<td>0</td>
</tr>
</tbody>
</table>
Motivation

Where were the last Winter Olympic Games held?

Sochi, Russia

20% of answers depend on when the question was asked.

How many gold medals did we win in the Winter Olympics 2018?

5

5% of questions specify a location where it was asked.

List of Winter Olympic Games

<table>
<thead>
<tr>
<th>No.</th>
<th>Year</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXII</td>
<td>2014</td>
<td>Sochi</td>
</tr>
<tr>
<td>XXIII</td>
<td>2018</td>
<td>Pyeongchang</td>
</tr>
<tr>
<td>XXIV</td>
<td>2022</td>
<td>Beijing</td>
</tr>
</tbody>
</table>

2018 Winter Olympics Medal Table

<table>
<thead>
<tr>
<th>Rank</th>
<th>NOC</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Australia (AUS)</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Austria (AUT)</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>Belarus (BLR)</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>Belgium (BEL)</td>
<td>0</td>
</tr>
</tbody>
</table>

[Analysis on Natural Questions, Kwitakowski et al, TACL 2019]
The answer to a question may depend on its extra-linguistic contexts (e.g., when and where it was asked).
Motivation: Advances in QA Benchmarking

Where was the last Winter Olympic Games held?

Why is this an study issue now?

Recent advances in NLP have opened new doors for benchmarking QA models

PyeongChang
Motivation: Advances in QA Benchmarking

Where was the last Winter Olympic Games held?

Benchmarking in the Past

2018 Winter Olympics

The 2018 Winter Olympics, officially known as the XXIII Olympic Winter Games (French: Les XXIIIᵉ Jeux olympiques d'hiver; [a] Korean: 제23회 동계 올림픽, romanized: Jeisipsamhoe Donggye Ollimpik) and commonly known as PyeongChang 2018 (Korean: 평창 2018), was an international winter multi-sport event that was held between 9 and 25 February 2018 in Pyeongchang County, Gangwon Province, South Korea, with the opening rounds for certain events held on 8 February 2018, the day before the opening ceremony.

Past benchmarks assume:
- The answer is a span in the provided passage
- All the necessary context is given in the document

PyeongChang
Motivation: Advances in QA Benchmarking

Where was the last Winter Olympic Games held?

Benchmarking in the Past

2018 Winter Olympics

The 2018 Winter Olympics, officially known as the XXIII Olympic Winter Games (French: Les XXIIIèmes Jeux olympiques d'hiver;[a] Korean: 제23회 동계 올림픽, romanized: Jeisipsamhoe Donggye Ollimpik) and commonly known as PyeongChang 2018 (Korean: 평창 2018), was an international winter multi-sport event that was held between 9 and 25 February 2018 in Pyeongchang County, Gangwon Province, South Korea, with the opening rounds for certain events held on 8 February 2018, the day before the opening ceremony.

Past benchmarks assume:

- The answer is a span in the provided passage
- All the necessary context is given in the document

Today’s benchmarks assume:

- The answer is out there somewhere…
- We have to assume the context
The SituatedQA Task

Temporally Dependent Question

Where were the last Winter Olympic Games held?

Geographically Dependent Question

How many gold medals did we win in the Winter Olympics 2018?

Objectives

1. Determine whether QA models understand how information changes across different contexts

2. Create QA systems that can produce the appropriate answer for a given context
The SituatedQA Task

Temporally Dependent Question
Where were the last Winter Olympic Games held?

Temporal Context (Timestamp)
- 2014
- 2018
- 2022

Geographically Dependent Question
How many gold medals did we win in the Winter Olympics 2018?

Geographical Context (Location)
- (US)
- (Brazil)
- (Norway)
The SituatedQA Task

**Temporally Dependent Question**

Where were the last Winter Olympic Games held?

**Temporal Context (Timestamp)**

- 2014
- 2015
- 2017
- 2021

**Answer**

Sochi

**Geographically Dependent Question**

How many gold medals did we win in the Winter Olympics 2018?

**Geographical Context (Location)**

- (US)
- (Norway)
- (Brazil)

**Answer**

- USA: 9
- Norway: 14
- Brazil: 0
Dataset: Sourcing Questions for SituatedQA

Temporal Dependence in QA

Temporally dependent questions are abundant in existing datasets

- Annotators provide the correct answer at the time of annotation
Dataset: Sourcing Questions for SituatedQA

**Temporal Dependence in QA**

*Temporally dependent questions* are *abundant* in existing datasets
- Annotators provide the correct answer at the time of annotation

**Geographically Dependence in QA**

*Geographically dependent questions* are *absent* in existing datasets
- There is no natural “correct” geographical context for annotators
Temporal Dependence in QA

Temporally dependent questions are **abundant** in existing datasets

- Annotators provide the correct answer at the time of annotation

Geographically Dependence in QA

Geographically dependent questions are **absent** in existing datasets

- There is no natural “correct” geographical context for annotators

To create such questions, we edit existing questions that specify a location

Who is the president of the US? → Who is the president?
Dataset: Geographical SituatedQA

We adopt a three stage pipeline for gathering annotations

0. Input Question
When was the last time states were created?
Dataset: Geographical SituatedQA

We adopt a three stage pipeline for gathering annotations

0. Input Question
When was the last time states were created?

1. Identification
Is this question geographically-dependent?
Yes
No
Dataset: Geographical SituatedQA

We adopt a three stage pipeline for gathering annotations

0. Input Question
When was the last time states were created?

1. Identification
Is this question geographically-dependent?
Yes
No

2. {Location / Answer} Collection
Identify locations and their corresponding answers

United States: 1959
Nigeria: 1996

When was the last time states were created?

Is this question geographically-dependent?
Yes
No

1959
1996
We adopt a three stage pipeline for gathering annotations.

0. Input Question
When was the last time states were created?

1. Identification
Is this question geographically-dependent?
Yes
No

2. {Location / Answer} Collection
Identify locations and their corresponding answers
- USA: 1959
- Nigeria: 1996

3. Validation
Is this {Location / Answer} pair correct?
Yes
No
We adopt a three stage pipeline for gathering annotations

0. Input Question
When was the last time states were created?

1. Identification
Is this question temporally-dependent?
Yes
No

2. Answer Timeline Collection
Create a timeline of the current and previous answers to the question

Previous Answer
Current Answer
Previous Answer Start Date
Current Answer Start Date

3. Validation
Is this answer timeline correct?
Yes
No
Mapping Answer Timelines to {Timestamps / Answers}

Static Questions

Static: Using temporally-independent questions by sampling a date after its original answer was collected

Example: Where is Belize as of 2020?
Mapping Answer Timelines to {Timestamps / Answers}

**Static Questions**

- **Data collection date**

**Current Answer**

- **Start Date**

**Previous Answer**

- **Start Date**

**Static: Using temporally-independent questions by sampling a date after its original answer was collected**

Example: Where is Belize as of 2020?

**Sampled: selecting a timestamp between an answer’s start and end date**

Example: Who was the most recently appointed supreme court justice as of March 25, 2019?

**Static Questions**

- **Using temporally-independent questions by sampling a date after its original answer was collected**

Example: Where is Belize as of 2020?

**Sampled: selecting a timestamp between an answer’s start and end date**

Example: Who was the most recently appointed supreme court justice as of March 25, 2019?
Mapping Answer Timelines to \{Timestamps / Answers\}

**Static: Using temporally-independent questions by sampling a date after its original answer was collected**

Example: Where is Belize as of 2020?

**Sampled: selecting a timestamp between an answer’s start and end date**

Example: Who was the most recently appointed supreme court justice as of March 25, 2019?

**Start: using an answer’s start date**

Example: Who was the most recently appointed supreme court justice as of October 26th 2020?
The SituatedQA Task

Temporally Dependent Question

Where were the last Winter Olympic Games held?

Geographical Context (Location)

Temporal Context (Timestamp)

Answer

Geographically Dependent Question

How many gold medals did we win in the Winter Olympics 2018?

Sourcing from NQ-Open, we create:

- **Temporal SituatedQA** containing 6.7K examples
- **Geographical SituatedQA** containing 5.5K examples
Experiments: Models

Closed-Book (BART)
I remember 🧠!
The answer is _____

Retrieval-Based (DPR)
You found me 🕵️!
The answer is _____
Experiments: Models

- **Zero-shot** models are trained on NQ-Open
- **Fine-tuned** models are first trained on NQ-Open, then also tuned on our collected training data
Experiments: Incorporating Context

Query modification is used to specify extra-linguistic context to a model.

- How many seasons are there for American Horror Story?
  - September, 2019
  - How many seasons are there for American Horror Story as of September, 2019?

- When was the last time states were created?
  - Nigeria
  - When was the last time states were created in Nigeria?
Results: Temporal SituatedQA

Closed-Book (BART)

Retrieval-Based (DPR)

Human Performance

NQ-Open Performance

<table>
<thead>
<tr>
<th>Type</th>
<th>Zero-Shot</th>
<th>Fine-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot</td>
<td>19.5</td>
<td>26.0</td>
</tr>
<tr>
<td>Fine-tuned</td>
<td>28.8</td>
<td>39.8</td>
</tr>
</tbody>
</table>

Answer Accuracy (EM)
Results: Temporal SituatedQA

**Closed-Book (BART)**

- Zero-Shot: 19.5
- Fine-tuned: 26.0

**Retrieval-Based (DPR)**

- Zero-Shot: 28.8
- Fine-tuned: 39.8

Human Performance: 17.2

NQ-Open Performance: 24.9

Legend:
- Static
- Start
- Sampled
Results: Temporal SituatedQA

Models struggle the most with contexts that are *between* answer transitions.
Results: Geographical SituatedQA

Closed-Book (BART)

Retrieval-Based (DPR)

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>Fine-tuned</th>
<th>Zero-Shot</th>
<th>Fine-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
<td>19.2</td>
<td>21.5</td>
<td>27.5</td>
<td>27.9</td>
</tr>
<tr>
<td>Rare</td>
<td>9.2</td>
<td>11.7</td>
<td>22.1</td>
<td>25.0</td>
</tr>
<tr>
<td>Overall</td>
<td>14.5</td>
<td>16.8</td>
<td>25.0</td>
<td>26.5</td>
</tr>
</tbody>
</table>

Answer Accuracy (EM)

- Zero-Shot
- Fine-tuned

Common, Rare, Overall
Results: Geographical SituatedQA

1. Models perform worse on rare locations

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>Fine-tuned</th>
<th>Zero-Shot</th>
<th>Fine-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Closed-Book</strong> (BART)</td>
<td>19.2</td>
<td>21.5</td>
<td>27.5</td>
<td>27.9</td>
</tr>
<tr>
<td><strong>Retrieval-Based</strong> (DPR)</td>
<td>9.2</td>
<td>11.7</td>
<td>22.1</td>
<td>25.0</td>
</tr>
<tr>
<td><strong>Common</strong></td>
<td>14.5</td>
<td>16.8</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td><strong>Rare</strong></td>
<td>11.7</td>
<td>16.8</td>
<td>25.0</td>
<td>26.5</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>16.8</td>
<td>17.7</td>
<td>26.5</td>
<td>26.5</td>
</tr>
</tbody>
</table>
Results: Geographical SituatedQA

Closed-Book (BART)

1. Models perform worse on rare locations
2. The performance gap between rare and common locations is less significant for retrieval-based methods

Retrieval-Based (DPR)

Answer Accuracy (EM)

- Zero-Shot: Δ = 10%
- Fine-tuned: Δ = 10%

- Zero-Shot: Δ = 5%
- Fine-tuned: Δ = 3%
Analysis: Geographical Bias

Q: What is the assumed context for a geographically-dependent question?

- We measure zero-shot performance on questions without the location.
- We separate the US and India, the two most frequent locations, from the remaining common locations.

![Chart showing answer accuracy for Closed-Book (BART) and Retrieval-Based (DPR) models.]

Answer Accuracy (EM)
Analysis: Geographical Bias

Q: What is the assumed context for a geographically-dependent question?

- We measure zero-shot performance on questions *without* the location.
- We separate the **US and India**, the two most frequent locations, from the remaining common locations.

*Models tend to assume the question is posed in the US.*
Analysis: Temporal Adaptation

Question

Do open-retrieval QA models that are trained on data collected in the past generalize to answering questions asked in the present?
Temporal Dependence of Models

SOTA QA models resolve context by using data from the same time period.

- Late 2018: NQ-open is collected
- Dec 2018: Wikipedia dump used by DPR
- Feb 2019: Pre-training data for LM

[Lazaridou et al, ArXiv 2021, Dhingra et al, ArXiv 2021]
SOTA QA models resolve context by using data from the same time period.

NQ-open is collected in late 2018.

Wikipedia dump used by DPR is collected in December 2018.

Pre-training data for LM is collected in February 2019.

Temporally dependence of models:

- SOTA QA models resolve context by using data from the same time period.
- NQ-open is collected in February 2021.
- SituatedQA data is collected in February 2021.

SituatedQA breaks this consistency.

[Lazaridou et al, ArXiv 2021, Dhingra et al, ArXiv 2021]
Temporal Dependence of Models

Example:
When was the last time the US hosted the Olympics?

Example:
What team does Jeremy Lin play for?

- Feb 2018: Wikipedia dump used by DPR
- Late 2018: NQ-open is collected
- Dec 2018: Pre-training data for LM
- Feb 2019: SituatedQA data is collected

[Lazaridou et al, ArXiv 2021, Dhingra et al, ArXiv 2021]
Analysis: Temporal Adaptation

Closed-Book (BART)

Answer Accuracy (EM)

- Zero-Shot
- Fine-Tuned

Stable
Updated

February 2018
Analysis: Temporal Adaptation

Closed-book models are unable to adapt to questions with updated answers.
Temporal Adaptation of Retrieval-based Systems

Stable

Updated

Dec 2018

Wikipedia dump used by DPR

Temporal Adaptation of Retrieval-based Systems

Analysis: Temporal Adaptation

**Closed-Book (BART)**

- Zero-Shot: 28.8
- Fine-Tuned: 28.1

**Retrieval-Based (DPR)**

- Zero-Shot: 27.3
- Zero-Shot (Updated): 11.5
- Fine-Tuned: 26.6

Answer Accuracy (EM)

- Stable
- Updated

February 2018

- Closed-Book (BART)
  - Zero-Shot: 5.3
  - Fine-Tuned: 6.7

- Retrieval-Based (DPR)
  - Zero-Shot: 7.2
  - Zero-Shot (Updated): 11.5
  - Fine-Tuned: 26.6

December 2018

- Closed-Book (BART)
  - Zero-Shot: 28.8
  - Fine-Tuned: 28.1

- Retrieval-Based (DPR)
  - Zero-Shot: 27.3
  - Zero-Shot (Updated): 11.5
  - Fine-Tuned: 26.6

February 2021
## Analysis: Temporal Adaptation

### Closed-Book (BART)

<table>
<thead>
<tr>
<th>Method</th>
<th>February 2018</th>
<th>December 2018</th>
<th>February 2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot</td>
<td>28.8</td>
<td>28.1</td>
<td>28.8</td>
</tr>
<tr>
<td>Fine-Tuned</td>
<td>5.3</td>
<td>6.7</td>
<td>17.2</td>
</tr>
</tbody>
</table>

### Retrieval-Based (DPR)

<table>
<thead>
<tr>
<th>Method</th>
<th>January 2018</th>
<th>December 2018</th>
<th>February 2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot</td>
<td>27.3</td>
<td>26.6</td>
<td>28.1</td>
</tr>
<tr>
<td>Zero-Shot</td>
<td>7.2</td>
<td>11.5</td>
<td>17.2</td>
</tr>
<tr>
<td>Fine-Tuned</td>
<td>11.5</td>
<td>28.1</td>
<td>17.2</td>
</tr>
</tbody>
</table>
Analysis: Temporal Adaptation

Retrieval-based methods also fail to generalize, even with access to an updated corpora

Closed-Book (BART)

Retrieval-Based (DPR)

<table>
<thead>
<tr>
<th></th>
<th>Zero-Shot</th>
<th>Zero-Shot</th>
<th>Fine-Tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>December 2018</td>
<td>27.3</td>
<td>26.6</td>
<td>28.1</td>
</tr>
<tr>
<td>February 2018</td>
<td>5.3</td>
<td>7.2</td>
<td>17.2</td>
</tr>
<tr>
<td>Updated</td>
<td>6.7</td>
<td>11.5</td>
<td>17.2</td>
</tr>
<tr>
<td>Stable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Updated</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Related Work:

**AmbigQA** explores ambiguity in QA, including ambiguity that cannot be resolved with context
  • Temporal-dependence is included under their definition of ambiguity

[Min et al., EMNLP 2020]

**CronQuestions** explores temporal dependence in knowledge-graph QA
  • Questions are generated from temporally-dependent relations within WikiData

[Saxena et al., ACL 2021]

**TempLama** explores creating temporally-aware pretrained language models
  • The authors train a version of T5 that utilizes temporal context (the year) during pretraining

[Dhingra et al., ArXiv 2021]

**GD-VCR** looks at geographical dependence in visual commonsense reasoning
  • This dataset explores forms of cultural and location-specific commonsense in Visual QA

[Yin et al., EMNLP 2021]
Data & code are available now!

Thank You!

SituatedQA: Incorporating Extra-Linguistic Contexts into QA

About

Answers to the same question may change depending on the extra-linguistic contexts (when and where the question was asked). To study this challenge, we introduce SituatedQA, an open-retrieval QA dataset where systems must produce the correct answer to a question given the temporal or geographical context. To construct SituatedQA we first identify such questions in existing QA datasets. We find that a significant proportion of information seeking questions have context-dependent answers (e.g., roughly 16.5% of NQ-Open). For such context-dependent questions, we then crowdsource alternative contexts and their corresponding answers. Our study shows that existing models struggle with questions where the context is not explicitly provided.

Previous Answer: Moderna, Pfizer

Current Answer: Moderna, Pfizer, J&J

Context Type: Temporal

Question: Which COVID-19 vaccines have been authorized for adults in the US?

Context: Dec 18, 2020

Answer: Moderna, Pfizer

Context: Apr 10, 2021

Answer: Moderna, Pfizer, J&J

SituatedQA.github.io