Understanding Dataset Design Choices for Multi-hop Reasoning

Jifan Chen and Greg Durrett
The University of Texas at Austin
Multi-hop Reasoning

**Question**: The Oberoi family is part of a hotel company that has a head office in what city?

**Doc 1**: The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

**Doc 2**: The Oberoi Group is a hotel company with its head office in Delhi.

Example picked from HotpotQA (Yang 2018)
Multi-hop Reasoning

**Question:** The Oberoi family is part of a hotel company that has a head office in what city?

**Doc 1**
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

**Doc 2**
The Oberoi Group is a hotel company with its head office in Delhi. ...
The Oberoi family is part of a hotel company that has a head office in what city?

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group. …

The Oberoi Group is a hotel company with its head office in Delhi. …

Example picked from HotpotQA (Yang 2018)
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.

The Oberoi Group is a hotel company with its head office in Delhi.

Question: The Oberoi family is part of a hotel company that has a head office in what city?

Same entity
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.

The Oberoi Group is a hotel company with its head office in Delhi.

Example picked from HotpotQA (Yang 2018)
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group...

The Oberoi Group is a hotel company with its head office in Delhi.

---

**Question:** The Oberoi family is part of a hotel company that has a head office in what city?

---

Example picked from HotpotQA (Yang 2018)
Multi-hop Reasoning

Question: The Oberoi family is part of a hotel company that has a head office in what city?

Doc 1: The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

Doc 2: The Oberoi Group is a hotel company with its head office in Delhi. ...

This is an idealized version of multi-hop reasoning. Do models need to do this to do well on this task?

Example picked from HotpotQA (Yang 2018)
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group. The Oberoi Group is a hotel company with its head office in Delhi.
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.

Question: The Oberoi family is part of a hotel company that has a head office in what city?

The Oberoi Group is a hotel company with its head office in Delhi.

High lexical overlap

Ignore the bridging entity and directly jump to the answer!

Example picked from HotpotQA (Yang 2018)
Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Multi-hop Reasoning

**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Doc 1**
Shirley Temple Black was an American actress, businesswoman, and singer ...
As an adult, she served as Chief of Protocol of the United States ...

**Doc 2**
Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer ...

**Doc 3**
Meet Corliss Archer is an American television sitcom that aired on CBS ...

Example picked from HotpotQA (Yang 2018)
Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Doc 1: Shirley Temple Black was an American actress, businesswoman, and singer. As an adult, she served as Chief of Protocol of the United States.

Doc 2: Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.

Doc 3: Meet Corliss Archer is an American television sitcom that aired on CBS.
Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Doc 1: Shirley Temple Black was an American actress, businesswoman, and singer ... As an adult, she served as Chief of Protocol of the United States ...

Doc 2: Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer ...

Doc 3: Meet Corliss Archer is an American television sitcom that aired on CBS ...

Example picked from HotpotQA (Yang 2018)
**Question:** What government position was held by the woman who portrayed Corliss Archer in the film *Kiss and Tell*?

**Doc 1:** Shirley Temple Black was an American actress, businesswoman, and singer ... As an adult, she served as Chief of Protocol of the United States ...

**Doc 2:** *Kiss and Tell* is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer ...

**Doc 3:** Meet Corliss Archer is an American television sitcom that aired on CBS ...

Example picked from HotpotQA (Yang 2018)
Multi-hop Reasoning

**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Doc 1:** Shirley Temple Black was an American actress, businesswoman, and singer ... As an adult, she served as Chief of Protocol of the United States ... Same entity

**Doc 2:** Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer ... Same entity

**Doc 3:** Meet Corliss Archer is an American television sitcom that aired on CBS ...

Example picked from HotpotQA (Yang 2018)
**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Doc 1:** Shirley Temple Black was an American actress, businesswoman, and singer ...

As an adult, she served as Chief of Protocol of the United States ...

**Doc 2:** Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.

**Doc 3:** Meet Corliss Archer is an American television sitcom that aired on CBS ...

Example picked from HotpotQA (Yang 2018)
Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Doc 1: Shirley Temple Black was an American actress, businesswoman, and singer ...
As an adult, she served as Chief of Protocol of the United States ...

Doc 2: Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer ...

Doc 3: Meet Corliss Archer is an American television sitcom that aired on CBS ...

No simple lexical overlap.

Example picked from HotpotQA (Yang 2018)
Multi-hop Reasoning

**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Doc 1**
Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States …

**Doc 2**
Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer …

**Doc 3**
Meet Corliss Archer is an American television sitcom that aired on CBS …
Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Doc 1: Shirley Temple Black was an American actress, businesswoman, and singer. As an adult, she served as Chief of Protocol of the United States.

Doc 2: Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.

Doc 3: Meet Corliss Archer is an American television sitcom that aired on CBS.

Example picked from HotpotQA (Yang 2018)
Multi-hop Reasoning

**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Doc 1**
Shirley Temple Black was an American actress, businesswoman, and singer...

As an adult, she served as Chief of Protocol of the United States...

**Doc 2**
Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer...

**Doc 3**
Meet Corliss Archer is an American television sitcom that aired on CBS...

Only one government position appears in the context!
Multi-hop Reasoning Datasets

How much multi-hop reasoning is really needed in the recently proposed multi-hop oriented datasets?

WikiHop (Welbl et al., 2018)  
HotpotQA (Yang et al., 2018)
# Multi-hop Reasoning Datasets

How much multi-hop reasoning is really needed in the recently proposed multi-hop oriented datasets?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Multiple-choice</th>
<th>Span-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiHop (Welbl et al., 2018)</td>
<td>Multiple-choice</td>
<td></td>
</tr>
<tr>
<td>HotpotQA (Yang et al., 2018)</td>
<td>Span-based</td>
<td></td>
</tr>
</tbody>
</table>
Multi-hop Reasoning Datasets

How much multi-hop reasoning is really needed in the recently proposed multi-hop oriented datasets?

WikiHop (Welbl et al., 2018)
- Multiple-choice
- Final answer is given as supervision

HotpotQA (Yang et al., 2018)
- Span-based
- Answer and supporting facts
Research Questions
Research Questions

Can a model identify the answer with only a set of candidates?

Government position → Chief of Protocol, actress, singer
Research Questions

Can a model identify the answer with only a set of candidates?

Government position ➔ Chief of Protocol, actress, singer

Can a model identify where the answer is in a single hop?

Oberoi Family ➔ Delhi
Research Questions

Can a model identify the answer with only a set of candidates?
- Government position ➔ Chief of Protocol, actress, singer

Can a model identify where the answer is in a single hop?
- Oberoi Family ➔ Delhi

Does the type of supervision matter?
- Span-based vs. Multiple-choice
Research Questions

Can a model identify the answer with only a set of candidates?
  Government position ➔ Chief of Protocol, actress, singer

Can a model identify where the answer is in a single hop?
  Oberoi Family ➔ Delhi

Does the type of supervision matter?
  Span-based vs. Multiple-choice
**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Answer:** Shirley Temple Black was an American actress, businesswoman, and singer. As an adult, she served as Chief of Protocol of the United States. 

- **Doc 1:** Shirley Temple Black was an American actress, businesswoman, and singer. As an adult, she served as Chief of Protocol of the United States.
- **Doc 2:** Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
- **Doc 3:** Meet Corliss Archer is an American television sitcom that aired on CBS.

**Chief of Protocol, businesswoman, ... actress**

*Kaushik and Lipton (2018)*
Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Results on WikiHop
Results on WikiHop

- Accuracy
  - Majority-candidate: 38.8
  - BiDAF: 42.9
  - Entity-GCN: 64.8
  - CFC: 66.4
  - BAG: 67.4

- Weak baselines vs. state-of-the-art
Results on WikiHop

- **Majority-candidate**: 38.8
- **BiDAF**: 42.9
- **Entity-GCN**: 64.8
- **CFC**: 66.4
- **BAG**: 67.4
- **NoContext**: 59.3

The bar chart compares different models on WikiHop, with **state-of-the-art** models outperforming **weak baselines**.
Results on WikiHop

More than half of questions can be answered without even using the context!
Results on WikiHop

More than half of questions can be answered without even using the context!

- SOTA models trained on this may learn question-answer correspondences, not multi-hop reasoning as advertised.
Research Questions

Can a model identify the answer with only a set of candidates?
Government position → Chief of Protocol, actress, singer

Can a model identify where the answer is in a single hop?
Oberoi Family → Dehli

Does the type of supervision matter?
Span-based vs. Multiple-choice
Finding the answer directly

**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

- **Doc 1:** Shirley Temple Black was an American actress, businesswoman, and singer. As an adult, she served as Chief of Protocol of the United States.

- **Doc 2:** Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.

- **Doc 3:** Meet Corliss Archer is an American television sitcom that aired on CBS.

**Options:**
- A: Chief of Protocol
- B: Businesswoman
- C: Actress

Example picked from HotpotQA (Yang 2018)
Finding the answer directly

**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

---

**Doc 1:** Shirley Temple Black was an American actress, businesswoman, and singer...
As an adult, she served as Chief of Protocol of the United States.

**Doc 2:** Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.

**Doc 3:** Meet Corliss Archer is an American television sitcom that aired on CBS...

---

A: Chief of Protocol  
B: Businesswoman  
C: Actress

There is no candidate set on HotpotQA

Example picked from HotpotQA (Yang 2018)
Find the answer by comparing each sentence with the question **separately**!

**Question:** The Oberoi family is part of a hotel company that has a head office in what city?

Doc 1
The Oberoi family is an Indian family that is ...

Doc 2
The Oberoi Group is a hotel company with its head office in Delhi.

Doc 3
Future Fibre Technologies a fiber technologies company ...
The Oberoi family... what city?

Answer prediction: Delhi

The Oberoi Group... in Delhi.

Future Fibre Technologies is a fibre...

The Oberoi family...
The Oberoi family

Future Fibre Technologies is a fibre...

The Oberoi family ... what city?

Softmax over all sentences is the only cross-sentence interaction

Answer prediction:

Delhi
<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF++</td>
<td>50.8</td>
</tr>
<tr>
<td>QFE</td>
<td>69.7</td>
</tr>
<tr>
<td>GRN</td>
<td>69.0</td>
</tr>
<tr>
<td>DFGN</td>
<td>68.1</td>
</tr>
<tr>
<td>Sentence Factored</td>
<td>58.7</td>
</tr>
</tbody>
</table>

Results on HotpotQA
Results on HotpotQA

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF++</td>
<td>58.7</td>
</tr>
<tr>
<td>QFE</td>
<td></td>
</tr>
<tr>
<td>GRN</td>
<td></td>
</tr>
<tr>
<td>DFGN</td>
<td></td>
</tr>
</tbody>
</table>

Sentence Factored
Results on HotpotQA

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF++</td>
<td>58.7</td>
</tr>
<tr>
<td>QFE</td>
<td>68.1</td>
</tr>
<tr>
<td>GRN</td>
<td>69.0</td>
</tr>
<tr>
<td>DFGN</td>
<td>69.7</td>
</tr>
</tbody>
</table>
Results on HotpotQA

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF++</td>
<td>58.7</td>
</tr>
<tr>
<td>QFE</td>
<td>68.1</td>
</tr>
<tr>
<td>GRN</td>
<td>69.0</td>
</tr>
<tr>
<td>DFGN</td>
<td>69.7</td>
</tr>
<tr>
<td>Sentence Factored</td>
<td>50.8</td>
</tr>
</tbody>
</table>
A simple single sentence reasoning model could solve more than half questions on HotpotQA.
The Oberoi family … what city?

Future Fibre Technologies …

The Oberoi Group … in Delhi.

Containing answer: 1
Sentence Factored Model

Containing answer:

1

Predict which sentence contains the answer

The Oberoi family ... what city?

The Oberoi Group ... in Delhi.

Future Fibre Technologies ...

The Oberoi family ...
Sentence Factored Model

- Predict which sentence contains the answer
- A simpler way to test how much multi-hop reasoning is needed

Containing answer:

1

The Oberoi family ...

Future Fibre Technologies ...

The Oberoi family ...

The Oberoi family ... what city?
Answer Sentence Prediction Results

<table>
<thead>
<tr>
<th></th>
<th>SQuAD</th>
<th>WikiHop</th>
<th>HotpotQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence-factored</td>
<td>88</td>
<td>66.1</td>
<td>57.2</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>22.1</td>
<td>6.5</td>
</tr>
</tbody>
</table>
Answers can be found for many questions without a multi-hop process.
Research Questions

Can a model identify the answer with only a set of candidates?
Government position  →  Chief of Protocol, actress, singer

Can a model identify where the answer is in a single hop?
Oberoi Family  →  Delhi

Does the type of supervision matter?
Span-based vs. Multiple-choice
Making HotpotQA Multiple-Choice

**Question:** The Oberoi family is part of a hotel company that has a head office in what city?

**Doc 1**
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.

**Doc 2**
The Oberoi Group is a hotel company with its head office in Delhi.

**Doc 3**
Future Fibre Technologies is a fiber technologies company with its US head office in Mountain View, Middle East, and its head office in Dubai.
Making HotpotQA Multiple-Choice

**Question**: The Oberoi family is part of a hotel company that has a head office in what city?

- **Doc 1**: The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …
- **Doc 2**: The Oberoi Group is a hotel company with its head office in Delhi. …
- **Doc 3**: Future Fibre Technologies a fiber technologies company with its US head office in Mountain View, Middle East head office in Dubai …

Randomly choose entities in the documents as candidates.

The Oberoi family, The Oberoi Group, Delhi, Future Fibre Technologies, US, Mountain View, Middle East, Dubai
Results of HotpotQA Multiple-Choice

Accuracy vs. Number of candidates

- **NoContext**
- **BiDAF++**

Accuracy decreases as the number of candidates increases.
Results of HotpotQA Multiple-Choice

Accuracy vs. Number of candidates

- **NoContext**
- **BiDAF++**
Results of HotpotQA Multiple-Choice

Accuracy vs Number of candidates

- NoContext
- BiDAF++

Accuracy:
- NoContext: 2.0 at 10 candidates, 2.9 at 20 candidates
- BiDAF++: 2.9 at 10 candidates, 2.0 at 20 candidates
Results of HotpotQA Multiple-Choice

<table>
<thead>
<tr>
<th>Number of candidates</th>
<th>NoContext</th>
<th>BiDAF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.0</td>
<td>2.9</td>
</tr>
<tr>
<td>20</td>
<td>3.8</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy vs. Number of candidates graph
Results of HotpotQA Multiple-Choice

Accuracy vs Number of candidates

- **NoContext**
  - 10 candidates: 2.0
  - 20 candidates: 2.9
  - 30 candidates: 3.8
  - 40 candidates: 2.5
  - 50 candidates: (data point not shown)

- **BiDAF++**
  - 10 candidates: (data point not shown)
  - 20 candidates: (data point not shown)
  - 30 candidates: (data point not shown)
  - 40 candidates: 2.5
  - 50 candidates: (data point not shown)
Results of HotpotQA Multiple-Choice

NoContext vs BiDAF++

Accuracy vs Number of candidates

- NoContext:
  - 10 candidates: 2.0
  - 20 candidates: 2.9
  - 30 candidates: 3.8
  - 40 candidates: 2.5
  - 50 candidates: 2.5

- BiDAF++:
  - 10 candidates: 2.0
  - 20 candidates: 2.9
  - 30 candidates: 3.8
  - 40 candidates: 2.5
  - 50 candidates: 2.5
Results of HotpotQA Multiple-Choice

More candidates does make the prediction harder, but BiDAF++ does not learn more than the NoContext baseline.
Can a model identify the answer with only a set of candidates?

Can a model identify where the answer is in a single hop?

Does the type of supervision matter?
Conclusions

Can a model identify the answer with only a set of candidates?

- Models can achieve high performance by “cheating” through extra candidate information.

Can a model identify where the answer is in a single hop?

Does the type of supervision matter?
Conclusions

Can a model identify the answer with only a set of candidates?

- Models can achieve high performance by “cheating” through extra candidate information.

Can a model identify where the answer is in a single hop?

- More than half of the questions in WikiHop and HotpotQA can be answered in a single hop.

Does the type of supervision matter?
Conclusions

Can a model identify the answer with only a set of candidates?
- Models can achieve high performance by “cheating” through extra candidate information.

Can a model identify where the answer is in a single hop?
- More than half of the questions in WikiHop and HotpotQA can be answered in a single hop.

Does the type of supervision matter?
- Span-based datasets may be less hackable.
Conclusions

Can a model identify the answer with only a set of candidates?

- Models can achieve high performance by “cheating” through extra candidate information.

Can a model identify where the answer is in a single hop?

- More than half of the questions in WikiHop and HotpotQA can be answered in a single hop.

Does the type of supervision matter?

- Span-based datasets may be less hackable.

*No results for WikiHop-Masked or HotpotQA fullwiki settings
Takeaways

When constructing multi-hop based dataset:
Takeaways

When constructing multi-hop based dataset:

- If using multiple-choice supervision, check the no-context baseline to see whether there are strong correlations between question and candidates.
Takeaways

When constructing multi-hop based dataset:

- If using multiple-choice supervision, check the no-context baseline to see whether there are strong correlations between question and candidates.

- Do an adversarial test using a sentence-factored model.
Takeaways

When constructing multi-hop based dataset:

- If using multiple-choice supervision, check the no-context baseline to see whether there are strong correlations between question and candidates.

- Do an adversarial test using a sentence-factored model.

Thank you!

Q & A