**Question Answering**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is the Victoria and Albert Museum located?</td>
<td>The V&amp;A is located in the Brompton district of the Royal Borough of Kensington and Chelsea. It is located in London.</td>
</tr>
<tr>
<td>How many permanent objects are there?</td>
<td>A permanent collection of over 4.5 million objects.</td>
</tr>
<tr>
<td>When was the Victoria and Albert museum founded?</td>
<td>It was founded in 1852.</td>
</tr>
<tr>
<td>Who is the museum named for?</td>
<td>Named after Queen Victoria and Prince Albert.</td>
</tr>
<tr>
<td>In which London borough is the Victoria and Albert Museum located?</td>
<td>Brompton district of the Royal Borough of Kensington and Chelsea.</td>
</tr>
</tbody>
</table>

[Rajpurkar et al 16]  

**paragraph-level**
Question Answering

Paragraph-level [Rajpurkar et al 16]

Corpus-level [TREC QA dataset]
Question Answering

[Paragraph-level]

[Rajpurkar et al 16]

[Corpus-level]

[TREC QA dataset]

Single long document
Research Question

• Given a question and a **long document**, how can we efficiently find an answer?

• State-of-the-art recurrent neural network is inappropriate to handle long document:
  
  • **Speed**: Sequential processing
  
  • **Effectiveness**: Often forgets earlier sentences
Question Answering
How long do effects of lorazepam last?
How long do effects of lorazepam last?
Question Answering

Reading document closely from start to the end is probably NOT the best strategy.

Can we search for relevant sentences and read them more carefully?
How long do effects of lorazepam last?
Task

How long do effects of lorazepam last?

Lorazepam is a benzodiazepine antianxiety medication that is used for short-term treatment of anxiety disorders and restless legs syndrome. The half-life of lorazepam is about 9 hours, meaning that it takes approximately nine hours for the body to reduce the concentration of the drug by half. However, the duration of the effects depends on various factors, such as individual response, dose, and route of administration.

Peak effects roughly coincide with peak serum levels, which occur 10 minutes after intravenous injection, up to 50 minutes after intramuscular injection, and 90 to 120 minutes after oral administration, but initial effects will be noted before this. A clinically relevant lorazepam dose will normally be effective for six to 12 hours, making it unsuitable for regular once-daily administration, so it is usually prescribed as two to four daily doses when taken regularly, but this may be extended to five or six, especially in the case of elderly patients.
This Work

Query (x) → Sentence Selection → Sentence set (\(\hat{\mathcal{S}}\)) → Answer Generation → Answer (y)

Document (d) → Sentence Selection
This Work

Query (x)

Sentence Selection

Sentence set (\(\hat{S}\))

Answer Generation

Answer (y)

Document (d)
This Work

- Coarse-to-Fine model for question answering
- Substantially faster (up to 6.7 times) with comparable accuracies
- Learning without direct supervision for evidence sentence
Related Work

• Coarse-to-Fine model for different applications
  (Charniak et al, 06, Cheng and Lapata, 16, Yang et al, 16, Lei et al, 16)

• Question answering
  (Servelyn and Mochitti 15, Yang et al, 16, Jurczyk et al, 16, dos Santos et al, 16, Sultan et al 16, Chen et al, 17)

• New datasets with multiple documents
  (SearchQA[Dunn et al 17], MS MARCO[Nguyen et al 16], TriviaQA[Joshi et al 17])
Data

• WikiReading (Hewlett et al, ACL 16)
  • Wikipedia InfoBox

• WikiReading-Long
  • Challenging WikiReading subset, longer documents

• WikiSuggest
  • Query suggest from Google, answered by Google snippets
WikiReading

- Taken from Wikipedia.
- Infobox properties and article.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Property</th>
<th>Document</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folkart Towers</td>
<td>Country</td>
<td>Folkart Towers are twin skyscrapers in Turkish city of Izmir.</td>
<td>Turkey</td>
</tr>
<tr>
<td>Canada</td>
<td>Located next to body of water</td>
<td>Canada is a country… extended from the Atlantic to the Pacific and northward into the Artic Ocean</td>
<td>Atlantic Ocean, Pacific, Arctic Ocean</td>
</tr>
<tr>
<td>Breaking Bad</td>
<td>Start time</td>
<td>Breaking Bad is a TV series… from January 20, 2008</td>
<td>20 January 2008</td>
</tr>
</tbody>
</table>
WikiReading

- Taken from Wikipedia.
- Infobox properties and article.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Property</th>
<th>Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folkart Towers</td>
<td>Country</td>
<td>Turkish city of Izmir</td>
</tr>
<tr>
<td>Canada</td>
<td>Located next to</td>
<td>body of water</td>
</tr>
<tr>
<td></td>
<td>body of water</td>
<td>Canadian continent extended from the Atlantic to the Pacific and northward into the Artic Ocean</td>
</tr>
<tr>
<td>Breaking Bad</td>
<td>Start time</td>
<td>Breaking Bad is a TV series… from January 20, 2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20 January 2008</td>
</tr>
</tbody>
</table>

✓ Large-scale dataset covering various domains

✓ Most documents are relatively short (480 words)
WikiReading-Long

- Pruned to have documents with at least 10 sentences
- Contains 1.97 million instances (~ 15% of original data)
- Single document contains 1200 words on average
WikiSuggest

• Question from Google’s user queries
• Answers from Google’s auto suggested answer
• Document from Wikipedia

Who is the first president of united states

President of the United States (1)

George Washington
WikiSuggest Examples

<table>
<thead>
<tr>
<th>Query</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>how many officials in a NFL football game</td>
<td>seven officials</td>
</tr>
<tr>
<td>the 11th tarot card</td>
<td>Major Arcana</td>
</tr>
<tr>
<td>what age Ronald Reagan become president</td>
<td>69 years</td>
</tr>
<tr>
<td>Ohio basketball coach</td>
<td>Saul Phillips</td>
</tr>
<tr>
<td>how old is Ed Marinaro</td>
<td>born on March 31, 1950</td>
</tr>
<tr>
<td>allers syndrome</td>
<td>Ehlers-Danlos</td>
</tr>
</tbody>
</table>
## WikiSuggest Examples

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many officials in a NFL football game</td>
<td>Seven officials</td>
</tr>
<tr>
<td>How old is Ronald Reagan become president</td>
<td>69 years</td>
</tr>
<tr>
<td>What is the Ohio basketball coach</td>
<td>Saul Phillips</td>
</tr>
<tr>
<td>How old is Ed Marinaro</td>
<td>Born on March 31, 1950</td>
</tr>
<tr>
<td>Ehlers-Danlos syndrome</td>
<td>Ehlers-Danlos</td>
</tr>
</tbody>
</table>

- **✓** Large-scale, noisy dataset covering various domain (3.5M)
- **✓** More diverse and natural questions
- **✓** Including systematically generated noise (~25%)
## Dataset Summary

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Examples</th>
<th># Unique Queries</th>
<th># of tokens / doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiReading</td>
<td>18M</td>
<td>867</td>
<td>0.5K</td>
</tr>
<tr>
<td>WikiReadingLong</td>
<td>2M</td>
<td>239</td>
<td>1.2K</td>
</tr>
<tr>
<td>WikiSuggest</td>
<td>3.5M</td>
<td>3.5M</td>
<td>5.9K</td>
</tr>
</tbody>
</table>
Model

Query (x)

Sentence Selection

Sentence set ($\hat{s}$)

Answer Generation

Answer (y)

Document (d)
Model

Query ($x$) → Sentence Selection → Sentence set ($\hat{s}$) → Answer Generation → Answer ($y$)

Document ($d$)

Cheap, fast model picking sentences relevant to the query

Expensive model (RNN) generating answers
Sentence Selection Model

- Take query and document as input
- Computes relevance score for each sentence ($P(s|x, d)$) to generate sentence set to pass on to answer generation model.
- Coarse and fast sentence representation (BoW)
Answer Generation Model
(Hewlett et al, ACL16)

• Given a sentence set and a query, generate answer strings

• RNN encoder-decoder model with placeholders
Learning

We do not have supervision for which sentences contain the information.
Can we know which sentence contains answer?

- Good heuristics:
  - Sentence with an answer string is the sentence that you should pay close attention to.

<table>
<thead>
<tr>
<th>Q: Where did Alexandro Friedmann die?</th>
<th>A: St. Petersburg</th>
<th>S: Alexandro Friedmann was born in St. Petersburg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: country</td>
<td>A: Turkey</td>
<td>S: Folkart Towers are twin skyscrapers in Turkish city of Izmir.</td>
</tr>
</tbody>
</table>

False Negative!
False Positive!
### Answer String Match Statistics

<table>
<thead>
<tr>
<th></th>
<th>Answer String Exists</th>
<th>Avg. # of Answer Match</th>
<th>Answer in First Sentence if answer exists</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiReading</td>
<td>47.1%</td>
<td>1.22</td>
<td>75.1%</td>
</tr>
<tr>
<td>WikiReading Long</td>
<td>50.4%</td>
<td>2.18</td>
<td>31.3%</td>
</tr>
<tr>
<td>WikiSuggest</td>
<td>100.0%</td>
<td>13.95</td>
<td>33.6%</td>
</tr>
</tbody>
</table>

False Negative

False Positive
Pipeline (Distant Supervision)

- Separate objective for two models

\[ \log p(s^* | x, d) + \log p(y^* | x, s^*) \]

- Gold sentence (\( s^* \)): First sentence with an answer string match or first sentence if answer string match does not exist.
Soft Attention

- Make a “blended” token representation by merging each sentence token weighted by its relevancy score $p(s|x)$.
- Allows end-to-end learning.

$$\log p(y^*|x, d) = \log p(y^*|x, \hat{s})$$
Hard Attention
(Reinforcement Learning)

• Action: Choosing a sentence

• Reward: Log probability of answer with chosen sentence

\[ R(s) = \log P(y^*|s, x) \]

\[ E[R] = \sum_s P(s|x) \cdot R(s) \]

\[ = \sum_s P(s|x) \cdot \log P(y^*|s, x) \]

• Can approximate the gradient with sampling (REINFORCE)

\[ \nabla \log P(y^*|\tilde{s}, x) + \log P(y^*|\tilde{s}, x) \cdot \nabla \log P(\tilde{s}|x) \]
Hard Attention
(Reinforcement Learning)

- Can be flexible on the number of sentences to pass on to the answer generation model

- Curriculum learning (Ross et al, AISTAT11)

  - Trained with pipeline objective at the beginning
Evaluation

• Answer accuracy:
  • exact match accuracy

• Efficiency:
  • time to finish document encoding
Comparison Systems

• First sentence baseline

• Answer generation baseline:
  • Input is the first 300 tokens.

• Heuristic oracle:
  • Input is the sentence with answer string or the first sentence when there is no answer match.
Accuracy Results

Accuracy Results for WikiReading, WikiReading Long, and WikiSuggest datasets for different models:

- WikiReading:
  - First: 71
  - SoftAttend: 71.6
  - Pipeline: 72.4
  - Reinforce: 73.9

- WikiReading Long:
  - First: 26.7
  - SoftAttend: 38.3
  - Pipeline: 36.8
  - Reinforce: 42.2

- WikiSuggest:
  - First: 44
  - SoftAttend: 45.4
  - Pipeline: 45.4
  - Reinforce: 45.8

Accuracy values are presented for each dataset and model.
Accuracy Results

Accuracy:

25 36 47 58 69 80

Dataset:

WikiReading  WikiReading-Long  WikiSuggest

First  SoftAttend  Pipeline  Reinforce  AnswerGen Only  Heuristic Oracle

Accuracy:

First:

WikiReading: 71 71.6 72.4 73.9 75.6 74.6

WikiReading-Long: 38.3 36.8 42.2 40.1 43.9 44

WikiSuggest: 26.7 46.7

SoftAttend:

WikiReading: 71 71.6 72.4 73.9 75.6 74.6

WikiReading-Long: 38.3 36.8 42.2 40.1 43.9 44

WikiSuggest: 26.7 46.7

Pipeline:

WikiReading: 71 71.6 72.4 73.9 75.6 74.6

WikiReading-Long: 38.3 36.8 42.2 40.1 43.9 44

WikiSuggest: 26.7 46.7

Reinforce:

WikiReading: 71 71.6 72.4 73.9 75.6 74.6

WikiReading-Long: 38.3 36.8 42.2 40.1 43.9 44

WikiSuggest: 26.7 46.7

AnswerGen Only:

WikiReading: 71 71.6 72.4 73.9 75.6 74.6

WikiReading-Long: 38.3 36.8 42.2 40.1 43.9 44

WikiSuggest: 26.7 46.7

Heuristic Oracle:

WikiReading: 71 71.6 72.4 73.9 75.6 74.6

WikiReading-Long: 38.3 36.8 42.2 40.1 43.9 44

WikiSuggest: 26.7 46.7
Speed Results

Batch Size

- AnswerGen Only
- Reinforce (K=2)
- Reinforce (K=1)

# of examples / seconds

- Batch Size 1: 7.52, 26.4, 50.6
- Batch Size 8: 30.88, 101.45, 184.56

AnswerGen Only
Reinforce (K=2)
Reinforce (K=1)
Speed Results

Up to 6.7 times speed-up!
## Error Analysis

<table>
<thead>
<tr>
<th>Error Type</th>
<th>WikiReading-Long</th>
<th>WikiSuggest</th>
</tr>
</thead>
<tbody>
<tr>
<td>No evidence in document</td>
<td>58%</td>
<td>16%</td>
</tr>
<tr>
<td>Error in answer generation</td>
<td>26%</td>
<td>30%</td>
</tr>
<tr>
<td>Error in sentence selection</td>
<td>16%</td>
<td>6%</td>
</tr>
<tr>
<td>Noisy QA pairs</td>
<td>0%</td>
<td>48%</td>
</tr>
</tbody>
</table>
Conclusion

• Coarse-to-Fine model for question answering

• Efficient model (up to 6.7 times speed up) with comparable accuracies

• Learning strategy without direct supervision for evidence sentence
Future Work

• Introduce more flexible document selection

• Extend to more challenging dataset (multi document setting)
Thank you!

Questions?
Sentence Selection Accuracy

- **WikiReading Long**
  - First: 31.3
  - SoftAttend: 70
  - Pipeline: 74.6
  - Reinforce: 74.4

- **WikiSuggest**
  - First: 42.6
  - SoftAttend: 49.9
  - Pipeline: 67.5
  - Reinforce: 67.3