Towards the **Machine Comprehension** of Text

Danqi Chen

Joint work with Christopher Manning, Jason Bolton, Adam Fisch, Antoine Bordes, Jason Weston

April 10, 2017
Towards the **Machine Comprehension of Text**

Towards the **Machine Comprehension of Text**: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013
Machine Comprehension

“A machine comprehends a passage of text if, for any question regarding that text that can be answered correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question.”
Machine Comprehension

Passage ($P$) + Question ($Q$) $\rightarrow$ Answer ($A$)
Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house……..

What city is Alyssa in?
Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

What city is Alyssa in? A Miami
Till 2015, we really haven’t had any statistical NLP systems that can understand even such simple passages.
Dataset vs. Model
Dataset vs. Model

Before 2015:
• MCTest (Richardson et al, 2013): 2600 questions
• ProcessBank (Berant et al, 2014): 500 questions
Dataset vs. Model

Before 2015:

• MCTest (Richardson et al, 2013): 2600 questions
• ProcessBank (Berant et al, 2014): 500 questions

After 2015:

• CNN/Daily Mail
• Children Book Test
• WikiReading
• LAMBADA
• SQuAD
• Who did What
• NewsQA
• MS MARCO
Dataset vs. Model

Before 2015:
- MCTest (Richardson et al, 2013): 2600 questions
- ProcessBank (Berant et al, 2014): 500 questions

After 2015:
- CNN/Daily Mail
- Children Book Test
- WikiReading
- LAMBADA
- SQuAD
- Who did What
- NewsQA
- Maluuba
- MS MARCO

More than 100k questions!
Dataset vs. Model

Before 2015:

- Lexical matching
- Logistic regression
Dataset vs. Model

Before 2015:
• Lexical matching
• Logistic regression

After 2015:
• Neural networks
Dataset vs. Model

**Before 2015:**
- Lexical matching
- Logistic regression

**After 2015:**
- **Attentive** Reader
- **Memory** Networks
- Gated-**attention** Reader
- ReasoNet
- **Match**-LSTM
- **Attention** Sum Reader
- **Attention**-over-**Attention** Reader
- Iterative **Attentive** Reader
- Dynamic **coattention** networks
- Bi-directional **Attention** Flow Network
- Multi-Perspective Context Matching
- ...
(Dhingra et al, 2016; Seo et al, 2017; Xiong et al, 2017)
(Dhingra et al., 2016; Seo et al., 2017; Xiong et al., 2017)
(Dhingra et al, 2016; Seo et al, 2017; Xiong et al, 2017)
This Talk
This Talk

- Our case studies in the **CNN/Daily Mail** Datasets and **Stanford Question Answering Dataset (SQuAD)**

What models do we actually need?
This Talk

- Our case studies in the **CNN/Daily Mail** Datasets and **Stanford Question Answering Dataset (SQuAD)**
  - What models do we actually need?

- **Open-domain** question answering using **full Wikipedia**
  - Are current MC models useful?
This Talk

• Our case studies in the CNN/Daily Mail Datasets and Stanford Question Answering Dataset (SQuAD)
  What models do we actually need?

• Open-domain question answering using full Wikipedia
  Are current MC models useful?

• Discussion: What have we achieved today? What are we still missing? What shall we do next?
CNN/Daily Mail Datasets

Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in "Star Wars" movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.
CNN/Daily Mail Datasets

**Story highlights**

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in "Star Wars" movies have gradually become more diverse

**CNN** — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.
CNN/Daily Mail Datasets

(Hermann et al, 2015)

Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in "Star Wars" movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.
if you feel a ripple in the force today, it may be the news that the official @entity6 is getting its first gay character. According to the sci-fi website @entity9, the upcoming novel "@entity11" will feature a capable but flawed @entity13 official named @entity14 who also happens to be a lesbian. The character is the first gay figure in the official @entity6 -- the movies, television shows, comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24, editor of "@entity6".

Characters in "@placeholder" movies have gradually become more diverse.
( @entity4 ) if you feel a ripple in the force today, it may be the news that the official @entity6 is getting its first gay character. According to the sci-fi website @entity9, the upcoming novel "@entity11" will feature a capable but flawed @entity13 official named @entity14 who "also happens to be a lesbian." The character is the first gay figure in the official @entity6 -- the movies, television shows, comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24, editor of "@entity6" characters in "@placeholder" movies have gradually become more diverse.

**CNN/Daily Mail Datasets**

CNN: 380k, Daily Mail: 879k training - free!
Stanford Attentive Reader

Bidirectional LSTMs

characters in "@placeholder"

movies have gradually become more diverse

(Chen et al, 2016)
Stanford Attentive Reader

Bidirectional LSTMs

characters in " @placeholder " movies have gradually become more diverse

(Chen et al, 2016)
Stanford Attentive Reader

characters in "@placeholder"
movies have gradually become more diverse

Bidirectional LSTMs

Characters in "@placeholder" movies have gradually become more diverse.
Stanford Attentive Reader

characters in "@placeholder"

movies have gradually become more diverse

Bidirectional LSTMs

characters in "@placeholder"
characters in "@placeholder"
movies have gradually become more diverse

Bidirectional LSTMs
Stanford Attentive Reader

characters in "@placeholder"

movies have gradually become more diverse

Bidirectional LSTMs

(Chen et al, 2016)
Stanford Attentive Reader

Bidirectional LSTMs

characters in " @placeholder " movies have gradually become more diverse

Attention

\[ \alpha_i = \text{softmax} \left( q^T W_s \tilde{p}_i \right) \]
Stanford Attentive Reader

characters in "@placeholder"
movies have gradually become more diverse

Attention

\[ \alpha_i = \frac{\text{softmax} (q^T W_s \tilde{p}_i)}{i} \]

\[ o = \sum_i \alpha_i \tilde{p}_i \]
Stanford Attentive Reader

Bidirectional LSTMs

Attention

\[ \alpha_i = \text{softmax}(q^T W_s \tilde{p}_i) \]

\[ o = \sum_i \alpha_i \tilde{p}_i \]

\[ a = \arg\max_{a \in p \cap E} W_a^T o \]

Characters in "@placeholder" movies have gradually become more diverse

(Chen et al, 2016)
Results
## Results

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>(Hermann et al, 2015)</td>
<td>NIPS’15</td>
<td>61.8</td>
</tr>
<tr>
<td>(Hill et al, 2016)</td>
<td>ICLR’16</td>
<td>63.4</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>(Hermann et al, 2015)</td>
<td>NIPS’15</td>
<td>61.8</td>
</tr>
<tr>
<td>(Hill et al, 2016)</td>
<td>ICLR’16</td>
<td>63.4</td>
</tr>
<tr>
<td>Ours: neural net</td>
<td>ACL’16</td>
<td>73.8</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>(Hermann et al, 2015)</td>
<td>NIPS’15</td>
<td>61.8</td>
</tr>
<tr>
<td>(Hill et al, 2016)</td>
<td>ICLR’16</td>
<td>63.4</td>
</tr>
<tr>
<td>Ours: neural net</td>
<td>ACL’16</td>
<td>73.8</td>
</tr>
<tr>
<td>Ours: neural net (ensemble)</td>
<td>ACL’16</td>
<td>77.2</td>
</tr>
</tbody>
</table>

(Chen et al, 2016)
# Results

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>(Hermann et al, 2015)</td>
<td>NIPS’15</td>
<td>61.8</td>
</tr>
<tr>
<td>(Hill et al, 2016)</td>
<td>ICLR’16</td>
<td>63.4</td>
</tr>
<tr>
<td>Ours: neural net</td>
<td>ACL’16</td>
<td><strong>73.8</strong></td>
</tr>
<tr>
<td>Ours: neural net (ensemble)</td>
<td>ACL’16</td>
<td><strong>77.2</strong></td>
</tr>
<tr>
<td>(Kobayashi et al, 2016)</td>
<td>NAACL’16</td>
<td>71.3</td>
</tr>
<tr>
<td>(Kadlec et al, 2016)</td>
<td>ACL’16</td>
<td>68.6</td>
</tr>
<tr>
<td>(Dhingra et al, 2016)</td>
<td>2016/6/5</td>
<td>73.0</td>
</tr>
<tr>
<td>(Sodorni et al, 2016)</td>
<td>2016/6/7</td>
<td>72.6</td>
</tr>
<tr>
<td>(Trischler et al, 2016)</td>
<td>2016/6/7</td>
<td>73.4</td>
</tr>
<tr>
<td>(Weissenborn, 2016)</td>
<td>2016/7/12</td>
<td>N/A</td>
</tr>
<tr>
<td>(Cui et al, 2016)</td>
<td>2016/7/15</td>
<td>73.1</td>
</tr>
</tbody>
</table>
Breakdown of the Examples

- Exact match
- Paraphrasing
- Partial clue
- Multiple sentences
- Coreference errors
- Ambiguous / hard

(Chen et al, 2016)
Exact Match

... it's clear @entity0 is leaning toward @entity60 ...

" it's clear @entity0 is leaning toward @placeholder, " says an expert who monitors @entity0

@entity60
... @entity0 called me personally to let me know that he wouldn't be playing here at @entity23, " @entity3 said ... 

@placeholder says he understands why @entity0 won't play at his tournament

@entity3
Partial Clue

P: @entity12 " @entity2 professed that his " @entity11 " is not a religious book . . . .

Q: a tv movie based on @entity2 's book " @placeholder " casts a @entity76 actor as @entity5

A: @entity11
Multiple sentences

"..." we got some groundbreaking performances, here too, tonight," @entity6 said. "we got @entity17, who will be doing some musical performances. he's doing a his-and-her duet all by himself."...

"he's doing a his-and-her duet all by himself," @entity6 said of @placeholder

@entity17
Coreference Error

... hip-hop star @entity246 saying on @entity247 that he was canceling an upcoming show for the @entity249 ...

rapper @placeholder "disgusted,"
cancels upcoming show for @entity280

@entity280 = @entity249 = SAEs

@entity246
Ambiguous / Hard

\[ P \]
... a small aircraft carrying @entity5, @entity6 and @entity7 "the @entity12" @entity3 crashed ...

\[ Q \]
pilot error and snow were reasons stated for @placeholder plane crash

\[ A \] @entity5
Breakdown of the Examples

- Exact match
- Paraphrasing
- Partial clue
- Multiple sentences
- Coreference errors
- Ambiguous / hard

(Chen et al, 2016)
Breakdown of the Examples

- Exact match
- Paraphrasing
- Partial clue
- Multiple sentences
- Coreference errors
- Ambiguous / hard

CNN: 100 samples

41
19
13
17
8
2

(Chen et al, 2016)
Breakdown of the Examples

Exact match
Paraphrasing
Partial clue
Multiple sentences
Coreference errors
Ambiguous / hard

CNN: 100 samples

<table>
<thead>
<tr>
<th></th>
<th>neural net</th>
<th>neural net (ensemble)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73.8</td>
<td>77.2</td>
</tr>
<tr>
<td></td>
<td>73.6</td>
<td>77.6</td>
</tr>
</tbody>
</table>

(Chen et al, 2016)
Per-category Accuracies

Neural net vs. Categorical Feature Classifier

Correctness (%)

<table>
<thead>
<tr>
<th>Category</th>
<th>Neural net</th>
<th>Categorical Feature Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM (13)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Para. (41)</td>
<td>95</td>
<td>78</td>
</tr>
<tr>
<td>Partial (19)</td>
<td>90</td>
<td>74</td>
</tr>
<tr>
<td>Multi (2)</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Coref. E (8)</td>
<td>38</td>
<td>50</td>
</tr>
<tr>
<td>Hard (17)</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

(Chen et al, 2016)
Per-category Accuracies

- EM (13) with 100% accuracy for both Neural net and Categorical Feature Classifier.
- Para. (41) with 95% Neural net and 78% Categorical Feature Classifier.
- Partial (19) with 90% Neural net and 74% Categorical Feature Classifier.
- Multi (2) with 50% accuracy for both.
- Coref. E (8) with 38% Neural net and 50% Categorical Feature Classifier.
- Hard (17) with 6% Neural net and 12% Categorical Feature Classifier.

Sources:
- (Chen et al, 2016)
Per-category Accuracies

- **Neural net**
- **Categorical Feature Classifier**

![Graph showing per-category accuracies with EM (13), Para. (41), Partial (19), Multi (2), Coref. E (8), and Hard (17).](image)

(Chen et al, 2016)
Per-category Accuracies

- Neural net
- Categorical Feature Classifier

- EM (13): 100 (Neural) 100 (Categorical)
- Para. (41): 95 (Neural) 78 (Categorical)
- Partial (19): 90 (Neural) 74 (Categorical)
- Multi (2): 50 (Neural) 50 (Categorical)
- Coref. E (8): 38 (Neural) 50 (Categorical)
- Hard (17): 6 (Neural) 12 (Categorical)

(Chen et al, 2016)
Per-category Accuracies

- **EM (13)**: 100%
- **Para. (41)**: Neural net 95%, Categorical Feature Classifier 78%
- **Partial (19)**: Neural net 90%, Categorical Feature Classifier 74%
- **Multi (2)**: 50% for both categories
- **Coref. E (8)**: 38% for neural net, 50% for Categorical Feature Classifier
- **Hard (17)**: 6% for neural net, 12% for Categorical Feature Classifier

(Chen et al, 2016)
Summary

- Simple biLSTM + attention model almost hits the capacity of this task; It is great for learning semantic matches.
Summary

• Simple biLSTM + attention model almost hits the capacity of this task; It is great for learning semantic matches.

• Great paradigm for creating large MC datasets
  • Still noisy and artificial (not real questions)
  • Not hard enough for reasoning and inference.
Summary

• Simple biLSTM + attention model almost hits the capacity of this task; It is great for **learning semantic matches**.

• Great paradigm for creating large MC datasets
  • Still noisy and artificial (not real questions)
  • Not hard enough for reasoning and inference.

Next: Does it work for a real QA problem?
Stanford Question Answering Dataset (SQuAD)

- Passage + Question → Answer
  - **Passage**: selected from Wikipedia
  - **Question**: crowdsourced
  - **Answer**: must be a span in the passage

Extractive Question Answering

(Rajpurkar et al, 2016)
He came to power by uniting many of the nomadic tribes of Northeast Asia. After founding the Mongol Empire and being proclaimed "Genghis Khan", he started the Mongol invasions that resulted in the conquest of most of Eurasia. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.
He came to power by uniting many of the nomadic tribes of Northeast Asia. After founding the Mongol Empire and being proclaimed "Genghis Khan", he started the Mongol invasions that resulted in the conquest of most of Eurasia. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.
Stanford Question Answering Dataset (SQuAD)

F1 score

- Logistic Regression: 51
- Best Single System: 80.8
- Human Performance: 91.2

Leaderboard at https://stanford-qa.com
Who did Genghis Khan unite before he began conquering the rest of Eurasia?

He came to power by uniting many of the nomadic tribes of Northeast Asia. After founding the Mongol Empire and being proclaimed "Genghis Khan", he started the Mongol invasions that resulted in the conquest of most of Eurasia. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmid and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.
Stanford Attentive Reader++

Bidirectional RNNs

Attention

\[ \alpha_i = \text{softmax} \left( q^T W_s \tilde{p}_i \right) \]
Stanford Attentive Reader++

Bidirectional RNNs

Attention

\[ \alpha_i = \text{softmax} \left( q^T W_s \tilde{p}_i \right) \]

→ predict start token
Stanford Attentive Reader++

Bidirectional RNNs

Attention

\[ \alpha_i = \text{softmax}(q^T W_s \tilde{p}_i) \]
→ predict start token

Attention

\[ \alpha'_i = \text{softmax}(q^T W'_s \tilde{p}_i) \]
→ predict end token
Missing Piece

Unfortunately, this simple model doesn’t work very well… F1 <= 60 :(
Missing Piece

Unfortunately, this simple model doesn’t work very well… F1 <= 60 :(

**Intuition:**
Most tokens in the passage are irrelevant. We should find a way to quickly zoom in and focus on the things that we need to learn.
Adding a simple feature

• Just add one 0/1 exact match feature to word embedding: whether the current token appears in the question
Adding a simple feature

- Just add one 0/1 exact match feature to word embedding: whether the current token appears in the question

![Graph showing comparison between word embedding and exact match features over epochs]

Dev F1 vs # Epoches with two lines: word embedding and exact match, with a peak of 75.6.
... vs. “Conditioned” Question Embeddings

Idea: signifies which question words are most relevant to each word in the passage.

\[
q_i^{align} = \sum_{j=1}^{n} a_{ij} q_j
\]

\[
a_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^{n} \exp(s_{ik})}
\]
… vs. “Conditioned” Question Embeddings

Idea: signifies which question words are most relevant to each word in the passage.

\[ q_{i}^{align} = \sum_{j=1}^{n} a_{ij}q_{j} \]

\[ a_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^{n} \exp(s_{ik})} \]

Similar counterparts in other models:

- (Lee et al, 2016): Passage-aligned question representation
- (Seo et al, 2016): context-to-query attention
- (Xiong et al, 2016): coattention
## More feature studies

<table>
<thead>
<tr>
<th>Feature Configuration</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word embedding only</td>
<td>58.9</td>
</tr>
<tr>
<td>+ Exact match</td>
<td>75.6</td>
</tr>
<tr>
<td>+ Conditioned question embedding</td>
<td>76.9</td>
</tr>
<tr>
<td>+ Both</td>
<td>77.9</td>
</tr>
</tbody>
</table>
## More feature studies

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word embedding only</td>
<td>58.9</td>
</tr>
<tr>
<td>+ Exact match</td>
<td>75.6</td>
</tr>
<tr>
<td>+ Conditioned question embedding</td>
<td>76.9</td>
</tr>
<tr>
<td>+ Both</td>
<td>77.9</td>
</tr>
</tbody>
</table>

**Similar but complementary**

“hard” vs “soft” alignment
A few more improvements

- Stack LSTM helps!

- Fix pre-trained word embeddings, only re-train the embeddings for important words (what, which, why, etc).
- Adding token-level linguistic features helps.
## Results (single model)

<table>
<thead>
<tr>
<th>Model Description</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>51.0</td>
</tr>
<tr>
<td>Fine-Grained Gating (Carnegie Mellon U)</td>
<td>73.3</td>
</tr>
<tr>
<td>Match-LSTM (Singapore Management U)</td>
<td>73.7</td>
</tr>
<tr>
<td>DCN (Salesforce)</td>
<td>75.9</td>
</tr>
<tr>
<td>BiDAF (UW &amp; Allen Institute)</td>
<td>77.3</td>
</tr>
<tr>
<td>Multi-Perspective Matching (IBM)</td>
<td>78.7</td>
</tr>
<tr>
<td>ReasoNet (MSR Redmond)</td>
<td>79.4</td>
</tr>
<tr>
<td>Ours</td>
<td>79.4</td>
</tr>
<tr>
<td>r-net (MSR Asia)</td>
<td>80.8</td>
</tr>
<tr>
<td>Human performance</td>
<td>91.2</td>
</tr>
</tbody>
</table>

(Chen et al, 2017)
What is the remaining 10%?

Q: What is the total number of professors, instructors, and lecturers at Harvard?

Harvard's 2,400 professors, lecturers, and instructors instruct 7,200 undergraduates and 14,000 graduate students. The school color is crimson, which is also the name of the Harvard sports teams and the daily newspaper, The Harvard Crimson. The color was unofficially adopted (in preference to magenta) by an 1875 vote of the student body, although the association with some form of red can be traced back to 1858, when Charles William Eliot, a young graduate student who would later become Harvard's 21st and longest-serving president (1869–1909), bought red bandanas for his crew so they could more easily be distinguished by spectators at a regatta.
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. **The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers** 24–10 to earn their third **Super Bowl title**. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.
A modern example of school discipline in North America and Western Europe relies upon the idea of an assertive teacher who is prepared to impose their will upon a class. Positive reinforcement is balanced with immediate and fair punishment for misbehavior and firm, clear boundaries define what is appropriate and inappropriate behavior. **Teachers are expected to respect their students; sarcasm and attempts to humiliate pupils are seen as falling outside of what constitutes reasonable discipline.**
Summary

• A simple feature does the trick.
  • Learning a representation related to the **paraphrased** nature of a question vs. the context around an answer.
Summary

• A simple feature does the trick.
  • Learning a representation related to the **paraphrased** nature of a question vs. the context around an answer.

• **The last 10% to go:** Syntax? Frames? Semantics?
Summary

• A simple feature does the trick.
  • Learning a representation related to the paraphrased nature of a question vs. the context around an answer.

• The last 10% to go: Syntax? Frames? Semantics?

Next: Can we leverage these MC models?
Open-domain Question Answering
Open-domain Question Answering

**SQuAD**
Q: How many of Warsaw's inhabitants spoke Polish in 1933? A: 833,500

**TREC**
Q: What U.S. state’s motto is “Live free or Die”? A: New Hampshire

**WebQuestions** (Berant et al, 2013)
Q: What part of the atom did Chadwick discover? A: neutron

**WikiMovies** (Miller et al, 2016)
Q: Who wrote the film Gigli? A: Martin Brest
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

WIKIPEDIA
Q: How many of Warsaw's inhabitants spoke Polish in 1933?
Q: How many of Warsaw's inhabitants spoke Polish in 1933?
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Reader → 833,500
Document Retriever

70-86% of questions we have that the answer segment appears in the top 5 articles
Evaluated using pre-trained model

- Pre-trained SQuAD

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exact Match (top-1 prediction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>27.1</td>
</tr>
<tr>
<td>TREC</td>
<td>19.7</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>11.8</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>24.5</td>
</tr>
</tbody>
</table>

(Chen et al, 2017)
Distant Supervision

(Q, A) $\rightarrow$ (P, Q, A)

if P is retrieved and A can be found in P
Distant Supervision

(Q, A) \rightarrow (P, Q, A)

if P is retrieved and A can be found in P

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train</th>
<th>#DS. Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>87,385</td>
<td>71,231</td>
</tr>
<tr>
<td>TREC</td>
<td>1,486</td>
<td>3,464</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>3,778</td>
<td>4,602</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>96,185</td>
<td>36,301</td>
</tr>
</tbody>
</table>

(Mintz et al, 2009)
Distant Supervision

TREC

• **Q:** What U.S. state’s motto is “Live free or Die”?  
• **A:** New Hampshire

(Chen et al, 2017)
Distant Supervision

TREC

- **Q:** What U.S. state’s motto is “Live free or Die”?  
- **A:** New Hampshire

**Live Free or Die**

*From Wikipedia, the free encyclopedia*

"Live Free or Die" is the official motto of the U.S. state of New Hampshire, adopted by the state in 1945.[1] It is possibly the best-known of all state mottos, partly because it conveys an assertive independence historically found in American political philosophy and partly because of its contrast to the milder sentiments found in other state mottos.
Distant Supervision

WebQuestions

- **Q:** What part of the atom did Chadwick discover?
- **A:** neutron
Distant Supervision

WebQuestions

- **Q:** What part of the atom did Chadwick discover?
- **A:** neutron

**Atom**

From Wikipedia, the free encyclopedia

The atomic mass of these isotopes varied by integer amounts, called the whole number rule. The explanation for these different isotopes awaited the discovery of the neutron, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932. Isotopes were then explained as elements with the same number of protons, but different numbers of neutrons within the nucleus.
Distant Supervision

WikiMovies

• **Q:** Who wrote the film Gigli?
• **A:** Martin Brest

(Chen et al, 2017)
Distant Supervision

WikiMovies

- **Q:** Who wrote the film Gigli?
- **A:** Martin Brest

*Gigli*

From Wikipedia, the free encyclopedia

*Gigli* (/ˈdʒiːli/, jee-lee) is a 2003 American romantic comedy film written and directed by **Martin Brest** and starring Ben Affleck, Jennifer Lopez, Justin Bartha, Al Pacino, Christopher Walken, and Lainie Kazan.
Results

Exact match (top-1 prediction)
Results

- Pre-trained SQuAD
- Fine-tuning
- Multi-task learning

**Results** (Chen et al, 2017)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exact match (top-1 prediction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>27.1</td>
</tr>
<tr>
<td>TREC</td>
<td>19.7</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>11.8</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>24.5</td>
</tr>
</tbody>
</table>
Results

- Pre-trained SQuAD
- Fine-tuning
- Multi-task learning

Exact match (top-1 prediction)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pre-trained SQuAD</th>
<th>Fine-tuning</th>
<th>Multi-task learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>27.1</td>
<td>25.7</td>
<td>34.3</td>
</tr>
<tr>
<td>TREC</td>
<td>19.7</td>
<td>19.5</td>
<td>11.8</td>
</tr>
<tr>
<td>WebQuestions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WikiMovies</td>
<td></td>
<td></td>
<td>34.3</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pre-trained SQuAD</th>
<th>Fine-tuning</th>
<th>Multi-task learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>27.1</td>
<td>28.4</td>
<td>29.8</td>
</tr>
<tr>
<td>TREC</td>
<td>19.7</td>
<td>25.7</td>
<td>25.4</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>11.8</td>
<td>19.5</td>
<td>20.7</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>24.5</td>
<td>34.3</td>
<td>36.5</td>
</tr>
</tbody>
</table>

**Exact match (top-1 prediction)**

(Chen et al, 2017)
Demo

Hi!
Demo

Hello! Please ask a question.

Hi!
Demo

Hello! Please ask a question.

Hi!

Who invented LSTM?
Demo

Hello! Please ask a question.

Who invented LSTM?

Hmmm. Tough one. I would say... Sepp Hochreiter
Demo

Hello! Please ask a question.

Who invented LSTM?

Hmmm. Tough one. I would say... Sepp Hochreiter

Where is Stanford University located?
Demo

Hello! Please ask a question.

Who invented LSTM?

Hmmm. Tough one. I would say... Sepp Hochreiter

Where is Stanford University located?

I'm quite confident the answer is... California
Demo

What is the highest mountain on Earth?
Demo

What is the highest mountain on Earth?

I'm quite confident the answer is... Mount Everest
Demo

What is the highest mountain on Earth?

I'm quite confident the answer is... Mount Everest

What year was the American Declaration of Independence?
Demo

What is the highest mountain on Earth?

I'm quite confident the answer is... Mount Everest

What year was the American Declaration of Independence?

I'm quite confident the answer is... 1776
Demo

What is the meaning of life?
Demo

What is the meaning of life?

I'm quite confident the answer is... What is the meaning in asking?
Demo

What is the meaning of life?

I'm quite confident the answer is... What is the meaning in asking?

Who will solve AI?
Demo

What is the meaning of life?

I'm quite confident the answer is... What is the meaning in asking?

Who will solve AI?

Hmmm. Tough one. I would say... Computers
Final words
Final words

• It is a great time to work “towards the machine comprehension of text”.
Final words

- It is a great time to work “towards the machine comprehension of text”.

- We should understand the nature of datasets and models. Start from simple models!
Final words

• It is a great time to work “towards the machine comprehension of text”.

• We should understand the nature of datasets and models. Start from simple models!

• Attention-based RNNs look promising, but it is still challenging to learn the subtlety of language?
Final words

• It is a great time to work “towards the machine comprehension of text”.

• We should understand the nature of datasets and models. Start from simple models!

• Attention-based RNNs look promising, but it is still challenging to learn the subtlety of language?

• We can leverage these machine comprehension models!
Thanks

Questions are welcome :)

55