What is Question Answering?

In question answering, we aim to build systems that can automatically answer questions posed by humans in natural language.
What information source does the system use for answering questions?

A single paragraph; All documents on the Web; A Knowledge Base; An image; ...
Question Answering — Taxonomy

What information source does the system use for answering questions?
- A single paragraph;
- All documents on the Web;
- A Knowledge Base;
- An image;

What is the type of the questions?
- Factoid vs. Non-Factoid;
- Open-Domain vs. Closed-Domain;
- Simple vs. Compositional;
- Natural vs. Cloze-style;
What information source does the system use for answering questions?
A single paragraph; All documents on the Web; A Knowledge Base; An image; …

What is the type of the questions?
Factoid vs. Non-Factoid; Open-Domain vs. Closed-Domain; Simple vs. Compositional; Natural vs. Cloze-style; …

What is the type of the answers?
Short text; Paragraph; List; Yes/No; …
Central Edinburgh

Calton Hill and the National Monument are situated in Central Edinburgh, east of Edinburgh’s New Town. Marked as a UNESCO World Heritage Site, Calton Hill has some of the city’s best views and if you get up early, the best sunrises. Calton Hill is also resident to some iconic Scottish monuments and buildings.
Question Answering — Applications

Smart Speaker Use Case Frequency January 2019

- Ask a question: 84.0% use, 66.0% moderate, 36.9% rarely
- Listen to streaming music service: 83.0% use, 69.9% moderate, 38.2% rarely
- Check the weather: 80.1% use, 61.4% moderate, 35.6% rarely
- Set an alarm: 62.4% use, 41.8% moderate, 23.5% rarely
- Set a timer: 62.4% use, 46.7% moderate, 22.9% rarely
- Listen to radio: 54.9% use, 40.5% moderate, 21.2% rarely
- Use a favorite Alexa skill / Google Action: 48.7% use, 35.0% moderate, 18.3% rarely
- Play game or answer trivia: 48.0% use, 29.1% moderate, 10.8% rarely
- Control smart home devices: 45.8% use, 33.3% moderate, 23.5% rarely
- Listen to news or sports: 43.8% use, 28.8% moderate, 13.4% rarely
- Search for product info: 41.2% use, 27.8% moderate, 10.8% rarely
Question: When did Beyoncé start becoming popular?

Answer: in the late 1990s

Beyoncé Giselle Knowles-Carter is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Beyoncé Giselle Knowles-Carter is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Beyoncé Giselle Knowles-Carter is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Question: When did Beyonce start becoming popular?
Question: When did Beyoncé start becoming popular?
Answer: in the late 1990s
Beyoncé Giselle Knowles-Carter is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Q: In what areas did Beyoncé compete in when she was young?

Answer: singing and dancing
Beyoncé Giselle Knowles-Carter is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Question: In what city and state did Beyoncé grow up?
Answer: Houston, Texas
Why work on QA?

Useful in many applications!
Why work on QA?

Useful in many applications!
Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.
Why work on QA?

Useful in many applications!
Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.
Many other NLP tasks can be reduced to Reading Comprehension:
Why work on QA?

Useful in many applications!
Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.

Many other NLP tasks can be reduced to Reading Comprehension:

**Machine Translation:**

**Question:** How do you say the following sentence in Italian?

**Paragraph:** The quick brown fox jumps over the lazy dog.
Why work on QA?

Useful in many applications!

Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.

Many other NLP tasks can be reduced to Reading Comprehension:

**Information Extraction:** (Barack Obama, educated_at, ?)

**Question:** where did Barack Obama graduate from?

**Paragraph:** Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organiser in Chicago.
Why work on QA?

Useful in many applications!

Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.

Many other NLP tasks can be reduced to Reading Comprehension:

**Part-of-Speech Tagging:**

**Question:** What is the part of speech of [runs] in the sentence?

**Paragraph:** He runs fast in the morning.
Why work on QA?

Useful in many applications!
Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.
Many other NLP tasks can be reduced to Reading Comprehension:

Math Word Problems:

**Question:** What is the solution to the following problem?

**Paragraph:** Lisa has 7 apples. She buys 12 more apples at the grocery store. Then, she gives 5 apples to her friend. How many apples does Lisa have now?
Why work on QA?

Useful in many applications!

Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.

Many other NLP tasks can be reduced to Reading Comprehension:

Language Modeling:

**Question:** What is the next word in the following sentence?

**Paragraph:** Despite the heavy rain, the match continued without any
Why work on QA?

Useful in many applications!

Reading Comprehension and QA are a great test bed for evaluating how well computer systems “understand” human language.

Many other NLP tasks can be reduced to Reading Comprehension:

Relation Extraction: (Elon Musk, ?, Tesla)

Question: What is the relationship between “Elon Musk” and “Tesla” in the text?

Paragraph: Elon Musk [...] is also known for his role in leading Tesla, Inc., where he serves as CEO and leads the company's innovative projects on electric vehicles and clean energy.
Stanford Question Answering Dataset (SQuAD)

100k annotated passage-question-answer triples

Large-scale supervised datasets were instrumental for training effective neural RC models.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall? gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? graupel

Where do water droplets collide with ice crystals to form precipitation? within a cloud
Stanford Question Answering Dataset (SQuAD)

100k annotated passage-question-answer triples

Large-scale supervised datasets were instrumental for training effective neural RC models.

Passages are selected from the English Wikipedia.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers."

What causes precipitation to fall? gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? graupel

Where do water droplets collide with ice crystals to form precipitation? within a cloud.
Stanford Question Answering Dataset (SQuAD)

100k annotated passage-question-answer triples

Large-scale supervised datasets were instrumental for training effective neural RC models

Passages are selected from the English Wikipedia

Crowdsourced questions

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall? gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? graupel

Where do water droplets collide with ice crystals to form precipitation? within a cloud
Stanford Question Answering Dataset (SQuAD)

100k annotated passage-question-answer triples

Large-scale supervised datasets were instrumental for training effective neural RC models

Passages are selected from the English Wikipedia

Crowdsourced questions

Each answer is a short segment of text (span) in the passage.

It does not include questions where the answer are not mentioned in the span, and unanswerable questions

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
 gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
 graupel

Where do water droplets collide with ice crystals to form precipitation?
 within a cloud
Stanford Question Answering Dataset (SQuAD)

100k annotated passage-question-answer triples

Large-scale supervised datasets were instrumental for training effective neural RC models

Passages are selected from the English Wikipedia

Crowdsourced questions

Each answer is a short segment of text (span) in the passage.

It does not include questions where the answer are not mentioned in the span, and unanswerable questions

SQuAD was very popular for quite some time (e.g., 2016-2018) — today is considered “almost solved” since neural models exceed human performance

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?

**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

**graupel**

Where do water droplets collide with ice crystals to form precipitation?

**within a cloud**
Stanford Question Answering Dataset (SQuAD)

Evaluation:

Metrics: **Exact Match** (EM; 0 or 1) and **F1** (partial credit)

Q: What did Tesla do in December 1878?

A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: left Graz and severed EM: max{0, 0, 0} = 0                     F1: max{0.67, 0.67, 0.61} = 0.67
Stanford Question Answering Dataset (SQuAD)

Evaluation:
Metrics: **Exact Match** (EM; 0 or 1) and **F1** (partial credit)
Multiple valid answers for each question

Q: What did Tesla do in December 1878?
A: {left Graz, left Graz, left Graz and severed all relations with his family}
Prediction: left Graz and severed

EM: max{0, 0, 0} = 0                     F1: max{0.67, 0.67, 0.61} = 0.67
Stanford Question Answering Dataset (SQuAD)

Evaluation:

Metrics: **Exact Match** (EM; 0 or 1) and **F1** (partial credit)

Multiple valid answers for each question

Each predicted answer is compared to each of the gold answer (some normalisation: a, an, the, punctuations are removed); we take the maximum EM and F1 scores. We then average EM and F1 over the whole dataset.
Stanford Question Answering Dataset (SQuAD)

Evaluation:

Metrics: **Exact Match** (EM; 0 or 1) and **F1** (partial credit)

Multiple valid answers for each question

Each predicted answer is compared to each of the gold answer (some normalisation: a, an, the, punctuations are removed); we take the maximum EM and F1 scores. We then average EM and F1 over the whole dataset.

**Estimated human performance:** EM 82.3, F1 91.2
Stanford Question Answering Dataset (SQuAD)

Evaluation:

Metrics: **Exact Match** (EM; 0 or 1) and **F1** (partial credit)

Multiple valid answers for each question

Each predicted answer is compared to each of the gold answer (some normalisation: a, an, the, punctuations are removed); we take the maximum EM and F1 scores. We then average EM and F1 over the whole dataset.

Estimated human performance: **EM 82.3**, **F1 91.2**

Q: What did Tesla do in December 1878?
A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: *left Graz and severed*

EM: max{0, 0, 0} = 0  
F1: max{0.67, 0.67, 0.61} = 0.67
Stanford Question Answering Dataset (SQuAD)
Training Neural RC Models

Problem:

**Input:** context/paragraph $C = \langle c_1, \ldots, c_n \rangle$, question $Q = \langle q_1, \ldots, q_m \rangle$, $c_i, q_j \in V$

**Output:** $1 \leq \text{answer start index} \leq \text{answer end index} \leq n$

Start and end Indices of the answer in the provided context/passage
Training Neural RC Models

Problem:

Input: context/paragraph $C = \langle c_1, \ldots, c_n \rangle$, question $Q = \langle q_1, \ldots, q_m \rangle$, $c_i, q_j \in V$

Output: $1 \leq$ answer start index $\leq$ answer end index $\leq n$

Start and end Indices of the answer in the provided context/passage

Back in the days, before LLMs:

LSTM-based solutions (2016-2018), e.g.

- Attentive Reader (Hermann et al. 2015)
- Bi-Directional Attention Flow (Seo et al. 2016)

More recently: fine-tuning BERT-like models for RC (2019+)
Bidirectional Attention Flow (BiDAF)
Bidirectional Attention Flow (BiDAF)

Contextual Embedding Layer:
Concatenate word embeddings (e.g., GloVe) and character embedding for each word in the context and query:
\[ e(c_i) = \text{emb}(c_i) \]
\[ e(q_i) = \text{emb}(q_i) \]
such that \( \text{emb}(x) = f([\text{GloVe}(x); \text{charEmb}(x)]) \)

Two bi-directional LSTMs to produce contextual embeddings for context and query:
\[ c = \text{BiLSTM}([e(c_1), \ldots, e(c_n)]) \]
\[ q = \text{BiLSTM}([e(q_1), \ldots, e(q_m)]) \]

Contextual Encoder

Diagram showing the flow of information from context to query and vice versa through attention layers.
Bidirectional Attention Flow (BiDAF)

Contextual Embedding Layer: Concatenate word embeddings (e.g., GloVe) and character embedding for each word in the context and query:

\[
\begin{align*}
e(c_i) &= \text{emb}(c_i) \\
e(q_i) &= \text{emb}(q_i)
\end{align*}
\]

such that

\[
\text{emb}(x) = f\left(\left[\text{GloVe}(x); \text{charEmb}(x)\right]\right)
\]
Bidirectional Attention Flow (BiDAF)

**Contextual Embedding Layer:** Concatenate word embeddings (e.g., GloVe) and character embedding for each word in the context and query:

\[
\begin{align*}
  e(c_i) &= \text{emb}(c_i) \\
  e(q_i) &= \text{emb}(q_i)
\end{align*}
\]

such that

\[
\text{emb}(x) = f\left(\left[\text{GloVe}(x); \text{charEmb}(x)\right]\right)
\]

Two bi-directional LSTMs to produce contextual embeddings for context and query:

\[
\begin{align*}
  c &= \text{BiLSTM} \left(\left[e(c_1), \ldots, e(c_n)\right]\right) \\
  q &= \text{BiLSTM} \left(\left[e(q_1), \ldots, e(q_m)\right]\right)
\end{align*}
\]
Context-to-Query Attention: for each context word, choose the most relevant words from the question words.
Context-to-Query Attention: for each context word, choose the most relevant words from the question words.

Question: Who is the leader of the United States?

Context: Joseph Biden is an [...] and current president of the US.
**Query-to-Context Attention:** for each question word, choose the most relevant words from the context words.
**Query-to-Context Attention:** for each question word, choose the most relevant words from the context words.

Context: **Carmona** scored the **only goal** of the **game** in a 1-0 over [...]

Question: Who scored the winning goal in the World Cup final?
Bidirectional Attention Flow (BiDAF)

Compute a **similarity score** for every context-query token pair \((c_i, q_j)\):

\[
S_{ij} = w_{\text{sim}}^T \left[ c_i; q_j; c_i \odot q_j \right]
\]
Bidirectional Attention Flow (BiDAF)

Compute a **similarity score** for every context-query token pair \((c_i, q_j)\):

\[
S_{ij} = \mathbf{w}_\text{sim}^T \left[ c_i; q_j; c_i \odot q_j \right]
\]

**Context-to-query attention** (find relevant question words for a context word):

\[
\alpha_{ij} = \text{softmax}_j \left( S_{ij} \right), \quad a_i = \sum_{j=1}^{m} \alpha_{ij} q_j
\]

**Distribution over question words**
Bidirectional Attention Flow (BiDAF)

Compute a similarity score for every context-query token pair \((c_i, q_j)\): 

\[
S_{ij} = \mathbf{w}_\text{sim}^\top \left[ c_i; q_j; c_i \odot q_j \right]
\]

Context-to-query attention (find relevant question words for a context word):

\[
\alpha_{ij} = \text{softmax}_j \left( S_{ij} \right), \quad a_i = \sum_{j=1}^{m} \alpha_{ij} q_j
\]

Query-to-context attention (find relevant context words for a question):

\[
\beta_i = \text{softmax} \left( \max_{\text{col}} \left( S_{ij} \right) \right), \quad b_i = \sum_{i=1}^{n} \beta_i c_i
\]
Bidirectional Attention Flow (BiDAF)

$$\alpha_{ij} = \text{softmax}_j(S_{ij}), \quad \mathbf{a}_i = \sum_{j=1}^{m} \alpha_{ij} \mathbf{q}_j$$  
Context-to-Query Attention

Query-to-Context Attention

$$\beta_i = \text{softmax}\left(\max_{\text{col}}(S_{ij})\right), \quad \mathbf{b}_i = \sum_{i=1}^{n} \beta_i \mathbf{c}_i$$
Bidirectional Attention Flow (BiDAF)

\[
\alpha_{ij} = \text{softmax}_j \left( S_{ij} \right), \quad \mathbf{a}_i = \sum_{j=1}^{m} \alpha_{ij} \mathbf{q}_j
\]

Context-to-Query Attention

\[
\mathbf{b}_i = \sum_{i=1}^{n} \beta_i \mathbf{c}_i
\]

Query-to-Context Attention

Output:

\[
\mathbf{g}_i = \left[ \mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}_i \right]
\]
**Bidirectional Attention Flow (BiDAF)**

**Modeling Layer:**
- Attention layer models interactions between query and context
- Modeling layer models interactions within context words

**Output Layer:**
- Two classifiers predict the start and end positions:
  - $P_{\text{start}} = \text{softmax}(w^\top_{\text{start}} [g_i; m_i])$
  - $P_{\text{end}} = \text{softmax}(w^\top_{\text{end}} [g_i; m'_i])$
  - $m'_i = \text{BiLSTM}(m_i)$

---

**Diagram:**
- Start and End boxes connected to Dense + Softmax and LSTM + Softmax layers.
- Arrows indicating flow from Modeling Layer to Output Layer.
- Modeling Layer depicted with LSTM blocks.

---

**Modeling and Output Layers**
Bidirectional Attention Flow (BiDAF)

Modeling Layer: passes the activations $g_i$ to a multi-layer bi-directional LSTM:

$$m_i = \text{BiLSTM}(g_i)$$
**Bidirectional Attention Flow (BiDAF)**

**Modeling Layer:** passes the activations $g_i$ to a multi-layer bi-directional LSTM:

$$m_i = \text{BiLSTM}(g_i)$$

**Output Layer:** two classifiers predict the start and end positions of the answer:

$$P_{\text{start}} = \text{softmax} \left( w_{\text{start}}^T [g_i; m_i] \right) \quad P_{\text{end}} = \text{softmax} \left( w_{\text{end}}^T [g_i; m'_i] \right)$$

with $m'_i = \text{BiLSTM}(m_i)$
Bidirectional Attention Flow (BiDAF)

Training Objective: maximise the likelihood of the true answer span delimited by $(s^*, e^*)$:

$$\arg \max_{\theta} \log P_{\text{start}}(s^*; \theta) + \log P_{\text{end}}(e^*; \theta)$$
Bidirectional Attention Flow (BiDAF)

Ablation:

- **BiDAF**: 77.3

<table>
<thead>
<tr>
<th></th>
<th>Single Model</th>
<th>Ensemble</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
<td>F1</td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>Logistic Regression Baseline$^a$</td>
<td>40.4</td>
<td>51.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dynamic Chunk Reader$^b$</td>
<td>62.5</td>
<td>71.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fine-Grained Gating$^c$</td>
<td>62.5</td>
<td>73.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Match-LSTM$^d$</td>
<td>64.7</td>
<td>73.7</td>
<td>67.9</td>
<td>77.0</td>
</tr>
<tr>
<td>Multi-Perspective Matching$^e$</td>
<td>65.5</td>
<td>75.1</td>
<td>68.2</td>
<td>77.2</td>
</tr>
<tr>
<td>Dynamic Coattention Networks$^f$</td>
<td>66.2</td>
<td>75.9</td>
<td>71.6</td>
<td>80.4</td>
</tr>
<tr>
<td>R-Net$^g$</td>
<td><strong>68.4</strong></td>
<td><strong>77.5</strong></td>
<td>72.1</td>
<td>79.7</td>
</tr>
<tr>
<td>BiDAF (Ours)</td>
<td>68.0</td>
<td>77.3</td>
<td><strong>73.3</strong></td>
<td><strong>81.1</strong></td>
</tr>
</tbody>
</table>

(a) Results on the SQuAD test set

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No char embedding</td>
<td>65.0</td>
<td>75.4</td>
</tr>
<tr>
<td>No word embedding</td>
<td>55.5</td>
<td>66.8</td>
</tr>
<tr>
<td>No C2Q attention</td>
<td>57.2</td>
<td>67.7</td>
</tr>
<tr>
<td>No Q2C attention</td>
<td>63.6</td>
<td>73.7</td>
</tr>
<tr>
<td>Dynamic attention</td>
<td>63.5</td>
<td>73.6</td>
</tr>
<tr>
<td>BiDAF (single)</td>
<td>67.7</td>
<td>77.3</td>
</tr>
<tr>
<td>BiDAF (ensemble)</td>
<td>72.6</td>
<td>80.7</td>
</tr>
</tbody>
</table>

(b) Ablations on the SQuAD dev set
Bidirectional Attention Flow (BiDAF)

Ablation:
- BiDAF: 77.3
- No word embeddings: 66.8
- No context-to-query attention: 67.7
- No query-to-context attention: 73.7
- No character embeddings: 75.4

<table>
<thead>
<tr>
<th>Model</th>
<th>Single EM</th>
<th>Single F1</th>
<th>Ensemble EM</th>
<th>Ensemble F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression Baseline$^a$</td>
<td>40.4</td>
<td>51.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dynamic Chunk Reader$^b$</td>
<td>62.5</td>
<td>71.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fine-Grained Gating$^c$</td>
<td>62.5</td>
<td>73.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Match-LSTM$^d$</td>
<td>64.7</td>
<td>73.7</td>
<td>67.9</td>
<td>77.0</td>
</tr>
<tr>
<td>Multi-Perspective Matching$^e$</td>
<td>65.5</td>
<td>75.1</td>
<td>68.2</td>
<td>77.2</td>
</tr>
<tr>
<td>Dynamic Coattention Networks$^f$</td>
<td>66.2</td>
<td>75.9</td>
<td>71.6</td>
<td>80.4</td>
</tr>
<tr>
<td>R-Net$^g$</td>
<td>68.4</td>
<td>77.5</td>
<td>72.1</td>
<td>79.7</td>
</tr>
<tr>
<td>BiDAF (Ours)</td>
<td>68.0</td>
<td>77.3</td>
<td>73.3</td>
<td>81.1</td>
</tr>
</tbody>
</table>

(a) Results on the SQuAD test set

<table>
<thead>
<tr>
<th>Ablation</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No char embedding</td>
<td>65.0</td>
<td>75.4</td>
</tr>
<tr>
<td>No word embedding</td>
<td>55.5</td>
<td>66.8</td>
</tr>
<tr>
<td>No C2Q attention</td>
<td>57.2</td>
<td>67.7</td>
</tr>
<tr>
<td>No Q2C attention</td>
<td>63.6</td>
<td>73.7</td>
</tr>
<tr>
<td>Dynamic attention</td>
<td>63.5</td>
<td>73.6</td>
</tr>
<tr>
<td>BiDAF (single)</td>
<td>67.7</td>
<td>77.3</td>
</tr>
<tr>
<td>BiDAF (ensemble)</td>
<td>72.6</td>
<td>80.7</td>
</tr>
</tbody>
</table>

(b) Ablations on the SQuAD dev set
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.
BERT-based Span-Based QA Models

**Input:** Question [sep] Passage

**Answer:** Predict the start token and the end token of the answer

---

**Figure 14.12** An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.
J&M, 3d Edition

**Input:** Question [sep] Passage

**Answer:** Predict the start token and the end token of the answer

\[
P_{\text{start}}(c_i) = \text{softmax}\left( w_{\text{start}}^T h_i \right)
\]

\[
P_{\text{end}}(c_i) = \text{softmax}\left( w_{\text{end}}^T h_i \right)
\]

\(h_i\) is the contextual representation of \(c_i\) produced by BERT
BERT-based Span-Based QA Models

Input: Question [sep] Passage

Answer: Predict the start token and the end token of the answer

\[
P_{\text{start}}(c_i) = \text{softmax}(w_\text{start}^T h_i)
\]

\[
P_{\text{end}}(c_i) = \text{softmax}(w_\text{end}^T h_i)
\]

\(h_i\) is the contextual representation of \(c_i\) produced by BERT

Training Objective: maximise the likelihood of the true answer span:

\[
\arg \max_\theta \log P_{\text{start}}(s*; \theta) + \log P_{\text{end}}(e*; \theta)
\]

BERT Parameters

Figure 14.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.
BERT-based Span-Based QA Models

Training Objective: maximise the likelihood of the true answer span:

$$\arg\max_{\theta} \log P_{\text{start}}(s^*; \theta) + \log P_{\text{end}}(e^*; \theta)$$

Results: Close to human performance (almost) without any architecture engineering/tweaking!

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human performance</td>
<td>91.2*</td>
<td>82.3*</td>
</tr>
<tr>
<td>BiDAF</td>
<td>77.3</td>
<td>67.7</td>
</tr>
<tr>
<td>BERT-base</td>
<td>88.5</td>
<td>80.8</td>
</tr>
<tr>
<td>BERT-large</td>
<td>90.9</td>
<td>84.1</td>
</tr>
</tbody>
</table>

Figure 14.12: An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.
BiDAF vs. BERT-based Models

\[ x^1, x^2, x^3, \ldots, x^T \]

\[ q^1, q^2, q^3, \ldots, q^J \]

\[ m^1, m^2, m^3, \ldots, m^T \]

\[ u^1, u^2, u^3, \ldots, u^J \]

\[ h^1, h^2, h^3, \ldots, h^T \]

~2.5M params
Several BiLSTMs
Trained from scratch (minus GloVe)

110M-330M params
Transformers (no recurrence)
Pre-Trained

Figure 14.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.
McMuffin

Wikipedia page:

**Product description**

In the US and Canada the standard McMuffin consists of a slice of Canadian bacon,[5] a griddle-fried egg, and a slice of American cheese on a toasted and buttered English muffin. The round shape of the egg is made by cooking it in a white plastic ring surrounded by an outer metal structure.[8][1]

**History**

The sandwich was invented in 1972.[5] Former McDonald's President Ray Kroc wrote that Herb Peterson and his assistant, Donald Greade, the operator of a McDonald's Santa Barbara franchise in Goleta, California,[5] asked Kroc to look at something, without giving details because it was:

... a crazy idea — a breakfast sandwich. It consisted of an egg that had been formed in a Teflon circle with the yolk broken, and was dressed with a slice of cheese and a slice of grilled ham. It was served open-faced on a toasted and buttered English muffin. The advent of the Egg McMuffin opened up a whole new area of potential business for McDonald’s, the breakfast trade.[5][1]
The first McDonald’s Corporate-authorized Egg McMuffin was served at the Belleville, New Jersey McDonald’s in 1972.
Natural Questions, Annotation Task

Question: when was the egg mcmuffin added to the menu

Step 1: Annotator selects context

The first McDonald’s Corporate-authorized Egg McMuffin was served at the Belleville, New Jersey McDonald’s in 1972.

Step 2: Annotator selects short answer, where applicable

1972
Stelae were essentially stone banners raised to glorify the king and record his deeds, although the earliest examples depict mythological scenes. Imagery developed throughout the Classic Period, [..]
Open-Domain Question Answering (ODQA):

We do not assume we are given a passage together with the question.

We can only access a large collection of documents (e.g., Wikipedia) — we don’t know which document contains the answer, and the goal is to answer any open-domain questions.

Both more challenging and more practical/useful!
Speech and Language Processing Ed. 3, Ch. 14 on QA 😊


(Optional; worth a reading!) The Bitter Lesson: https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf