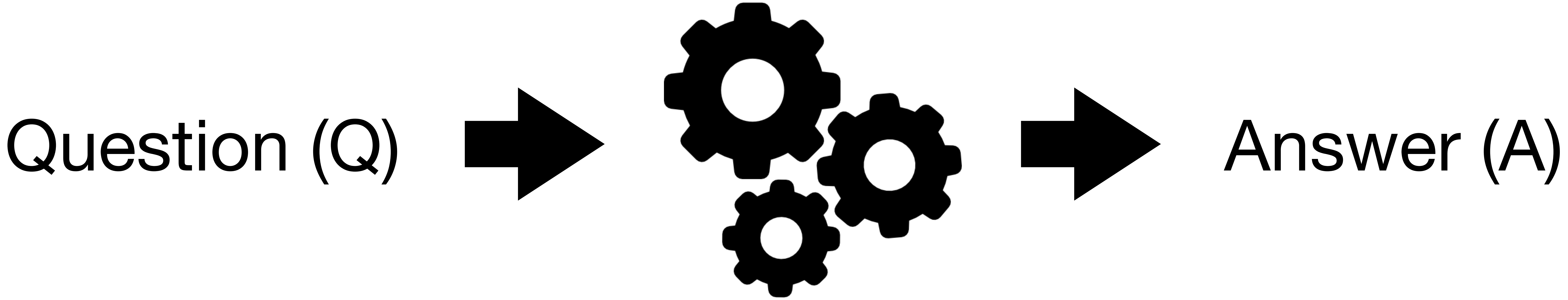
Natural Language Understanding, Generation, and Machine Translation

Lecture 18: Question Answering

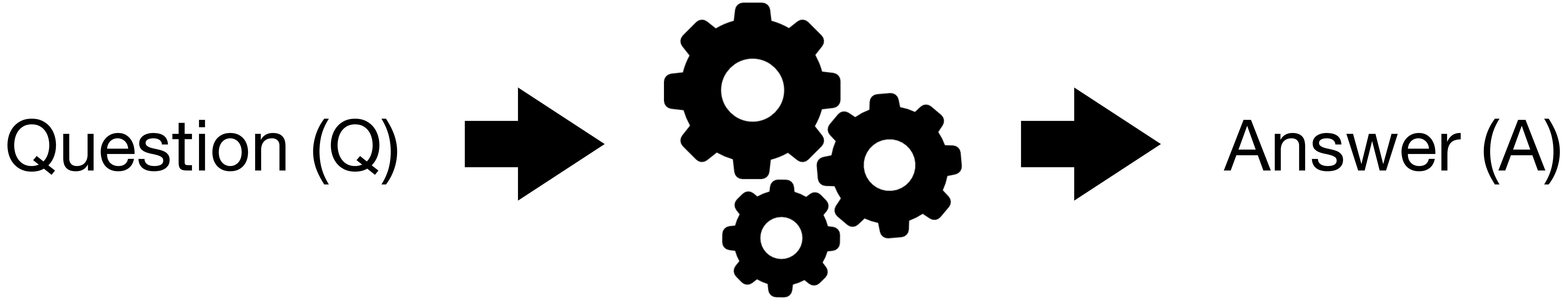
Pasquale Minervini
p.minervini@ed.ac.uk
March 1st, 2024

What is Question Answering?



In question answering, we aim to build systems that can **automatically** answer questions posed by humans in **natural language**.

Question Answering — Taxonomy

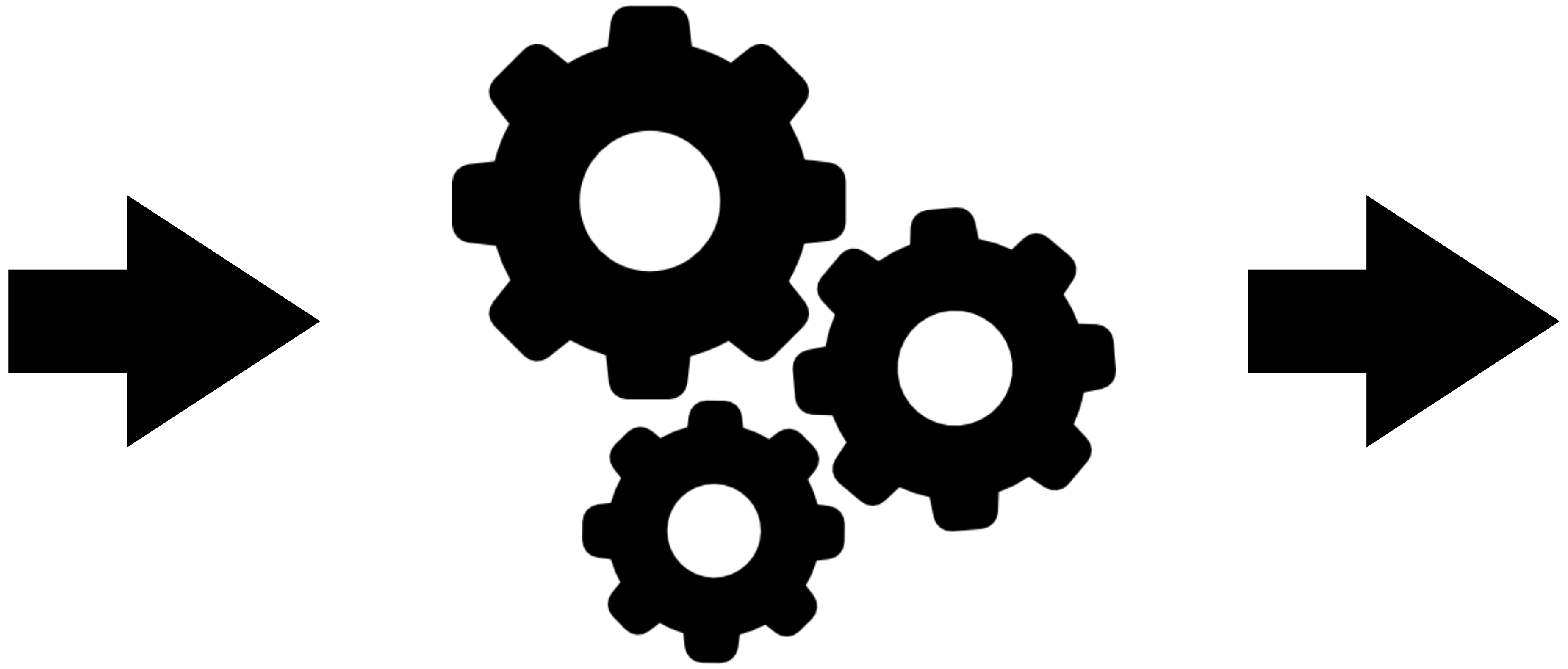


What information source does the system use for answering questions?

A single paragraph; All documents on the Web; A Knowledge Base; An image; …

Question Answering — Taxonomy

Question (Q)



Answer (A)

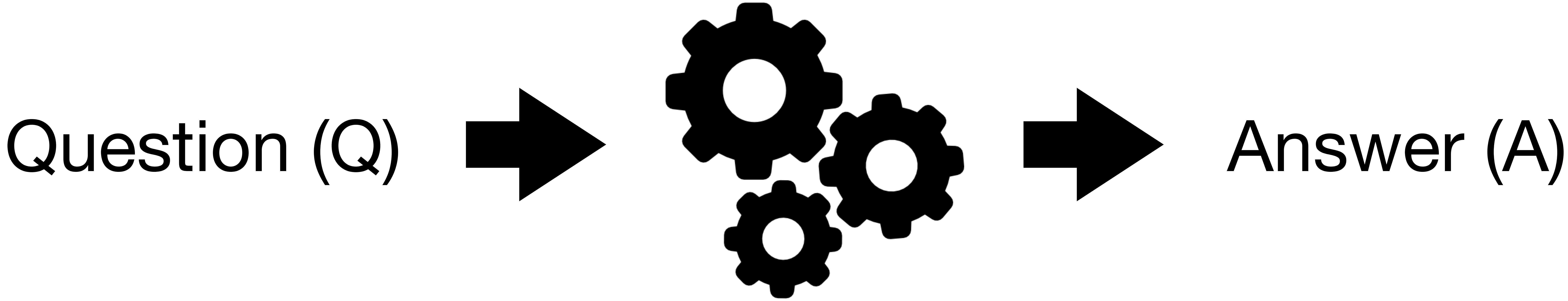
What information source does the system use for answering questions?

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; …

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?

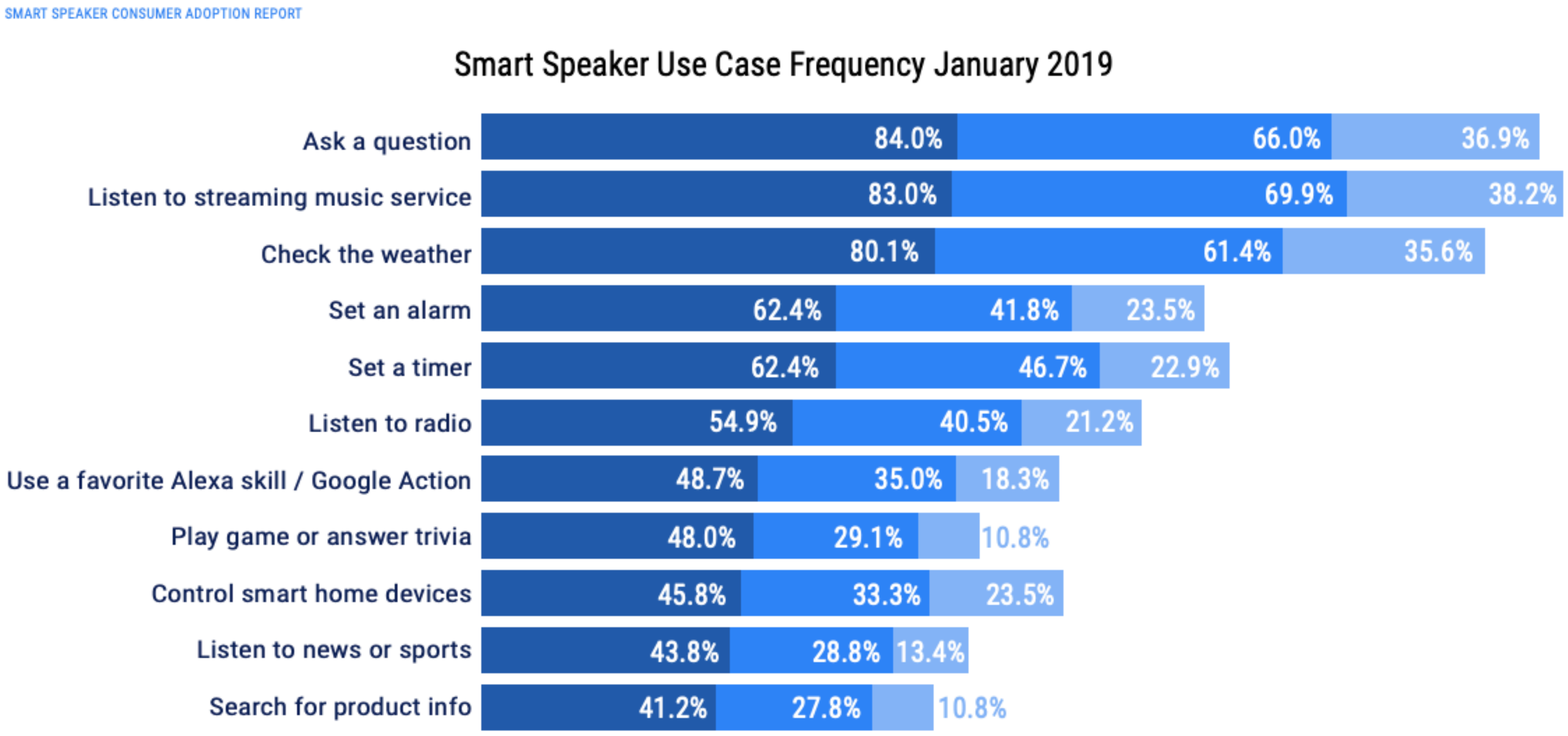
Factoid vs. Non-Factoid; Open-Domain vs. Closed-Domain; Simple vs. Compositional; Natural vs. Cloze-style; …

What is the type of the answers?

Short text; Paragraph; List; Yes/No; …

Question Answering — Applications

Question Answering — Applications



Reading Comprehension

Reading Comprehension: comprehend a passage of text, and answer questions about its content. $(P, Q) \rightarrow A$

Answer

Paragraph

Question

Reading Comprehension

Reading Comprehension: comprehend a passage of text, and answer questions about its content. $(P, Q) \rightarrow A$

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".

Reading Comprehension

Reading Comprehension: comprehend a passage of text, and answer questions about its content. $(P, Q) \rightarrow A$

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?

Reading Comprehension

Reading Comprehension: comprehend a passage of text, and answer questions about its content. $(P, Q) \rightarrow A$

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?

Answer: in the late 1990s

Reading Comprehension

Reading Comprehension: comprehend a passage of text, and answer questions about its content. $(P, Q) \rightarrow A$

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 Beyonce compete in when she was young?

Answer: singing and dancing

Reading Comprehension

Reading Comprehension: comprehend a passage of text, and answer questions about its content. $(P, Q) \rightarrow A$

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 Beyonce 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)

Stanford Question Answering Dataset (SQuAD)

Evaluation:

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

Multiple valid answers for each question

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: $1 \leq$ answer start index $\leq$ 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

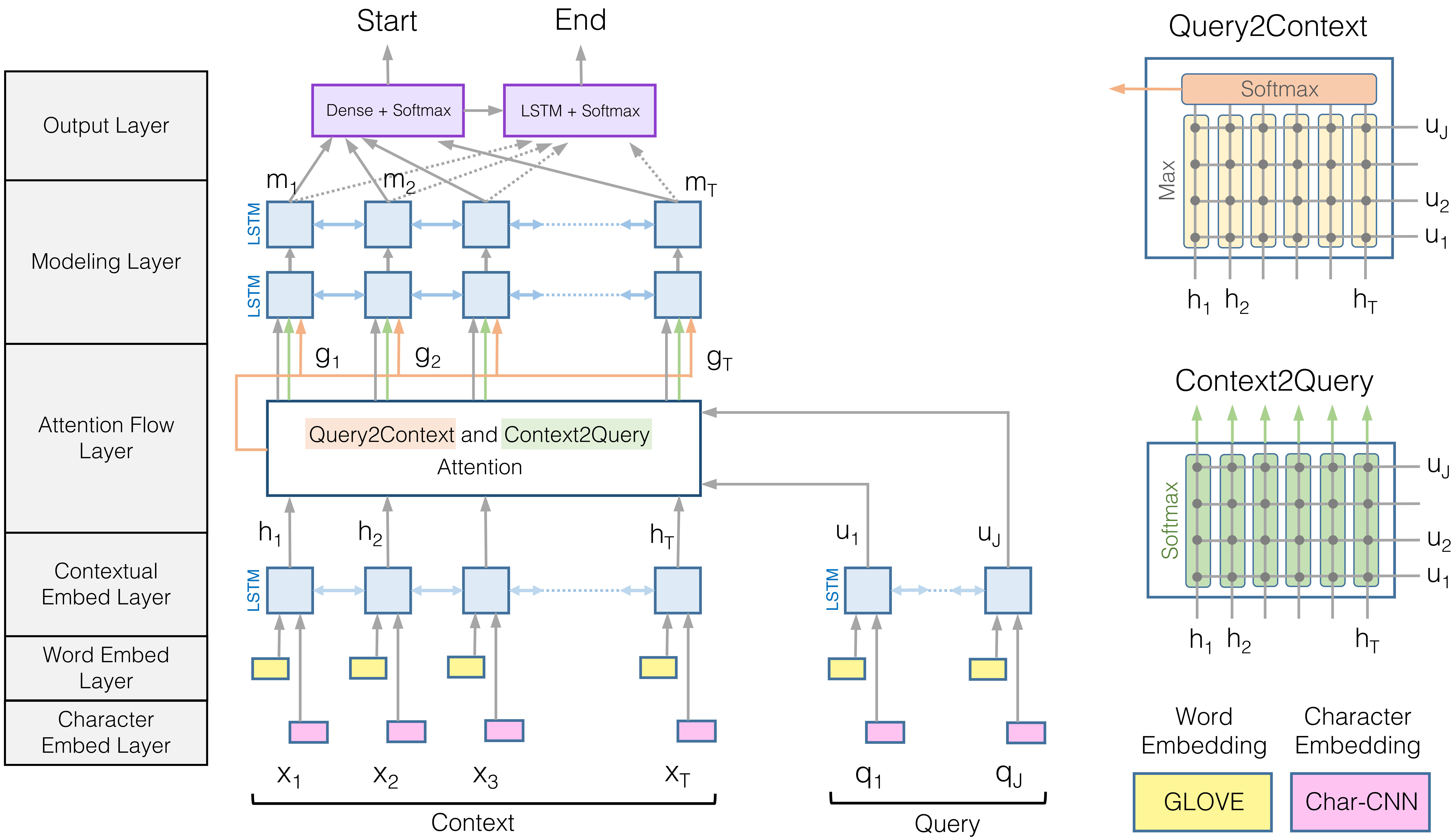
Back in the days, before LLMs:

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

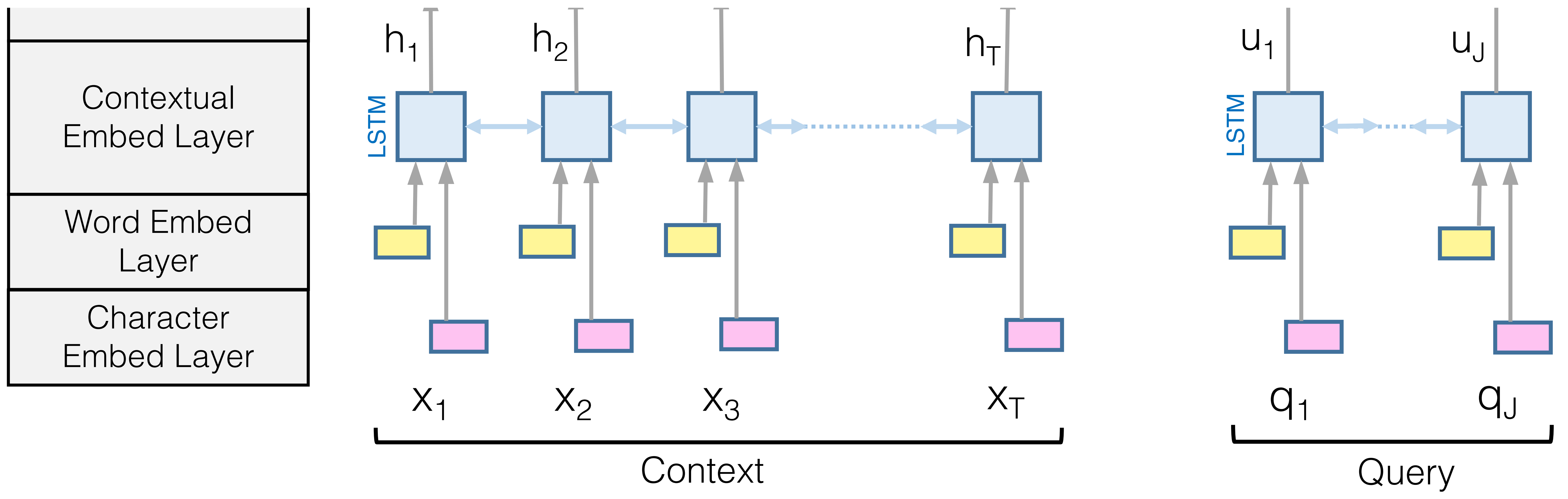
- Attentive Reader (Hermann et a. 2015)
- Bi-Directional Attention Flow (Seo et al. 2016)

More recently: fine-tuning BERT-like models for RC (2019+)

Bidirectional Attention Flow (BiDAF)

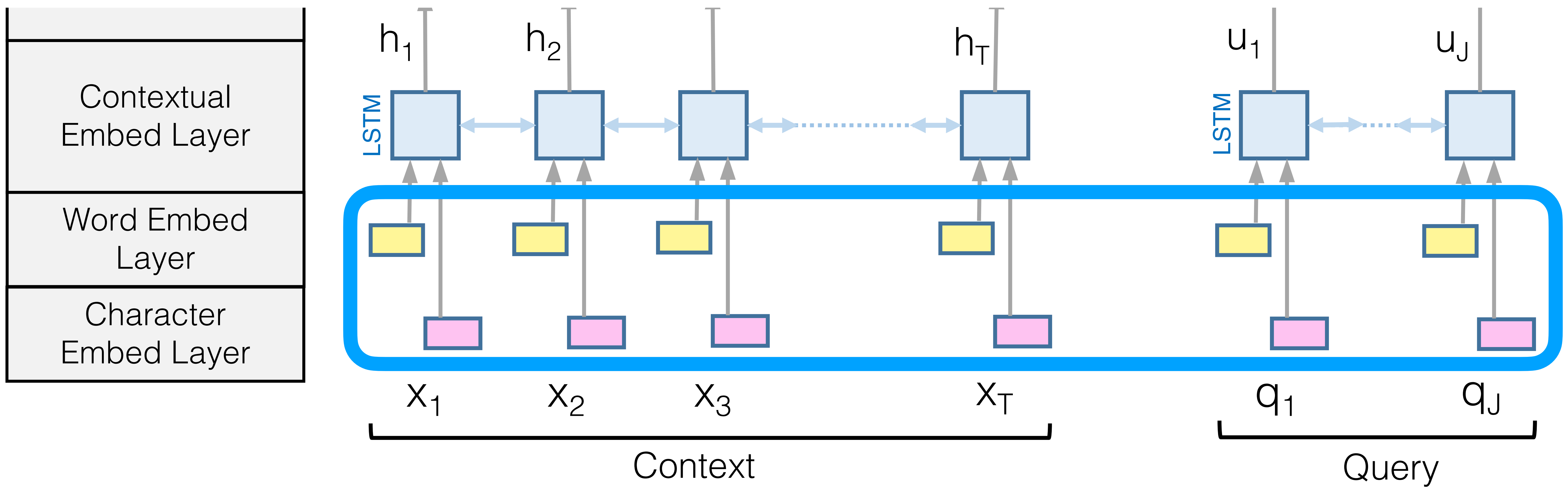


Bidirectional Attention Flow (BiDAF)



Contextual Encoder

Bidirectional Attention Flow (BiDAF)



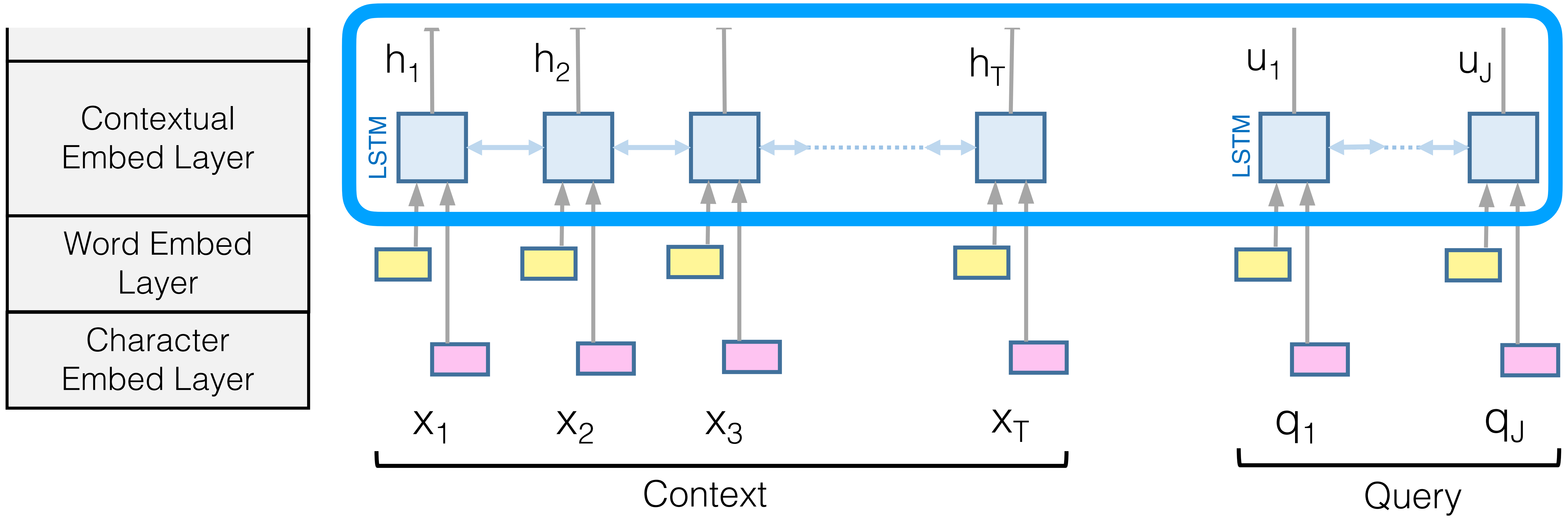
Contextual Encoder

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 $\text{emb}(x) = f\Big(\big[\text{GloVe}(x); \text{charEmb}(x)\big]\Big)$

Bidirectional Attention Flow (BiDAF)



Contextual Encoder

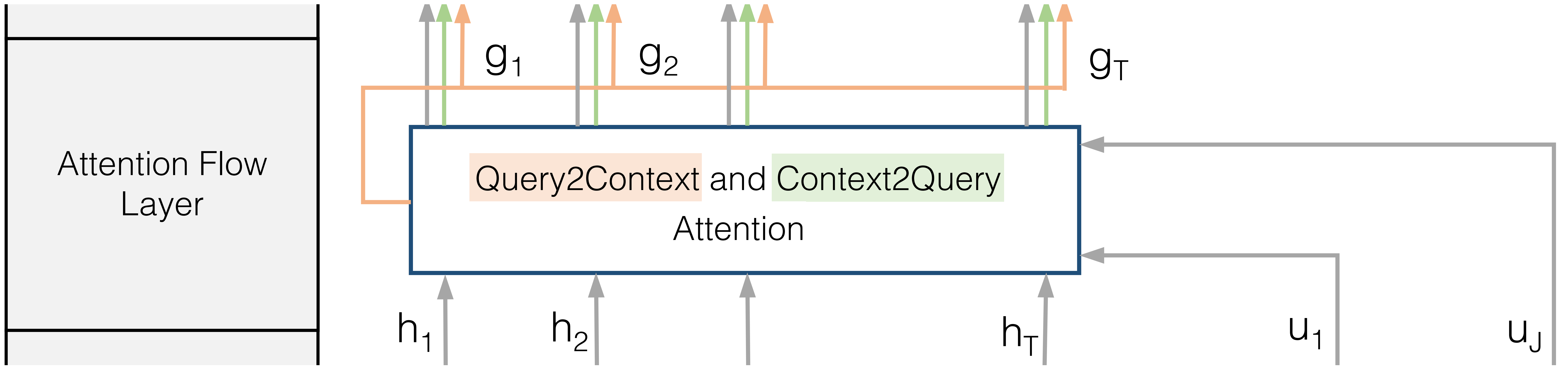
Contextual Embedding Layer: Concatenate word embeddings (e.g., GloVe) and character embedding for each word in the context and query:

$$e(c_i) = \text{emb}(c_i) \qquad e(q_i) = \text{emb}(q_i) \qquad \text{such that} \quad \text{emb}(x) = f\left(\left[\text{GloVe}(x); \text{charEmb}(x)\right]\right)$$

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

$$\mathbf{c} = \text{BiLSTM}\left(\left[e(c_1), \ldots, e(c_n)\right]\right) \qquad \mathbf{q} = \text{BiLSTM}\left(\left[e(q_1), \ldots, e(q_m)\right]\right)$$

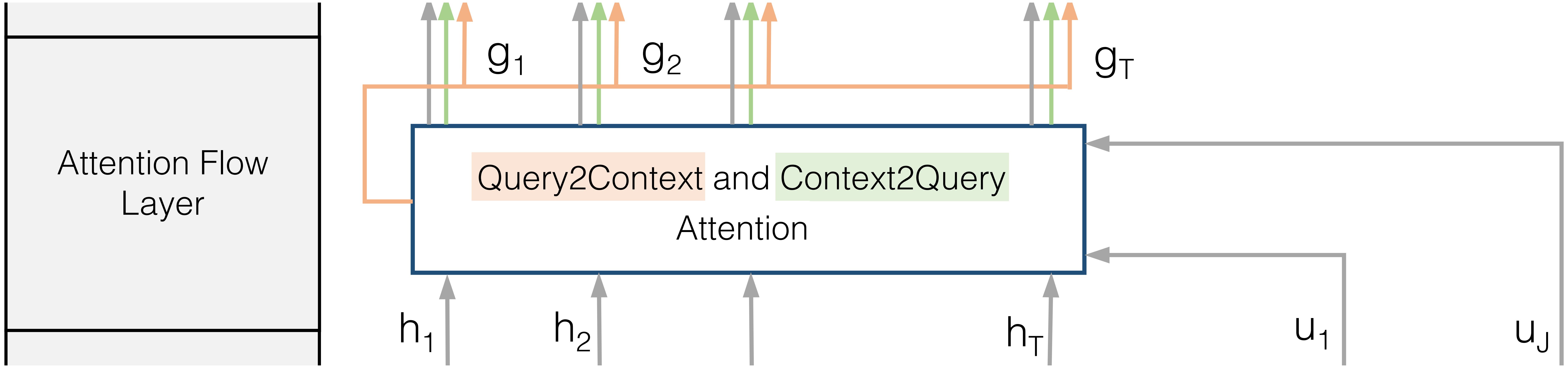
Bidirectional Attention Flow (BiDAF)



Context-to-Query Attention

Context-to-Query Attention: for each context word, choose the most relevant words from the question words.

Bidirectional Attention Flow (BiDAF)



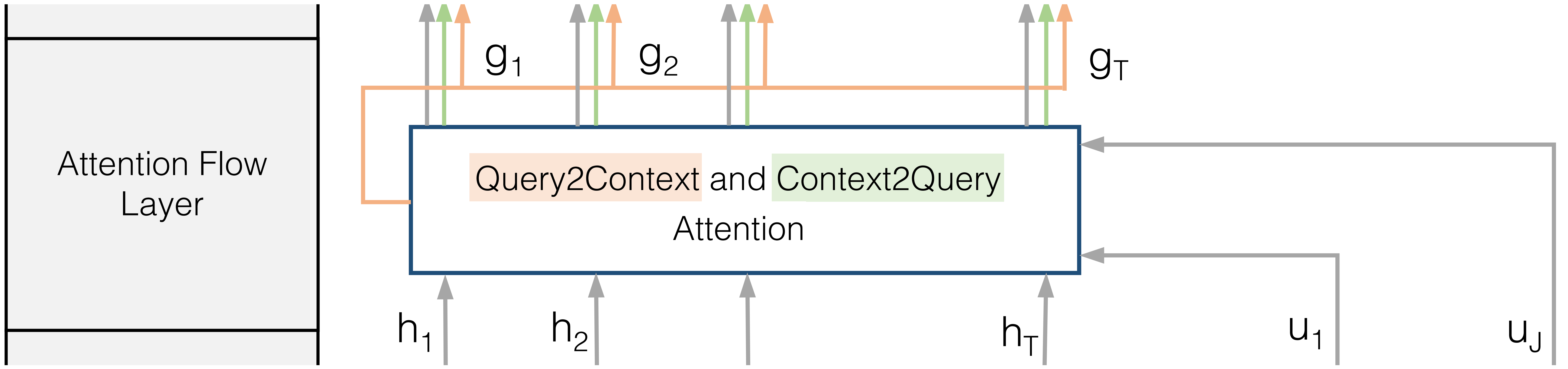
Context-to-Query Attention

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.

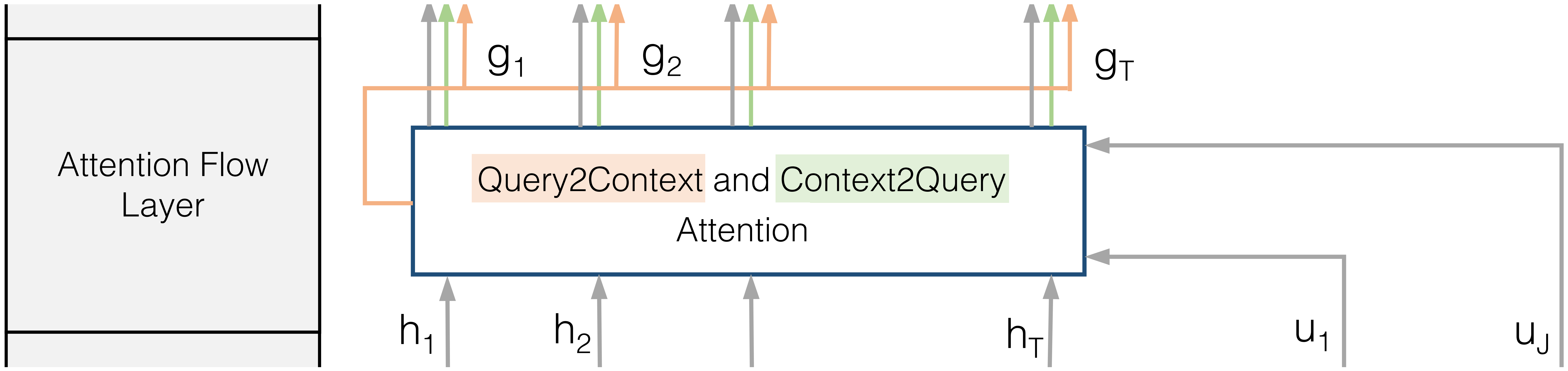
Bidirectional Attention Flow (BiDAF)



Query-to-Context Attention

Query-to-Context Attention: for each question word, choose the most relevant words from the context words.

Bidirectional Attention Flow (BiDAF)



Query-to-Context Attention

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 $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{ij} = \mathbf{w}_{\text{sim}}^{\top} \left[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j\right]$$

Bidirectional Attention Flow (BiDAF)

Compute a **similarity score** for every context-query token pair $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{ij} = \mathbf{w}_{\text{sim}}^{\top} \left[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j\right]$$

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

Distribution over question words

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

Bidirectional Attention Flow (BiDAF)

Compute a **similarity score** for every context-query token pair $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{ij} = \mathbf{w}_{\text{sim}}^{\top} \left[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j\right]$$

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

Distribution over question words

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

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

Distribution over context words

$$\beta_i = \text{softmax}\left(\text{max}_{\text{col}}\left(S_{ij}\right)\right), \quad \mathbf{b}_i = \sum_{i=1}^{n} \beta_i \mathbf{c}_i$$

Bidirectional Attention Flow (BiDAF)

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

Context-to-Query Attention

Query-to-Context Attention

$$\beta_i = \mathrm{softmax}\left(\max_{\mathrm{col}}\left(S_{ij}\right)\right), \quad \boxed{\mathbf{b}_i} = \sum_{i=1}^{n} \beta_i\mathbf{c}_i$$

Bidirectional Attention Flow (BiDAF)

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

Context-to-Query Attention

Query-to-Context Attention

$$\beta_i = \mathrm{softmax}\left(\max_{\mathrm{col}}\left(S_{ij}\right)\right), \quad \boxed{\mathbf{b}_i} = \sum_{i=1}^{n} \beta_i\mathbf{c}_i$$

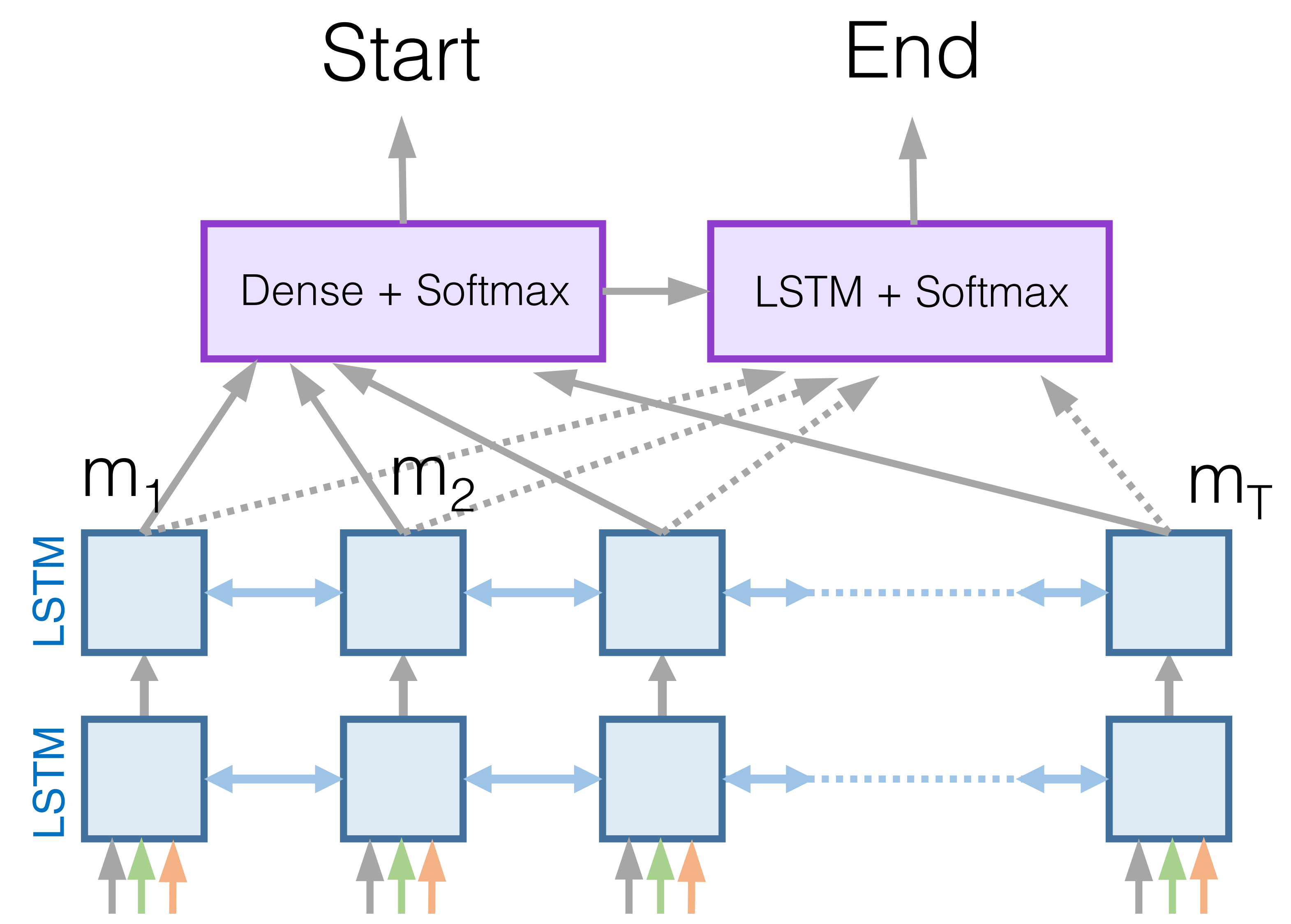
Output:

$$\boxed{\mathbf{g}_i} = \left[\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}_i\right]$$

Bidirectional Attention Flow (BiDAF)

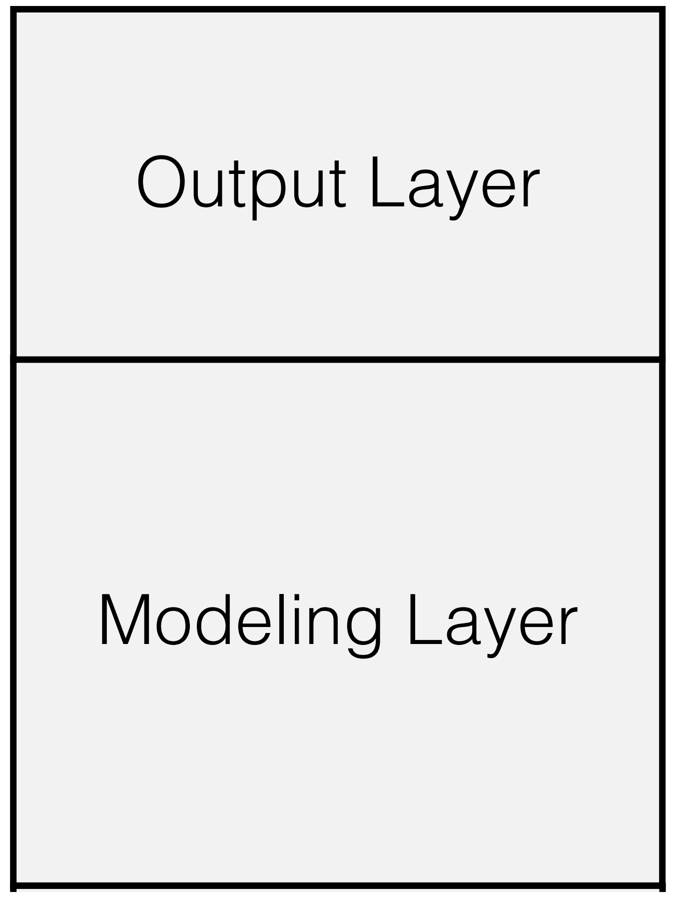
Output Layer

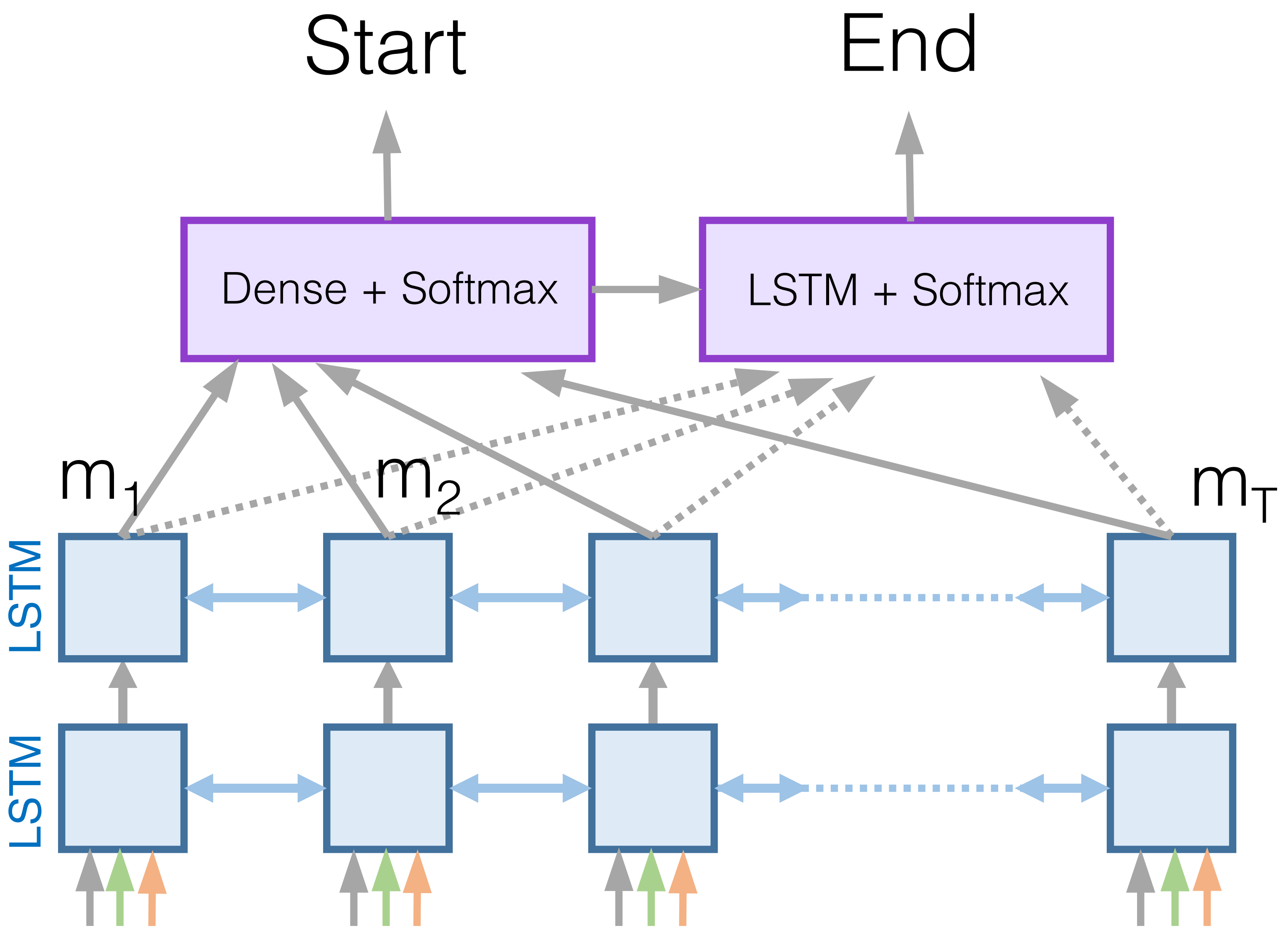
Modeling Layer



Modeling and Output Layers

Bidirectional Attention Flow (BiDAF)

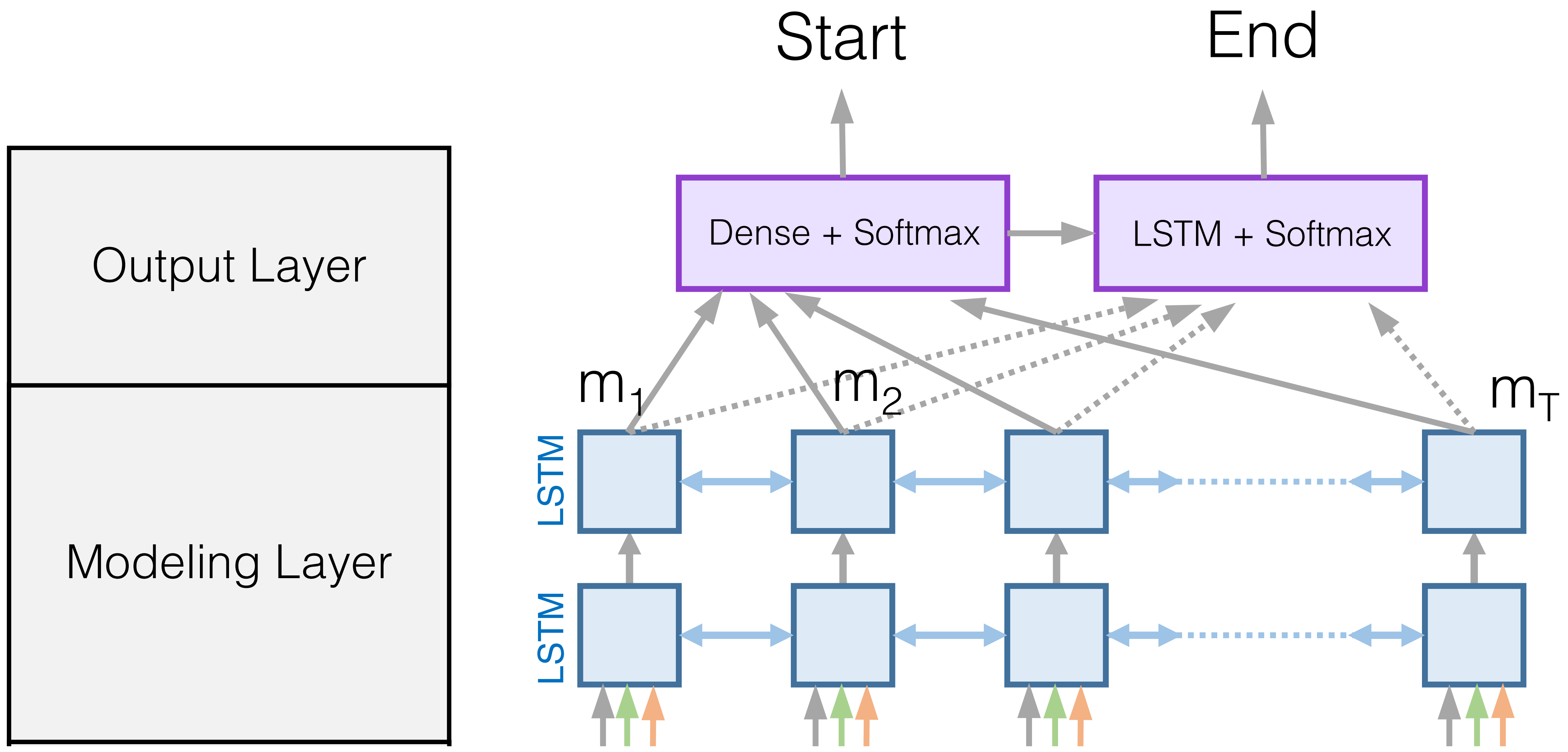




Modeling Layer: passes the activations $\mathbf{g}_i$ to a multi-layer bi-directional LSTM:

$$\mathbf{m}_i = \text{BiLSTM}(\mathbf{g}_i)$$

Bidirectional Attention Flow (BiDAF)



Modeling Layer: passes the activations $\mathbf{g}_i$ to a multi-layer bi-directional LSTM:

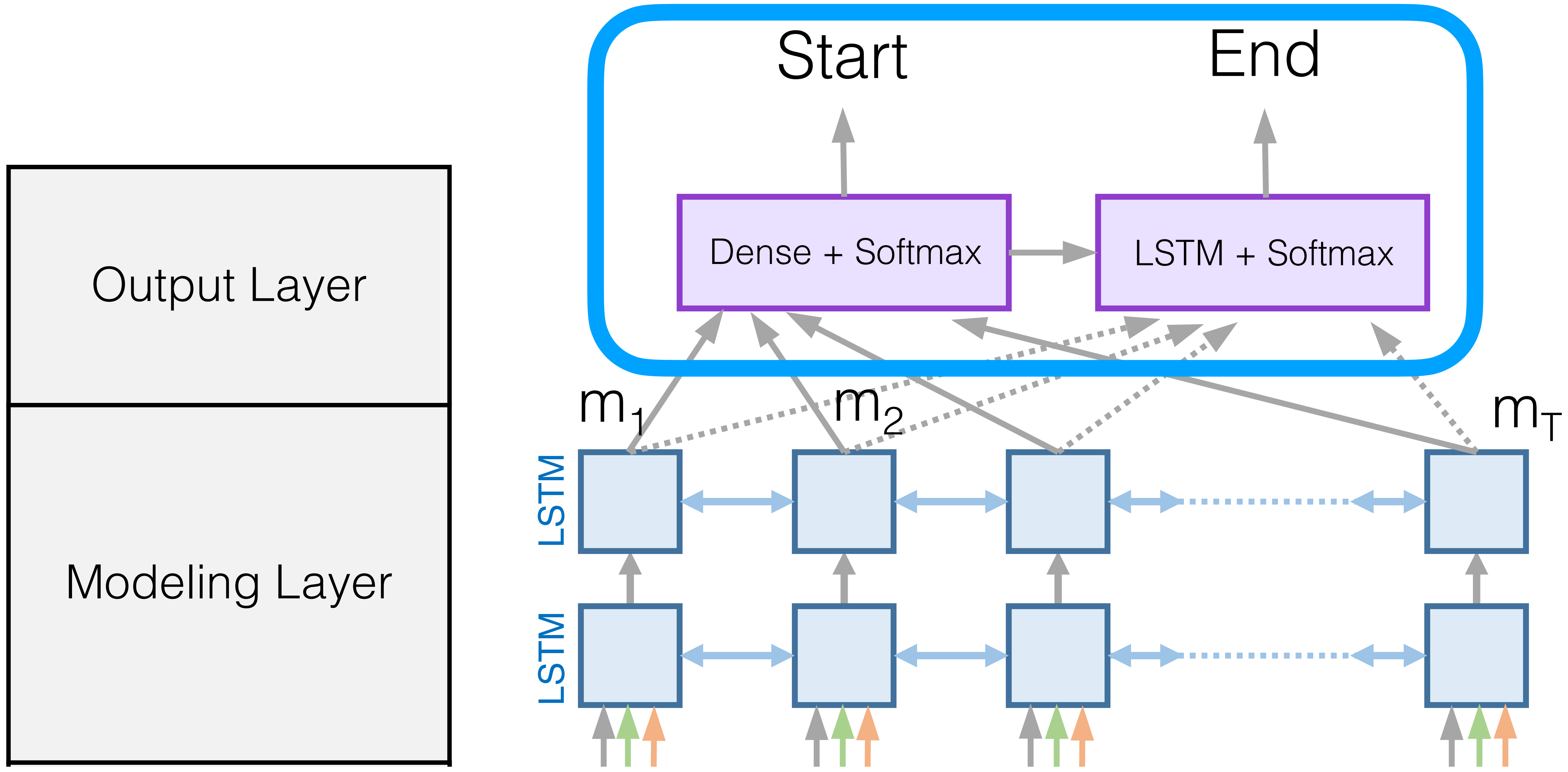
$$\mathbf{m}_i = \text{BiLSTM}(\mathbf{g}_i)$$

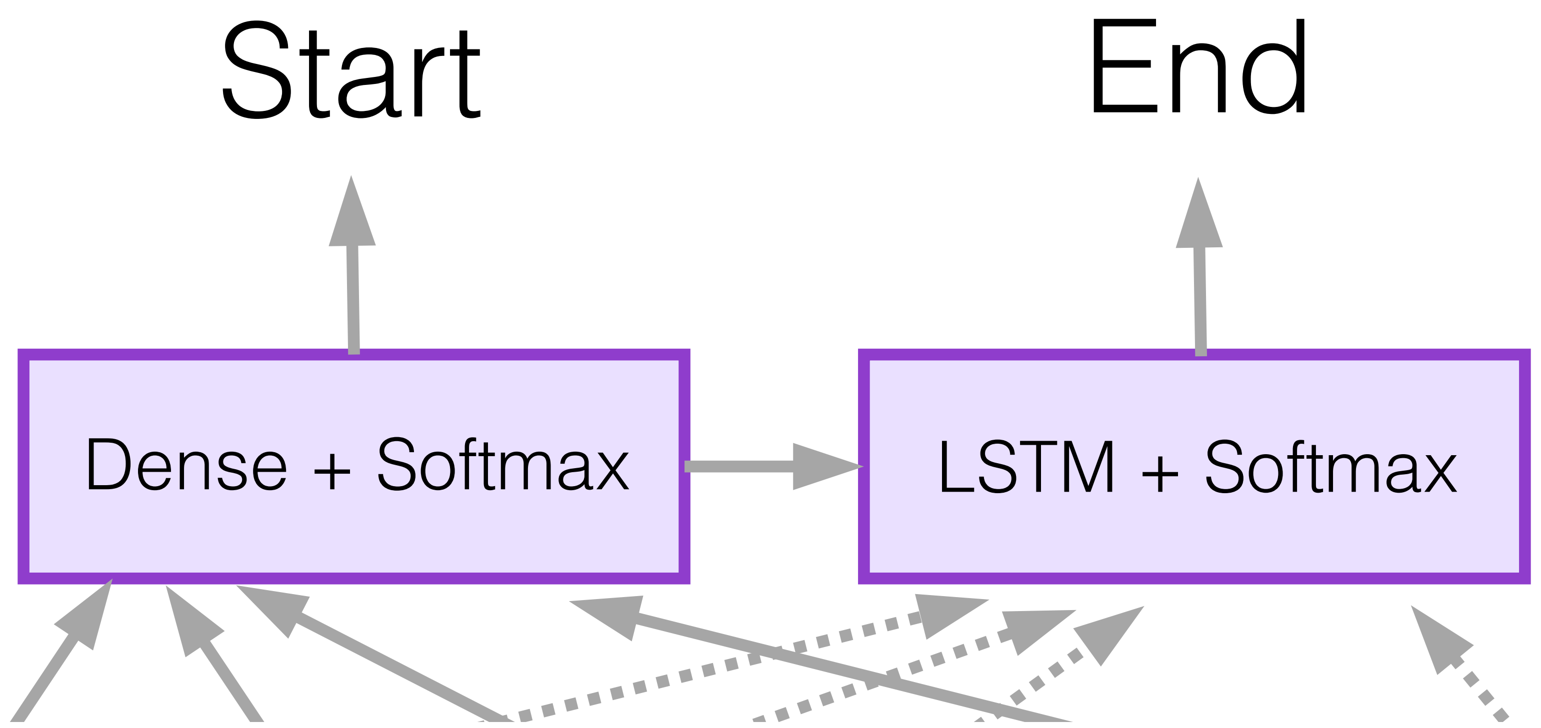
Output Layer: two classifiers predict the start and end positions of the answer:

$$P_{\text{start}} = \text{softmax}\left(\mathbf{w}_{\text{start}}^{\top}[\mathbf{g}_i; \mathbf{m}_i]\right) \quad P_{\text{end}} = \text{softmax}\left(\mathbf{w}_{\text{end}}^{\top}[\mathbf{g}_i; \mathbf{m}'_i]\right)$$

$$\text{with } \mathbf{m}'_i = \text{BiLSTM}(\mathbf{m}_i)$$

Bidirectional Attention Flow (BiDAF)





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

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

Answer start and end indices

Model parameters

Bidirectional Attention Flow (BiDAF)

Ablation:

- BiDAF: **77.3**

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

(a) Results on the SQuAD test set

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BiDAF (single)	67.7	77.3
BiDAF (ensemble)	72.6	80.7

(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

	Single Model EM	Single Model F1	Ensemble EM	Ensemble F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BIDAF (Ours)	68.0	77.3	**73.3**	**81.1**

(a) Results on the SQuAD test set

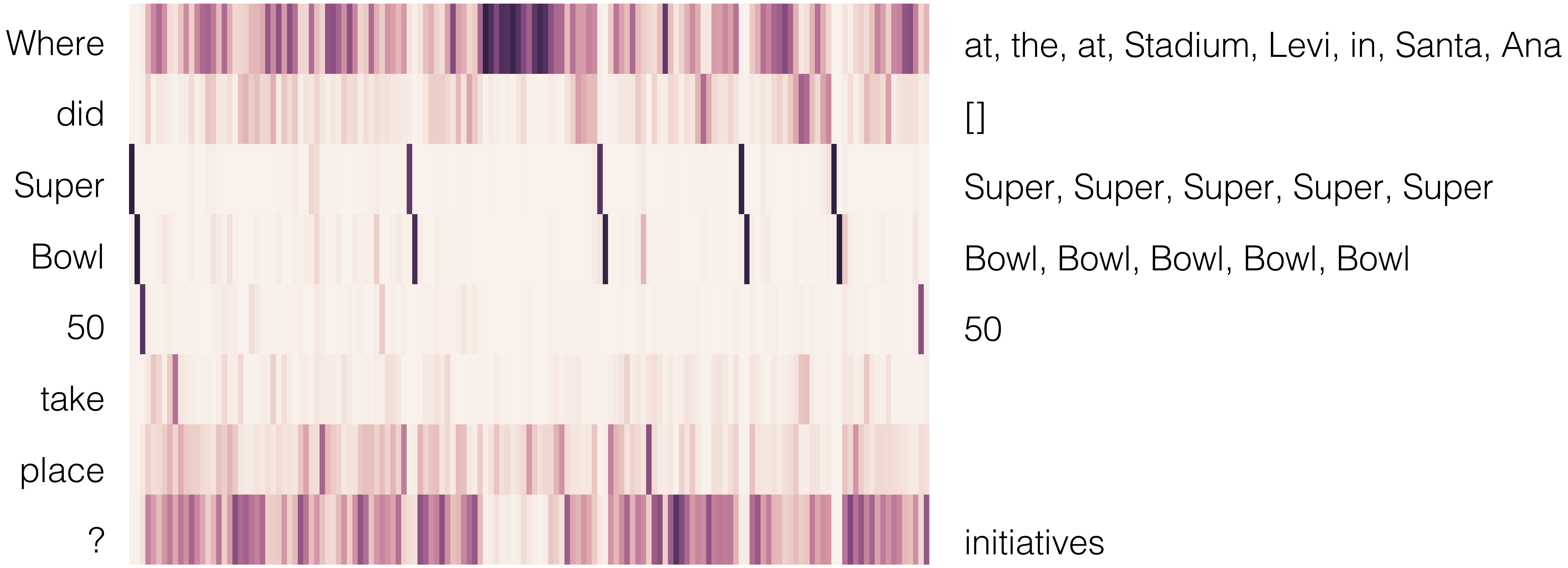
	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

(b) Ablations on the SQuAD dev set

Bidirectional Attention Flow (BiDAF)

Attention Visualisation

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

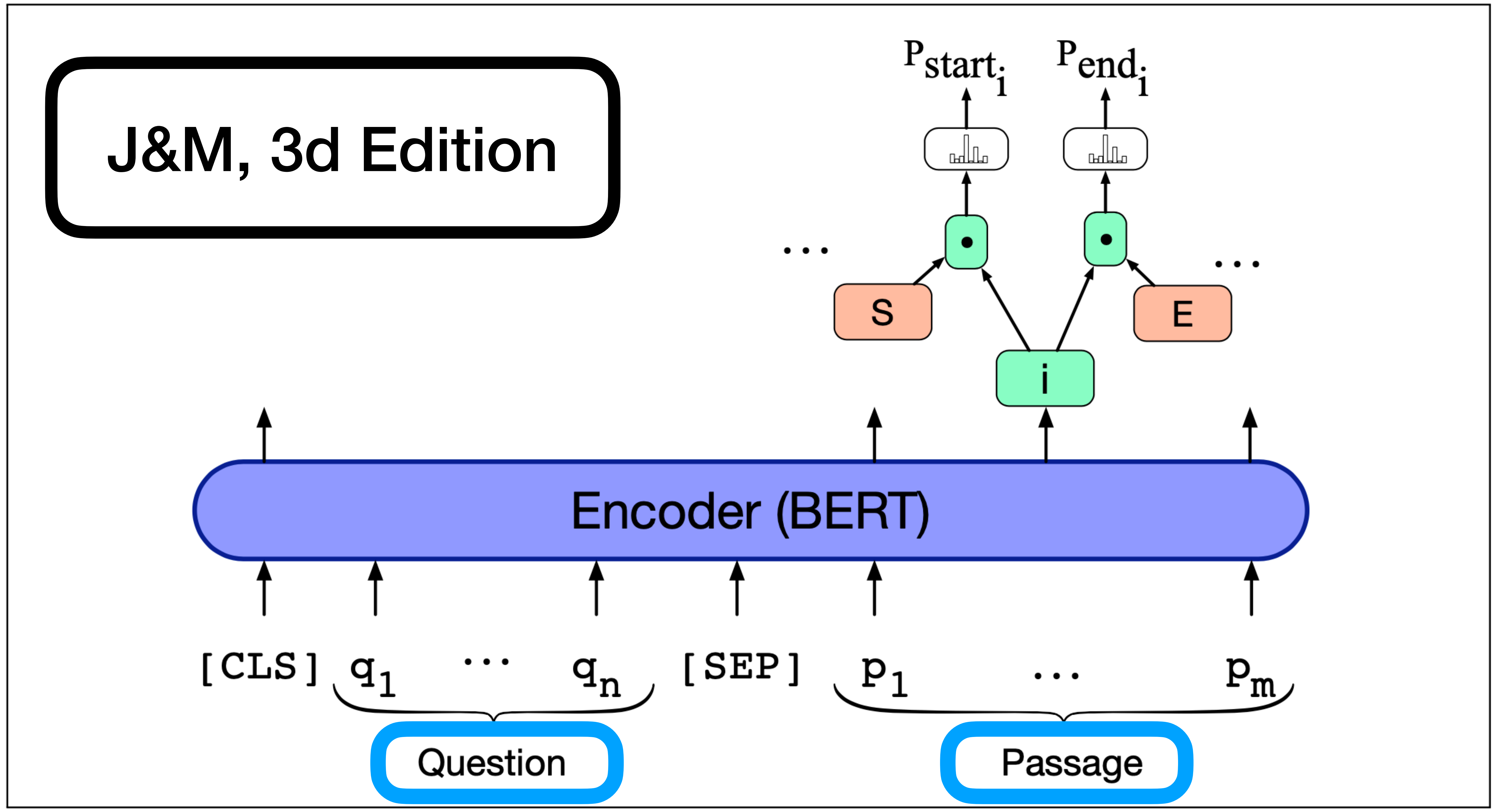


Figure 14.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.

Input: Question [sep] Passage

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

BERT-based Span-Based QA Models

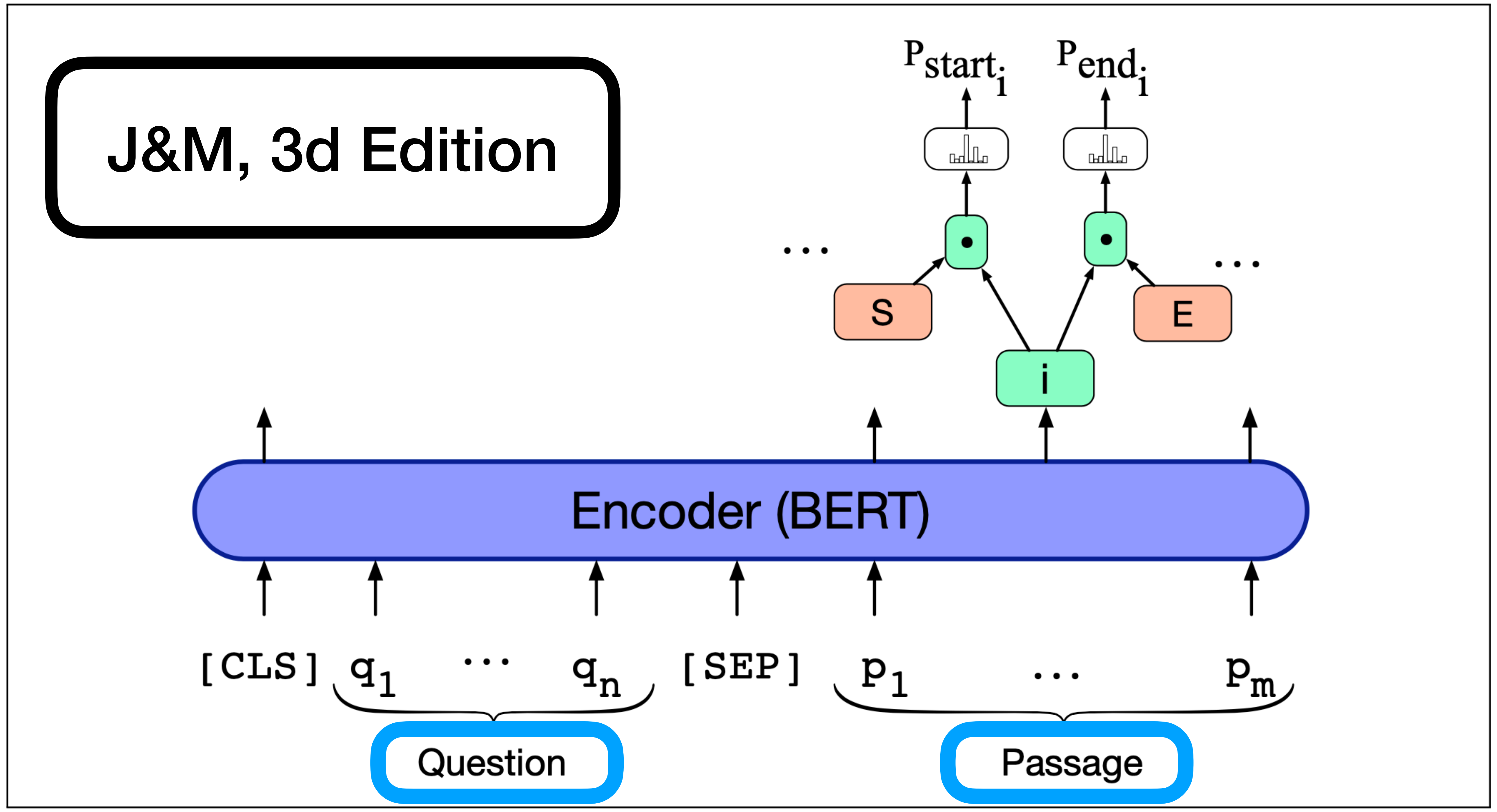


Figure 14.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.

Input: Question [sep] Passage

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

$$P_{\text{start}}(c_i) = \text{softmax}\left(\mathbf{w}_{\text{start}}^{\top}\mathbf{h}_i\right)$$

$$P_{\text{end}}(c_i) = \text{softmax}\left(\mathbf{w}_{\text{end}}^{\top}\mathbf{h}_i\right)$$

$\mathbf{h}_i$ is the contextual representation of c_i produced by BERT

BERT-based Span-Based QA Models

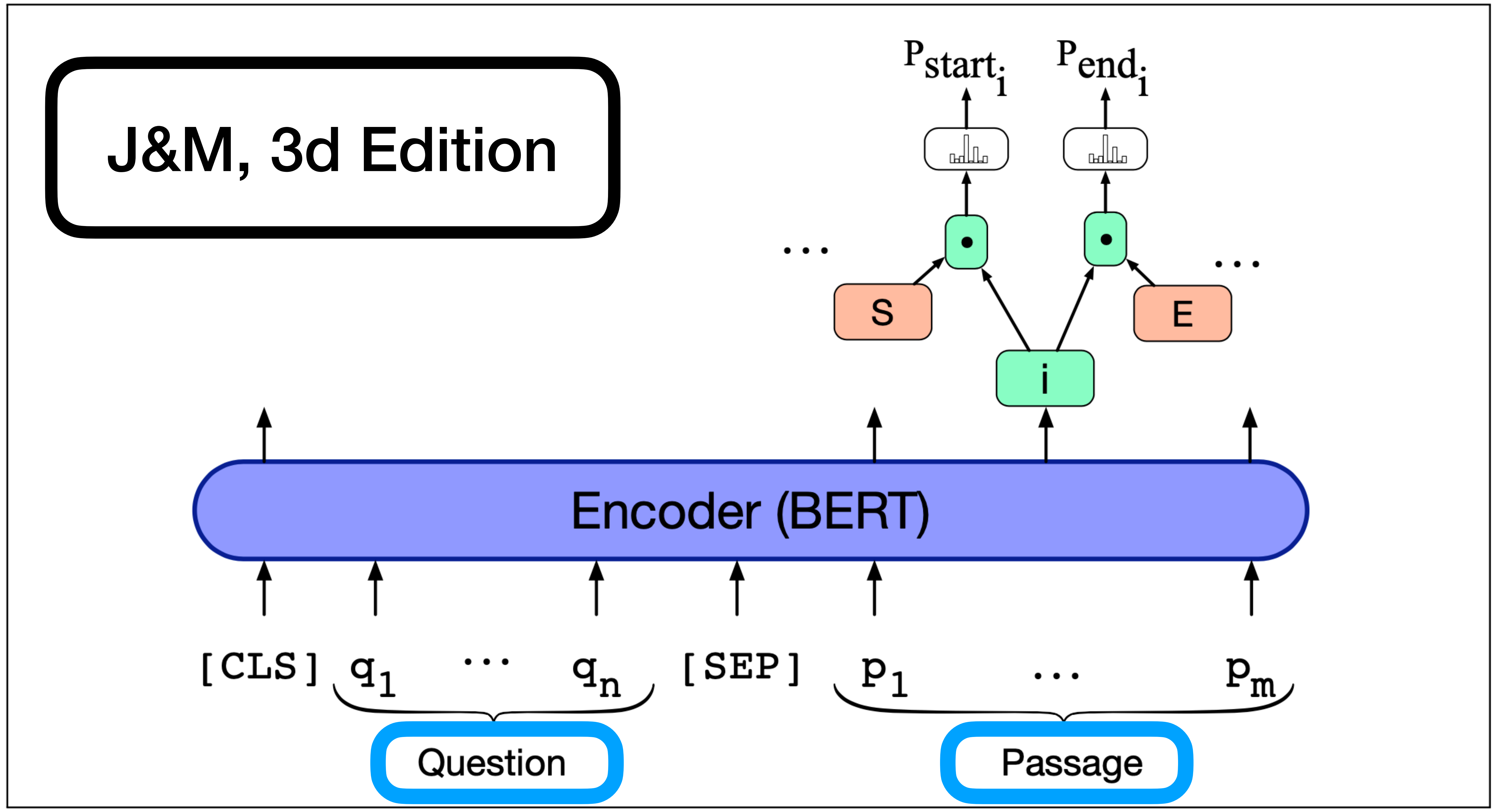


Figure 14.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.

Input: Question [sep] Passage

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

$$P_{\text{start}}(c_i) = \text{softmax}\left(\mathbf{w}_{\text{start}}^{\top}\mathbf{h}_i\right)$$

$$P_{\text{end}}(c_i) = \text{softmax}\left(\mathbf{w}_{\text{end}}^{\top}\mathbf{h}_i\right)$$

$\mathbf{h}_i$ 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 θ

BERT-based Span-Based QA Models

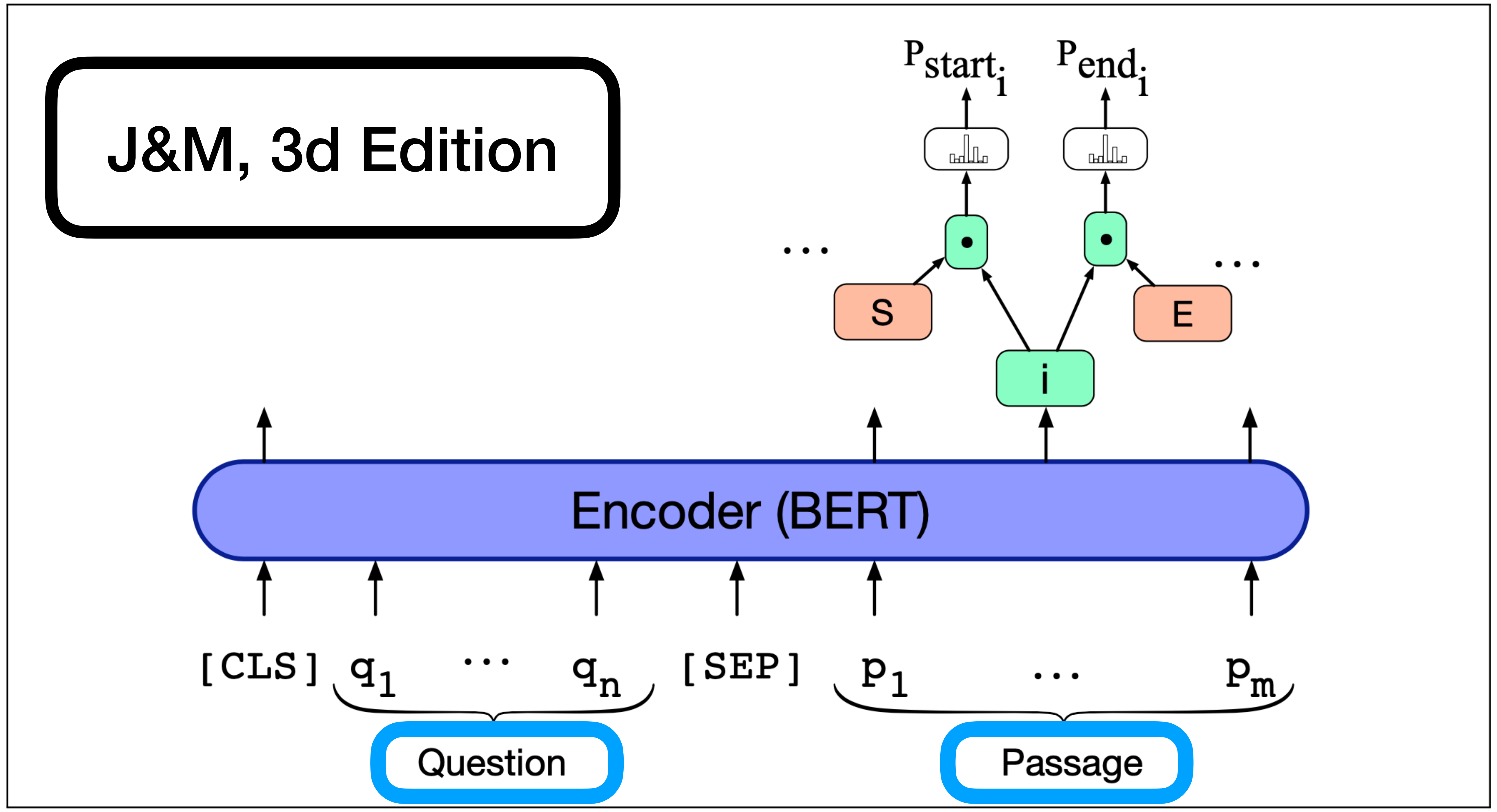


Figure 14.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.

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

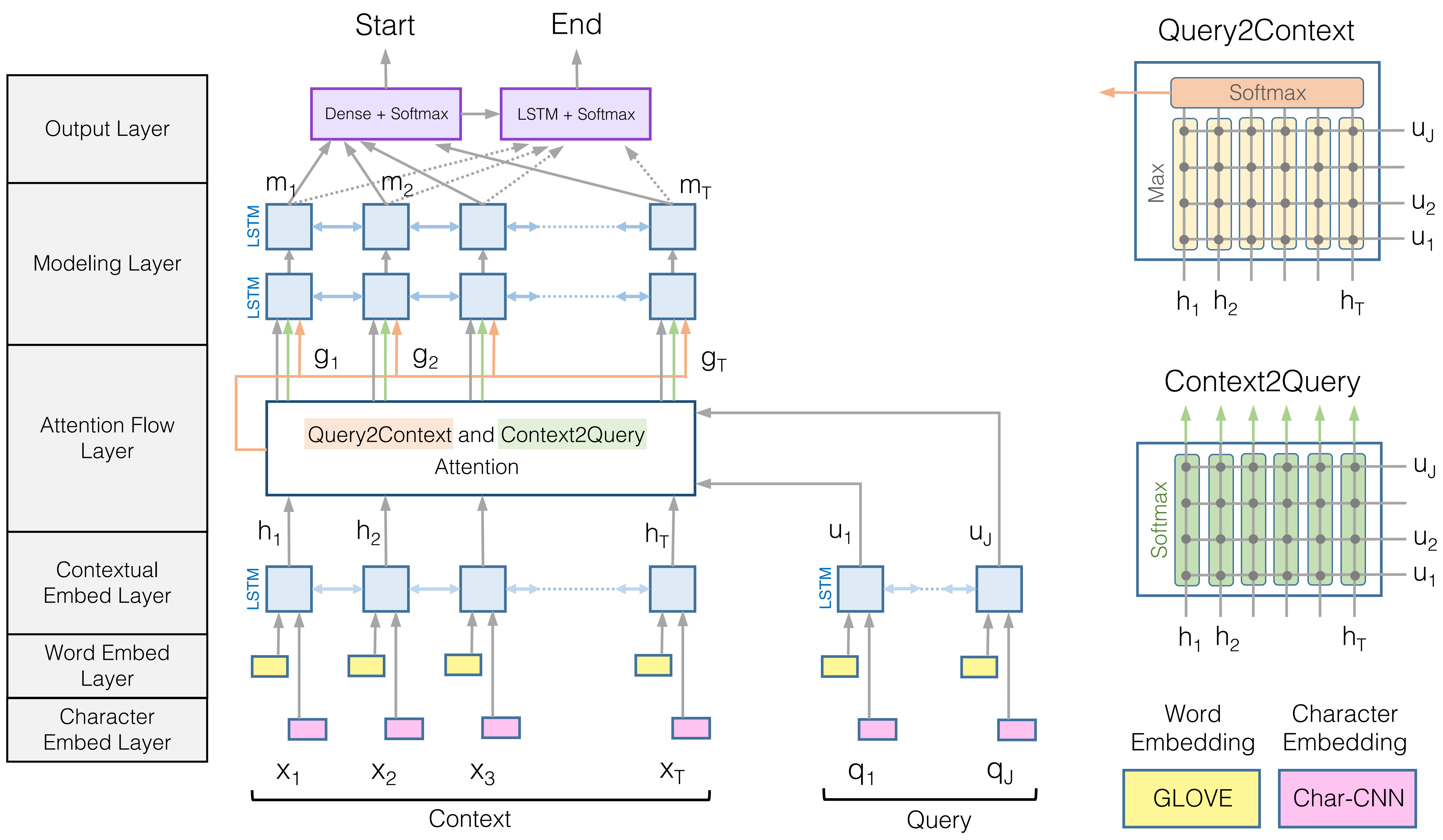
	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1

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

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

BERT Parameters

BiDAF vs. BERT-based Models



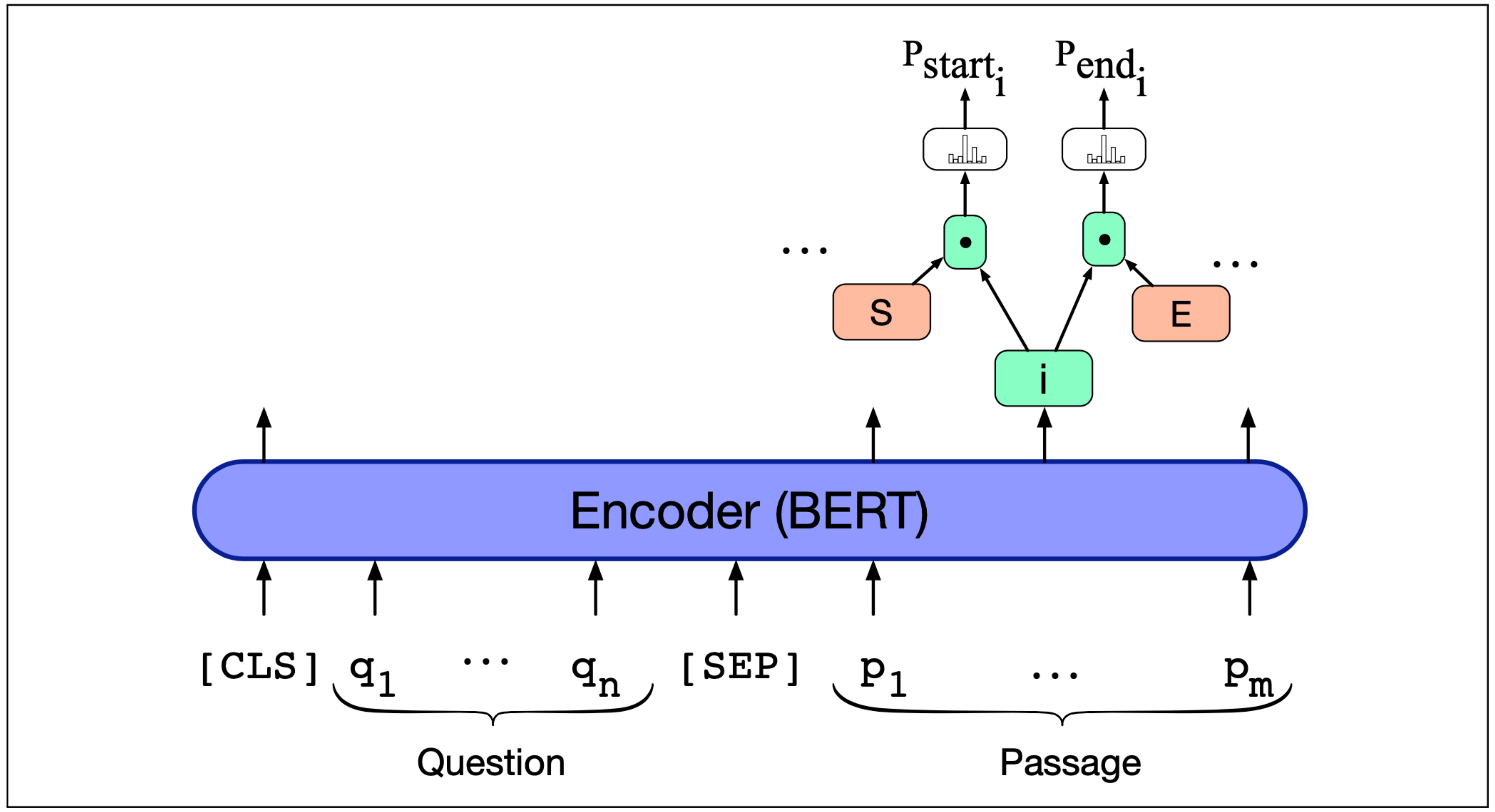


Figure 14.12 An encoder model (using BERT) for span-based question answering from reading-comprehension-based question answering tasks.

~2.5M params

Several BiLSTMs

Trained from scratch (minus GloVe)

110M-330M params

Transformers (**no recurrence**)

Pre-Trained

Natural Questions, Annotation Task

Question: when was the egg mcmuffin added to the menu

Question from the Google Query Stream

Wikipedia page:

McMuffin

11 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

McMuffin is a family of breakfast sandwiches sold by the international fast food restaurant chain McDonald's. The **Egg McMuffin** is the signature sandwich, which was invented in 1972 by Herb Peterson to resemble eggs benedict, a traditional American breakfast dish with English muffins, ham, eggs and hollandaise sauce.[1]

Product description

In the US and Canada the standard McMuffin consists of a slice of Canadian bacon,[2] 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.[3][1]

History

The sandwich was invented in 1972.[4] Former McDonald's President Ray Kroc wrote that Herb Peterson and his assistant, Donald Greadel, 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.[3][1]

McMuffin

Nutritional value per 1 sandwich, 7.1 oz (200 g)		
Energy	300 kcal (1,300 kJ)	
Carbohydrates	30 g (10%)	
Sugars	16 g	
Dietary fiber	2 g (8%)	
Fat	18 g (37%)	
Saturated	5 g (24%)	
Trans	0†	
Protein	18 g	
Vitamins	Quantity	%DV†
Vitamin A equiv.	90 μg	11%
Vitamin C	0 mg	0%
Vitamin E	0 mg	0%

Natural Questions, Annotation Task

Question: when was the egg mcmuffin added to the menu

Question from the Google Query Stream

Wikipedia page:

McMuffin

11 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

McMuffin is a family of breakfast sandwiches sold by the international fast food restaurant chain McDonald's. The **Egg McMuffin** is the signature sandwich, which was invented in 1972 by Herb Peterson to resemble eggs benedict, a traditional American breakfast dish with English muffins, ham, eggs and hollandaise sauce.[1]

Product description

In the US and Canada the standard McMuffin consists of a slice of Canadian bacon,[2] 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.[3][1]

History

The sandwich was invented in 1972.[4] Former McDonald's President Ray Kroc wrote that Herb Peterson and his assistant, Donald Greadel, 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.[3][1]

McMuffin

Nutritional value per 1 sandwich, 7.1 oz (200 g)		
Energy	300 kcal (1,300 kJ)	
Carbohydrates	30 g (10%)	
Sugars	16 g	
Dietary fiber	2 g (8%)	
Fat	18 g (37%)	
Saturated	5 g (24%)	
Trans	0†	
Protein	18 g	
Vitamins	Quantity	%DV†
Vitamin A equiv.	90 μg	11%
Vitamin C	0 mg	0%
Vitamin E	0 mg	0%

Step 1: Annotator selects context

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

Question from the Google Query Stream

Wikipedia page:

McMuffin

11 languages

Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

McMuffin is a family of breakfast sandwiches sold by the international fast food restaurant chain McDonald's. The **Egg McMuffin** is the signature sandwich, which was invented in 1972 by Herb Peterson to resemble eggs benedict, a traditional American breakfast dish with English muffins, ham, eggs and hollandaise sauce.[1]

Product description

In the US and Canada the standard McMuffin consists of a slice of Canadian bacon,[2] 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.[3][1]

History

The sandwich was invented in 1972.[4] Former McDonald's President Ray Kroc wrote that Herb Peterson and his assistant, Donald Greadel, 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.[3][1]

McMuffin

Nutritional value per 1 sandwich, 7.1 oz (200 g)		
Energy	300 kcal (1,300 kJ)	
Carbohydrates	30 g (10%)	
Sugars	16 g	
Dietary fiber	2 g (8%)	
Fat	18 g (37%)	
Saturated	5 g (24%)	
Trans	0†	
Protein	18 g	
Vitamins	**Quantity**	**%DV†**
Vitamin A equiv.	90 μg	11%
Vitamin C	0 mg	0%
Vitamin E	0 mg	0%

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

Natural Questions, Example

Question: what might you find on a mayan monument

Wikipedia page:

Maya stelae

11 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Maya stelae (singular ***stela***) are monuments that were fashioned by the Maya civilization of ancient Mesoamerica. They consist of tall, sculpted stone shafts and are often associated with low circular stones referred to as altars, although their actual function is uncertain.[2] Many stelae were sculpted in low relief,[3] although plain monuments are found throughout the Maya region.[4] The sculpting of these monuments spread throughout the Maya area during the Classic Period (250–900 AD),[2] and these pairings of sculpted stelae and circular altars are considered a hallmark of Classic Maya civilization.[5] The earliest dated stela to have been found *in situ* in the Maya lowlands was recovered from the great city of Tikal in Guatemala.[6] During the Classic Period almost every Maya kingdom in the southern lowlands raised stelae in its ceremonial centre.[4]

Stela 51 from Calakmul, dating to 731, is the best preserved monument from the city. It depicts the king Yuknoom Took' K'awiil.[1]

Stela H, a high-relief in-the-round sculpture from Copán in Honduras

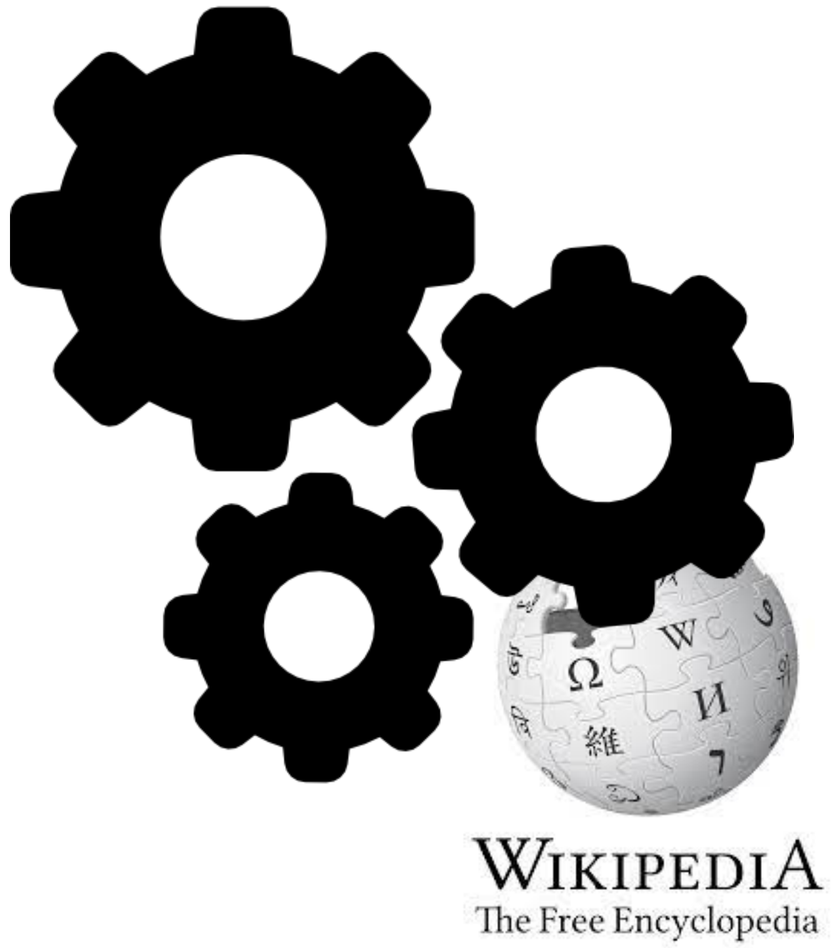
Stelae became closely associated with the concept of divine kingship and declined at the same time as this institution. The production of stelae by the Maya had its origin around 400 BC and continued through to the end of the Classic Period, around 900, although some monuments were reused in the Postclassic (c. 900–1521). The major city of Calakmul in Mexico raised the greatest number of stelae known from any Maya city, at least 166, although they are very poorly preserved.[7]

Hundreds of stelae have been recorded in the Maya region,[8] displaying a wide stylistic variation.[4] Many are upright slabs of limestone sculpted on one or more faces,[4] with available surfaces sculpted with figures carved in relief and with hieroglyphic text.[3] Stelae in a few sites display a much more three-dimensional appearance where locally available stone permits, such as at Copán and Toniná.[4] Plain stelae do not appear to have been painted nor overlaid with stucco decoration,[9] but most Maya stelae were probably brightly painted in red, yellow, black, blue and other colours.[10]

Stelae were essentially stone banners raised to glorify the king and record his deeds,[11] although the earliest examples depict mythological scenes.[12] Imagery developed throughout the Classic Period, with Early Classic stelae (c. 250–600) displaying non-Maya characteristics from the 4th century onwards, with the introduction of imagery linked to the central Mexican metropolis of Teotihuacan.[13] This influence receded in the 5th century although

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

Question (Q) → 

→ Answer (A)

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!

Reading List

Speech and Language Processing Ed. 3, Ch. 14 on QA 🙂

https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

SQuAD: 100,000+ Questions for Machine Comprehension of Text, **https://arxiv.org/abs/1606.05250** (SQuAD)

(Optional) Know What You Don't Know: Unanswerable Questions for SQuAD, **https://arxiv.org/abs/1806.03822** (SQuAD v2)

Bidirectional Attention Flow for Machine Comprehension, **https://arxiv.org/abs/1611.01603** (BiDAF)

Natural Questions: A Benchmark for Question Answering Research, **https://aclanthology.org/Q19-1026/** (NQ)

Latent Retrieval for Weakly Supervised Open Domain Question Answering **https://arxiv.org/abs/1906.00300**

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