Natural Language Processing with Deep Learning
CS224N/Ling284

Christopher Manning

based on slides by Danqi Chen, Princeton University

Lecture 11: Question Answering

Lecture plan

1. What is question answering? (10 mins)

2. Reading comprehension (50 mins) Your default final project!

   ✓ How to answer questions over **a single passage of text**

3. Open-domain (textual) question answering (20 mins)

   ✓ How to answer questions over **a large collection of documents**

Due today:
Ass 4
Final project proposal
Hopefully Azure is working okay for everyone now 🤞

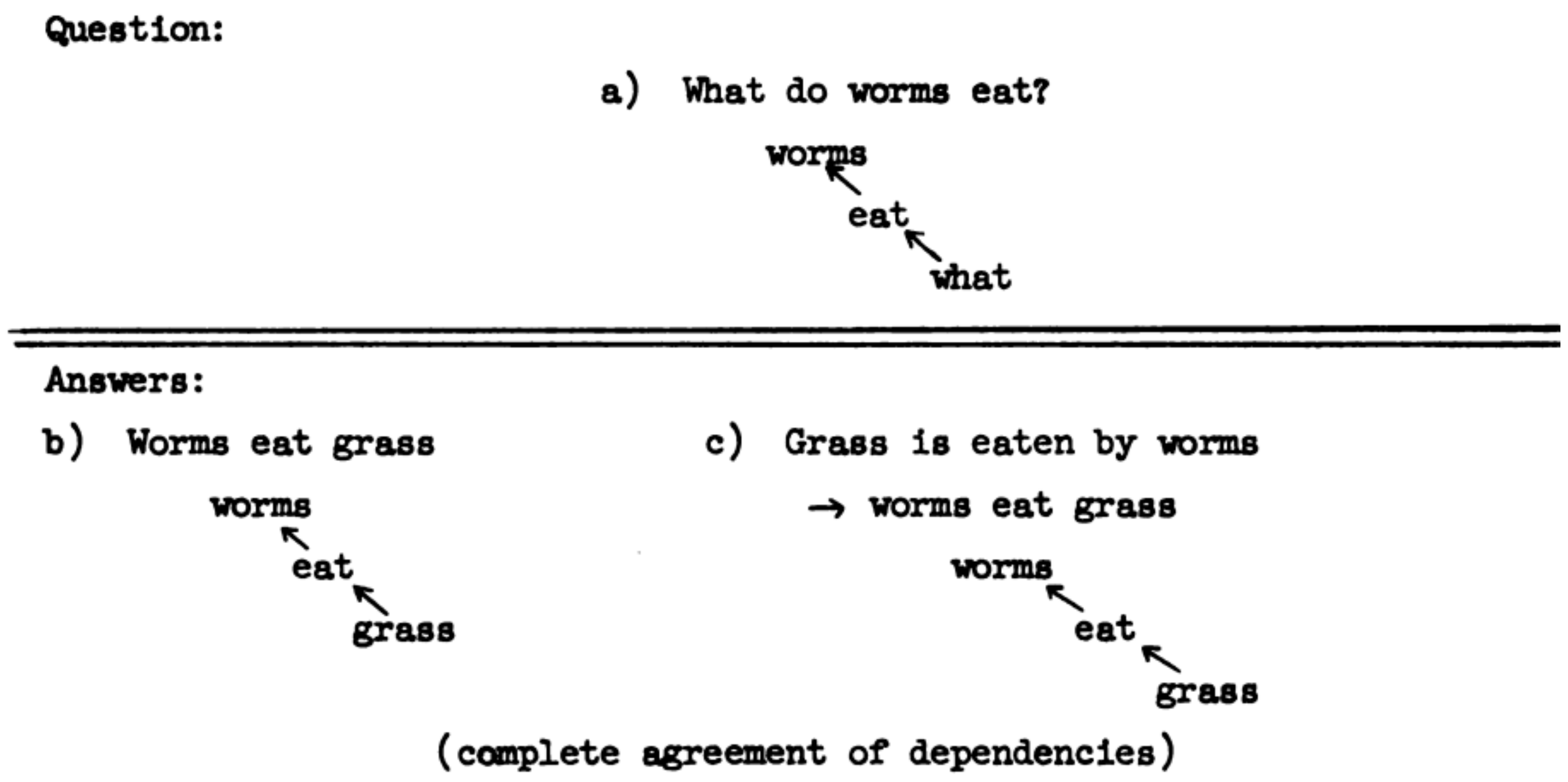
1. What is question answering?

Question (Q) →  → Answer (A)

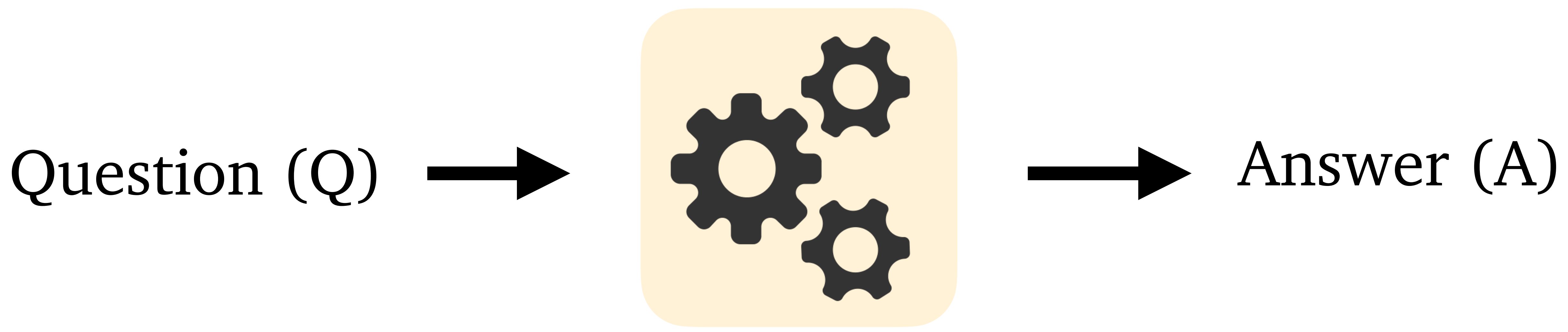
The goal of question answering is to build systems that **automatically** answer questions posed by humans in a **natural language**

The earliest QA systems dated back to 1960s!

(Simmons et al., 1964)



Question answering: a taxonomy



- What information source does a system build on?
  - A text passage, all Web documents, knowledge bases, tables, images..
- Question type
  - Factoid vs non-factoid, open-domain vs closed-domain, simple vs compositional, ..
- Answer type
  - A short segment of text, a paragraph, a list, yes/no, …

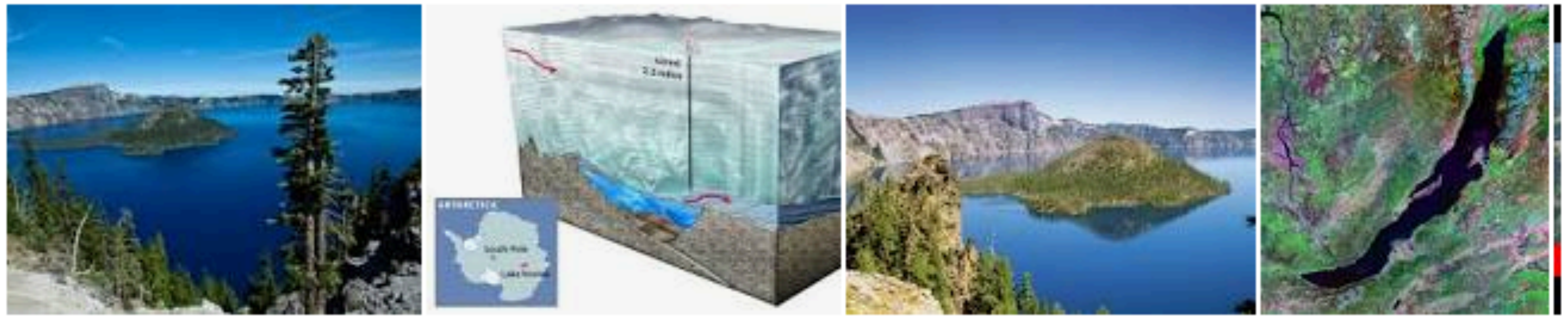
Lots of practical applications

Google

Where is the deepest lake in the world?

All Maps Images News Videos More Settings Tools

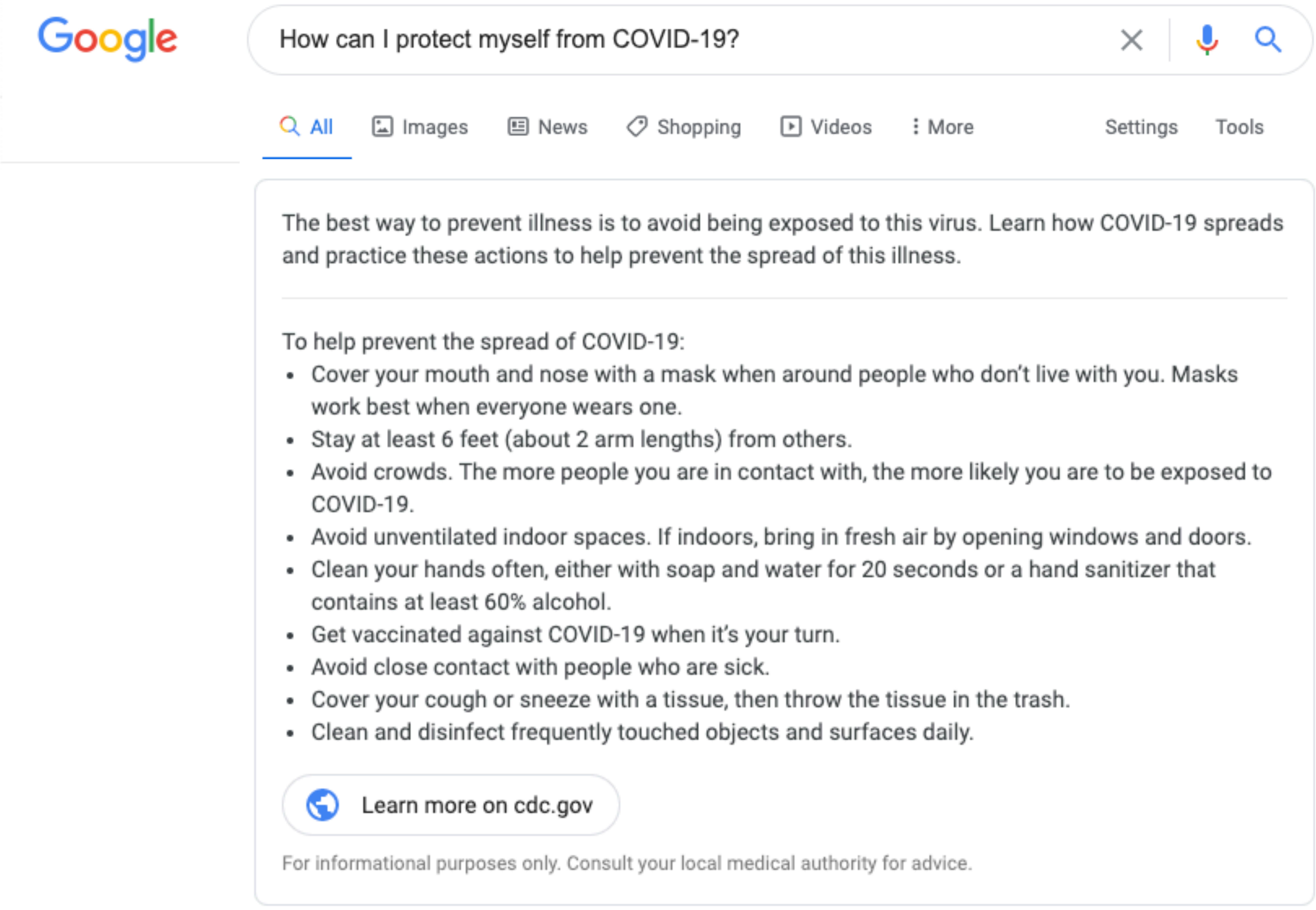
About 21,100,000 results (0.71 seconds)



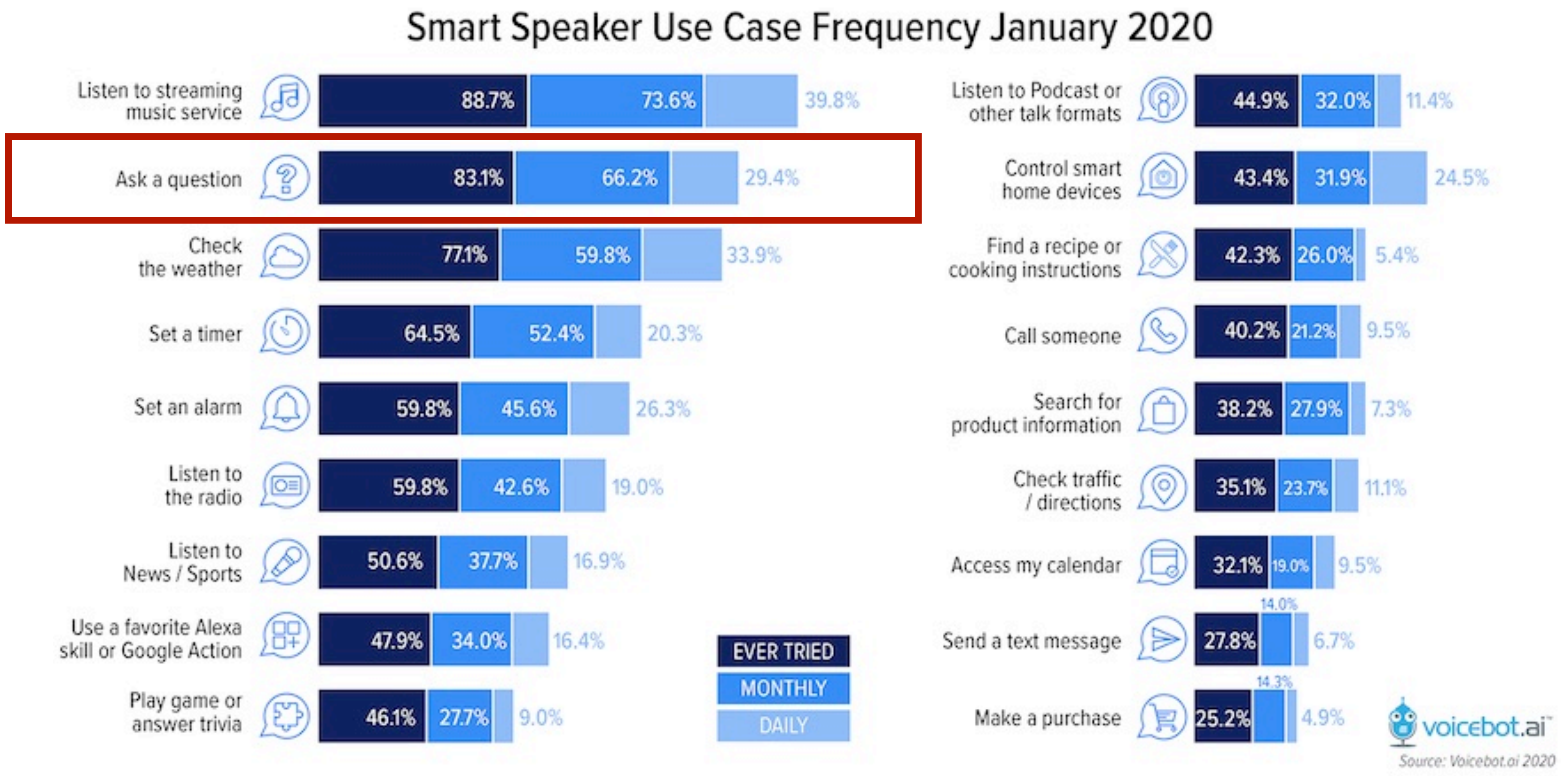
Siberia

Lake **Baikal**, in Siberia, holds the distinction of being both the deepest lake in the world and the largest freshwater lake, holding more than 20% of the unfrozen fresh water on the surface of Earth.

Lots of practical applications



Lots of practical applications



2011: IBM Watson beat Jeopardy champions

IBM Watson defeated two of Jeopardy's greatest champions in 2011

IBM Watson beat Jeopardy champions

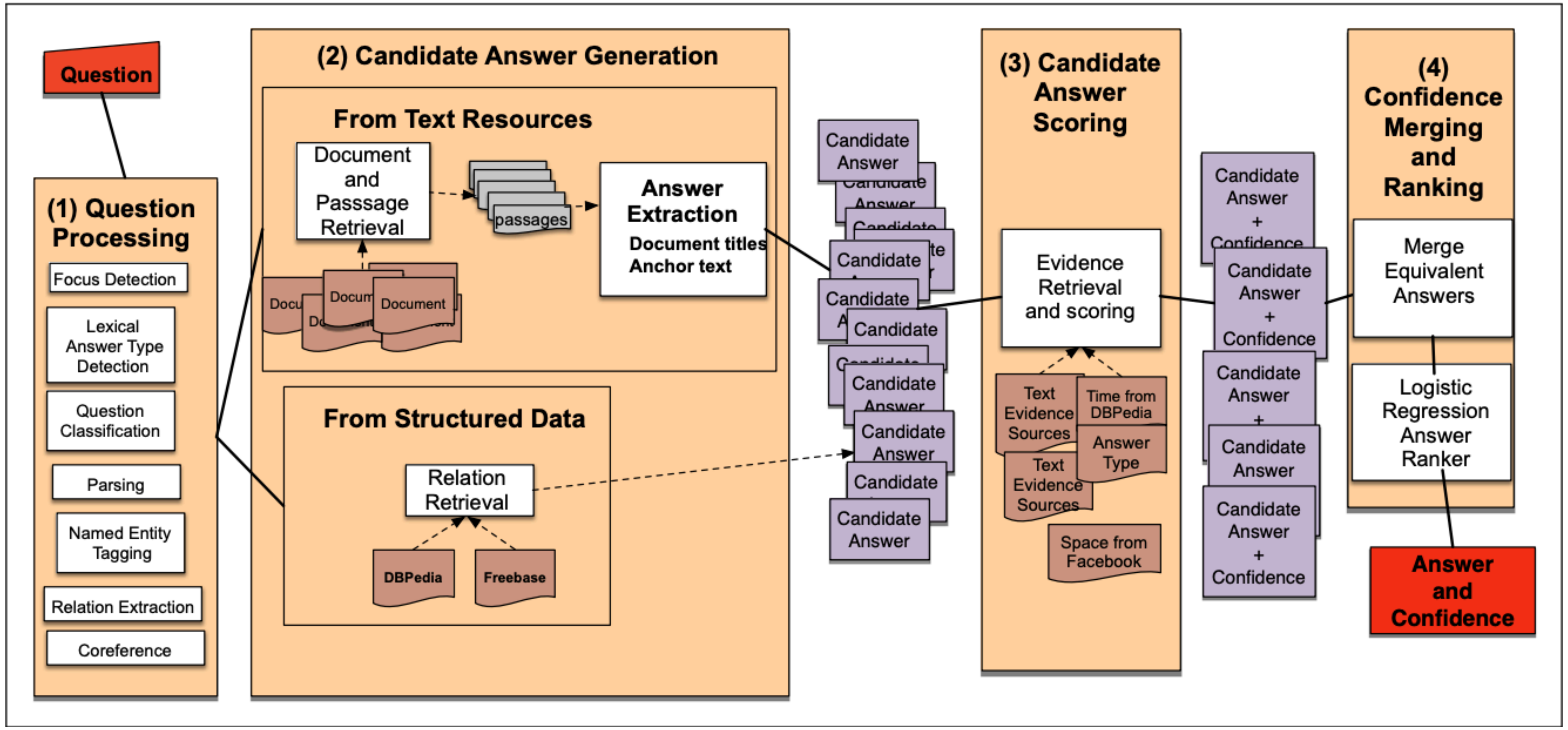


Image credit: J & M, edition 3

(1) Question processing, (2) Candidate answer generation, (3) Candidate answer scoring, and (4) Confidence merging and ranking.

Question answering in deep learning era

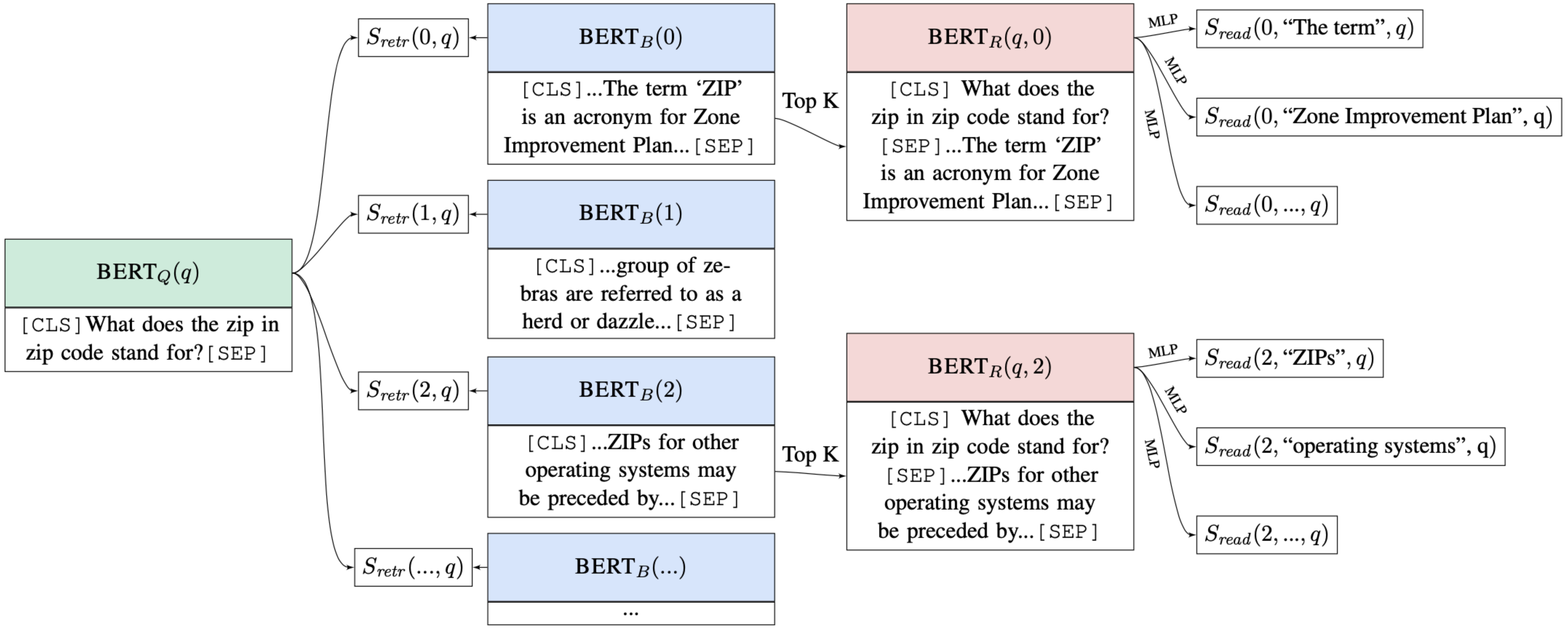


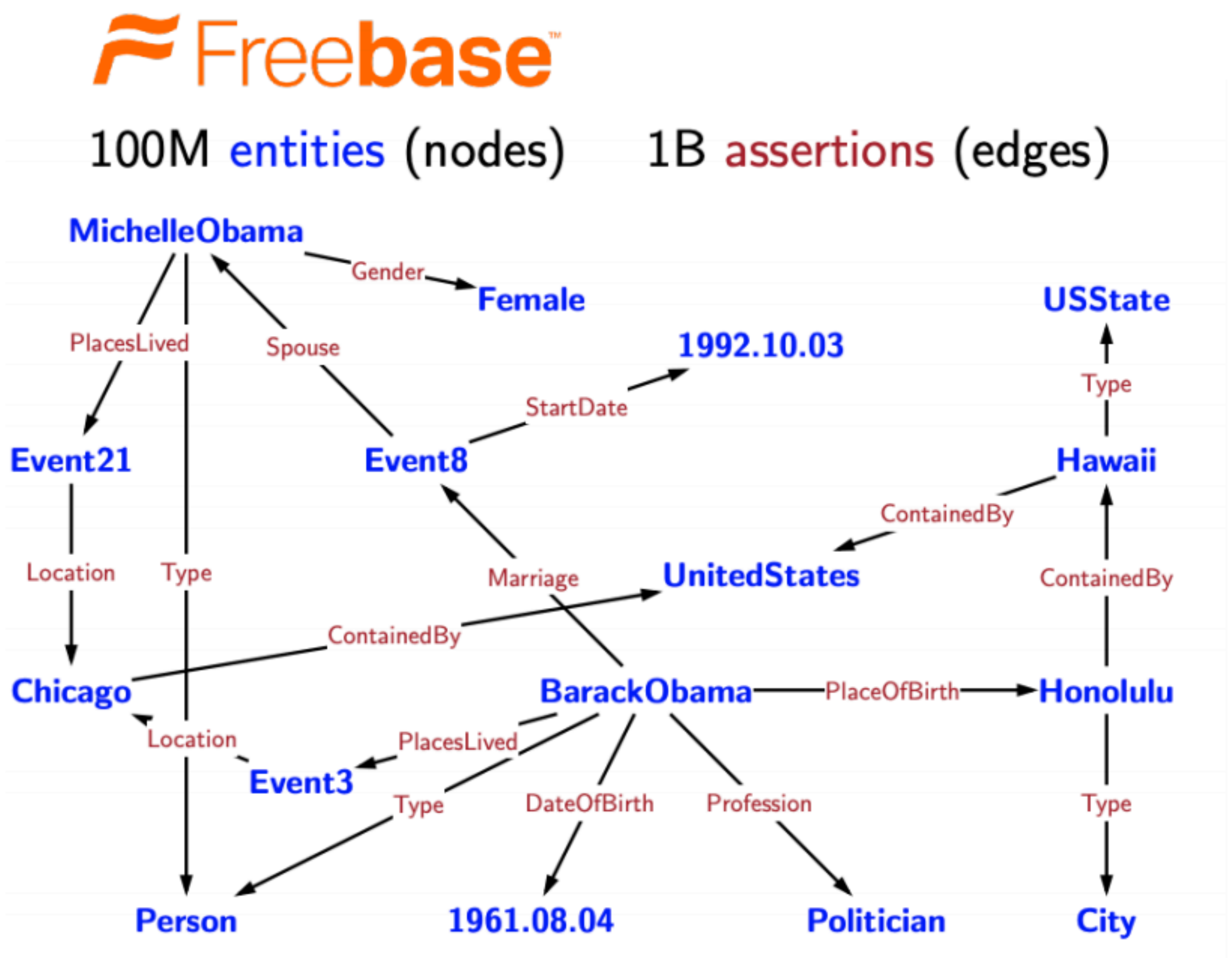
Image credit: (Lee et al., 2019)

Almost all the state-of-the-art question answering systems are built on top of end-to-end training and pre-trained language models (e.g., BERT)!

Beyond textual QA problems

Today, we will mostly focus on how to answer questions based on **unstructured text**.

Knowledge based QA



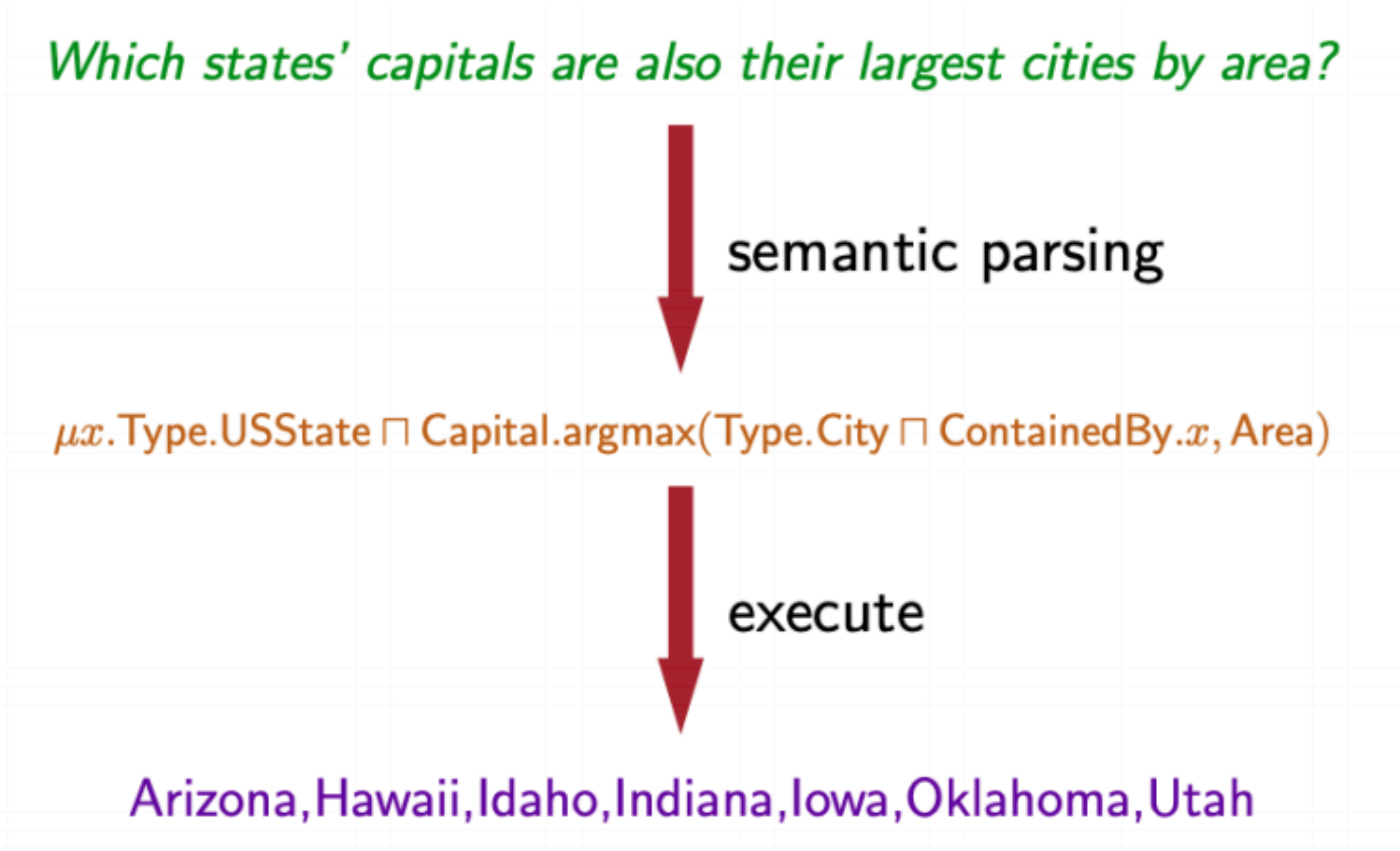


Image credit: Percy Liang

Beyond textual QA problems

Today, we will mostly focus on how to answer questions based on **unstructured text**.

Visual QA

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

(Antol et al., 2015): Visual Question Answering

2. Reading comprehension

Reading comprehension = comprehend a passage of text and answer questions about its content $(P, Q) \longrightarrow A$

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

A: German

2. Reading comprehension

Reading comprehension: building systems to comprehend a passage of text and answer questions about its content (P, Q) $\longrightarrow$ A

Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

A: Hindi

Why do we care about this problem?

- Useful for many practical applications
- Reading comprehension is an important testbed for evaluating how well computer systems understand human language
    - Wendy Lehnert 1977: “Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**.”
- Many other NLP tasks can be reduced to a reading comprehension problem:

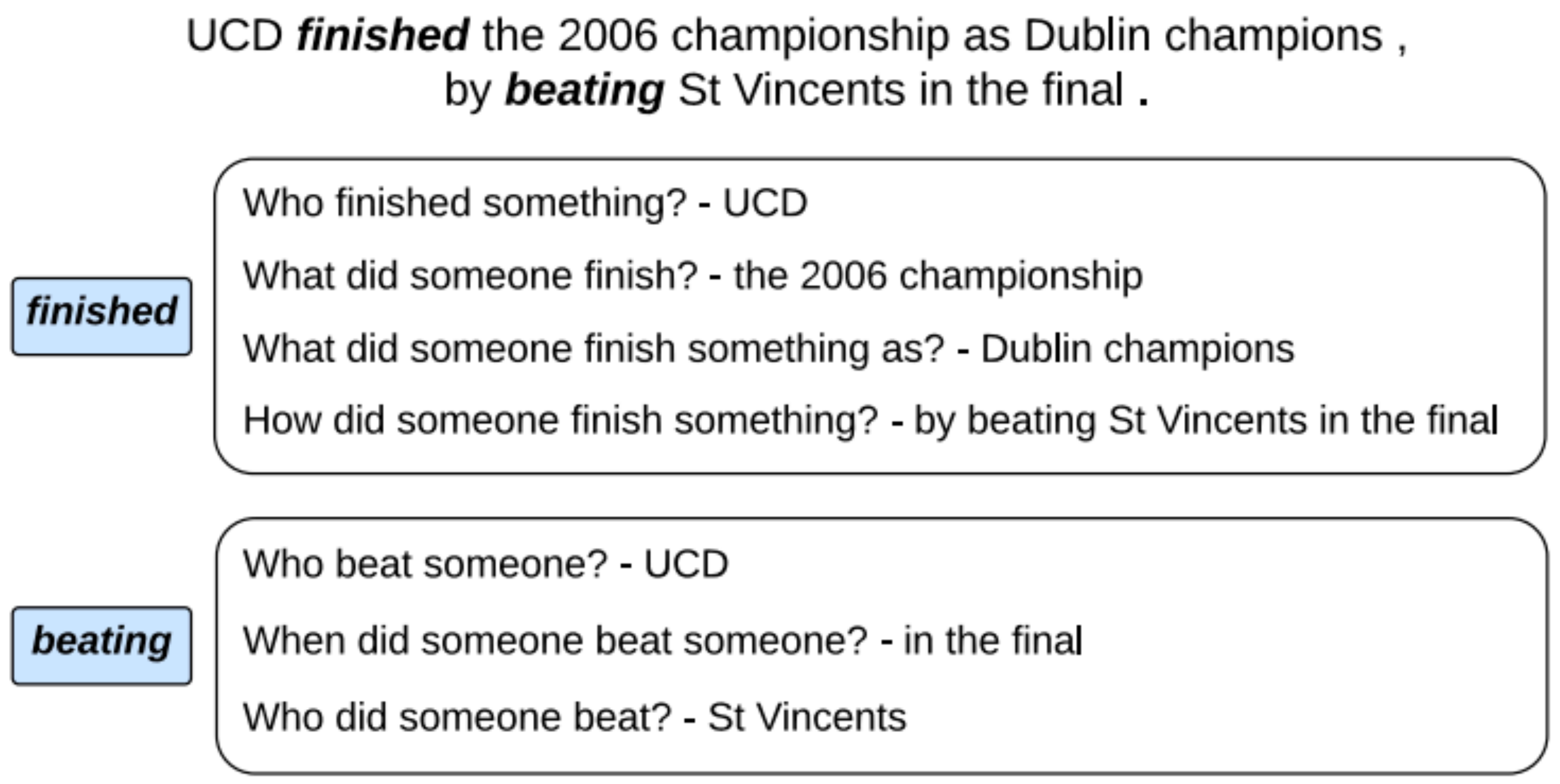
Information extraction

(Barack Obama, educated_at, ?)

> Question: Where did Barack Obama graduate from?
>
> Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.

(Levy et al., 2017)

Semantic role labeling



(He et al., 2015)

Stanford question answering dataset (SQuAD)

- 100k annotated (passage, question, answer) triples

  Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!

- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.

  This is a limitation— not all the questions can be answered in this way!

- SQuAD still remains the most popular reading comprehension dataset; it is "almost solved" today and the state-of-the-art exceeds the estimated 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

(Rajpurkar et al., 2016): SQuAD: 100,000+ Questions for Machine Comprehension

Stanford question answering dataset (SQuAD)

- **Evaluation**: exact match (0 or 1) and F1 (partial credit).
- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers.
- We compare the predicted answer to *each* gold answer (a, an, the, punctuations are removed) and take max scores. Finally, we take the average of all the examples for both exact match and F1.
- 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 served}

Exact match: $\max\{0, 0, 0\} = 0$

F1: $\max\{0.67, 0.67, 0.61\} = 0.67$

Other question answering datasets

- TriviaQA: Questions and answers by trivia enthusiasts. Independently collected web paragraphs that contain the answer and seem to discuss question, but no human verification that paragraph supports answer to question

- Natural Questions: Question drawn from frequently asked Google search questions. Answers from Wikipedia paragraphs. Answer can be substring, yes, no, or NOT_PRESENT. Verified by human annotation.

- HotpotQA. Constructed questions to be answered from the whole of Wikipedia which involve getting information from two pages to answer a multistep query:
  Q: Which novel by the author of "Armada" will be adapted as a feature film by Steven Spielberg?   A: *Ready Player One*

Neural models for reading comprehension

How can we build a model to solve SQuAD?

(We are going to use **passage, paragraph and context,** as well as **question** and **query** interchangeably)

- Problem formulation
  - Input: $C = (c_1, c_2, \ldots, c_N)$, $Q = (q_1, q_2, \ldots, q_M)$, $c_i, q_i \in V$
  - Output: $1 \leq \text{start} \leq \text{end} \leq N$

N~100, M ~15

answer is a span in the passage

- A family of LSTM-based models with attention (2016–2018)

  Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), BiDAF (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..

- Fine-tuning BERT-like models for reading comprehension (2019+)

LSTM-based vs BERT models

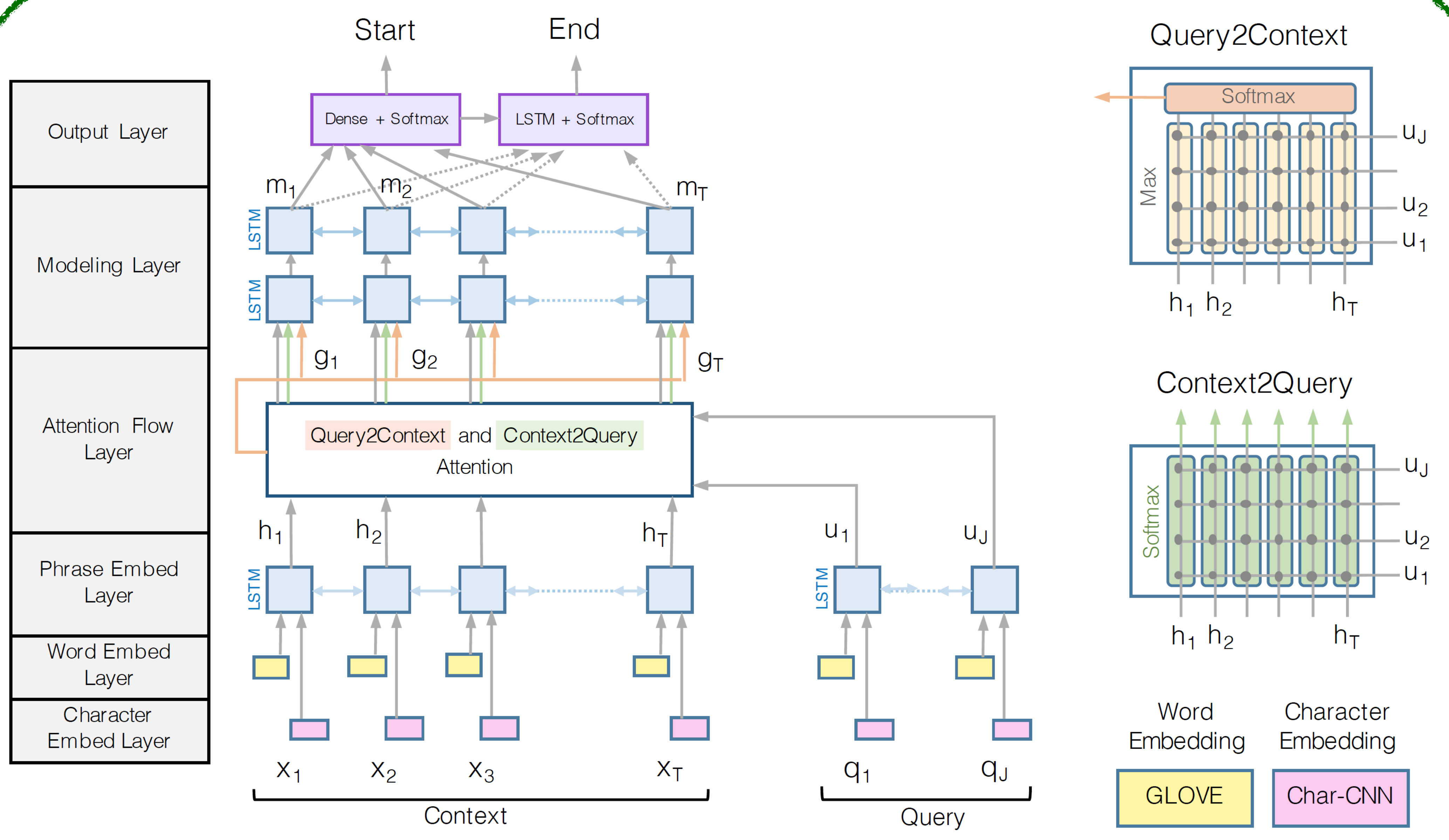


Image credit: (Seo et al, 2017)

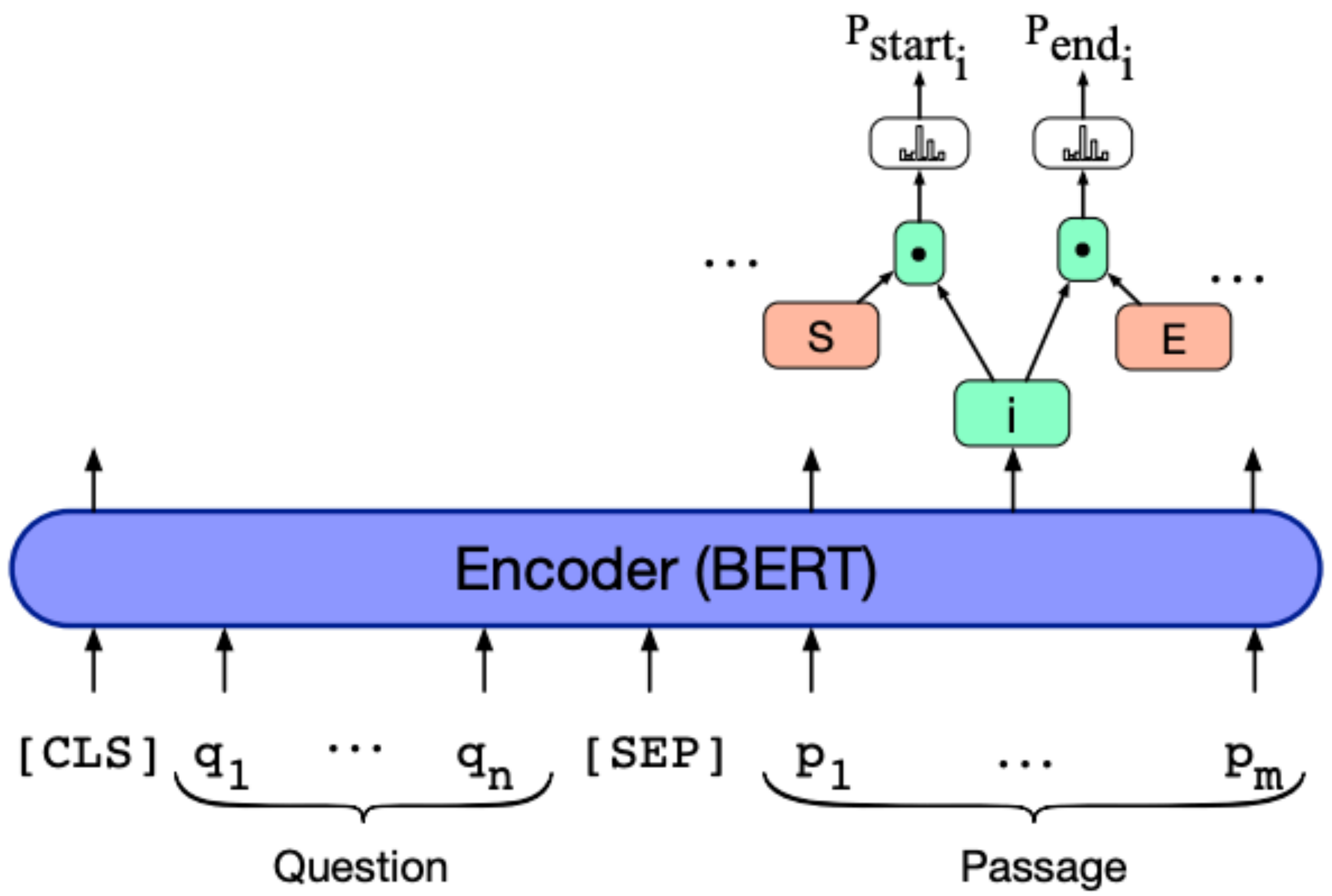


Image credit: J & M, edition 3

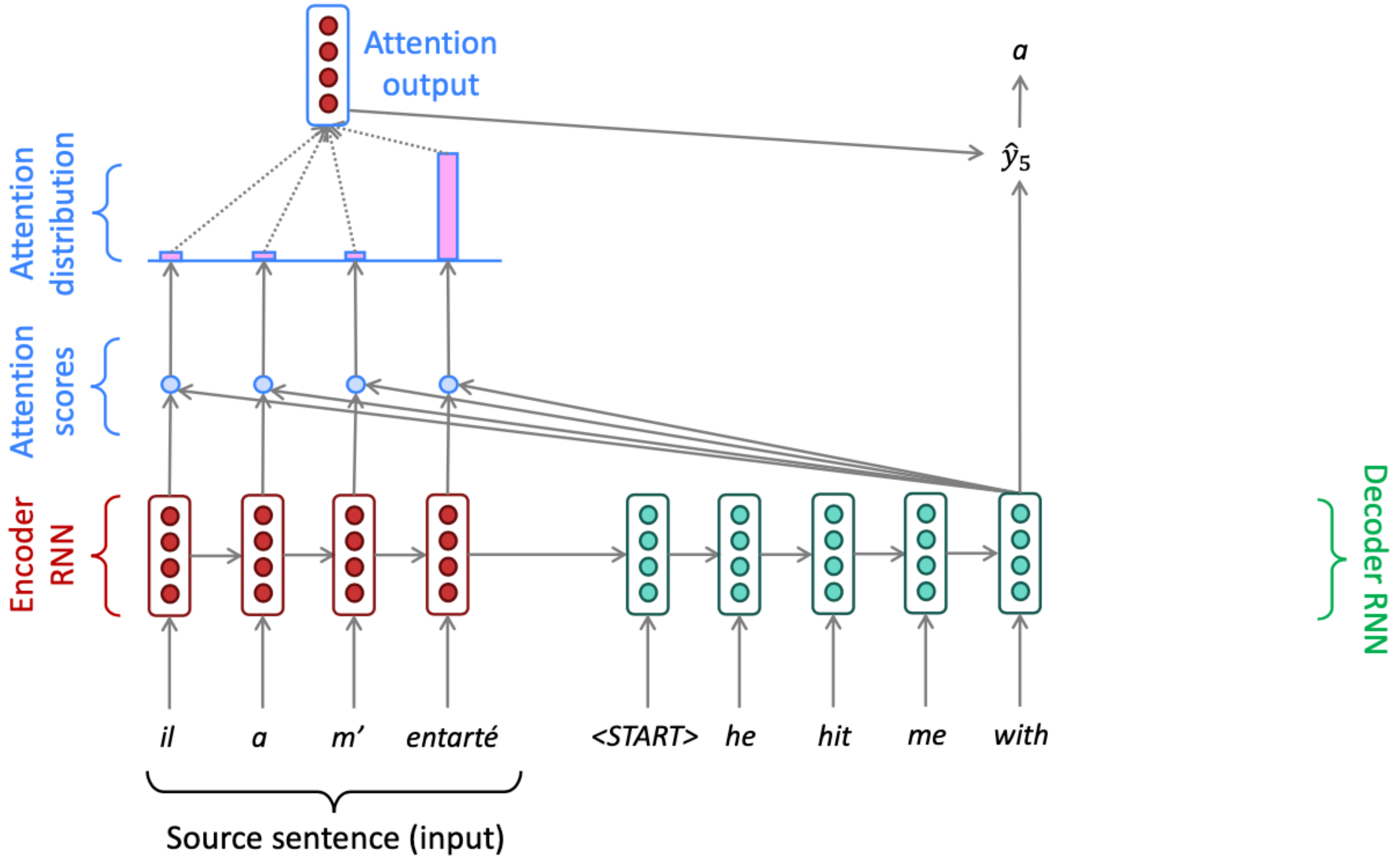
Recap: seq2seq model with attention

- Instead of source and target sentences, we also have two sequences: passage and question (lengths are unbalanced)

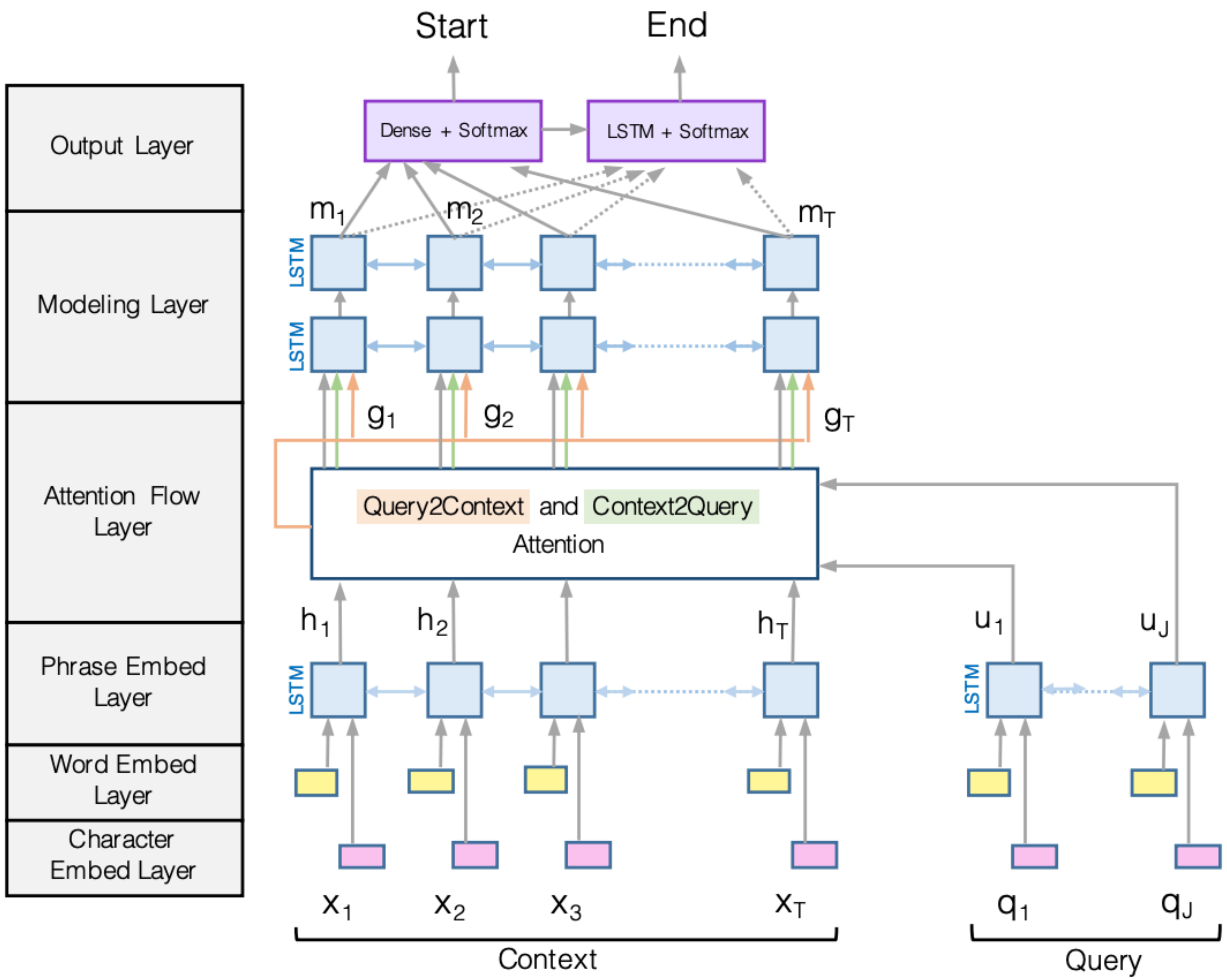
- We need to model which words in the passage are most relevant to the question (and which question words)

  Attention is the key ingredient here, similar to which words in the source sentence are most relevant to the current target word…

- We don't need an autoregressive decoder to generate the target sentence word-by-word. Instead, we just need to train two classifiers to predict the start and end positions of the answer!

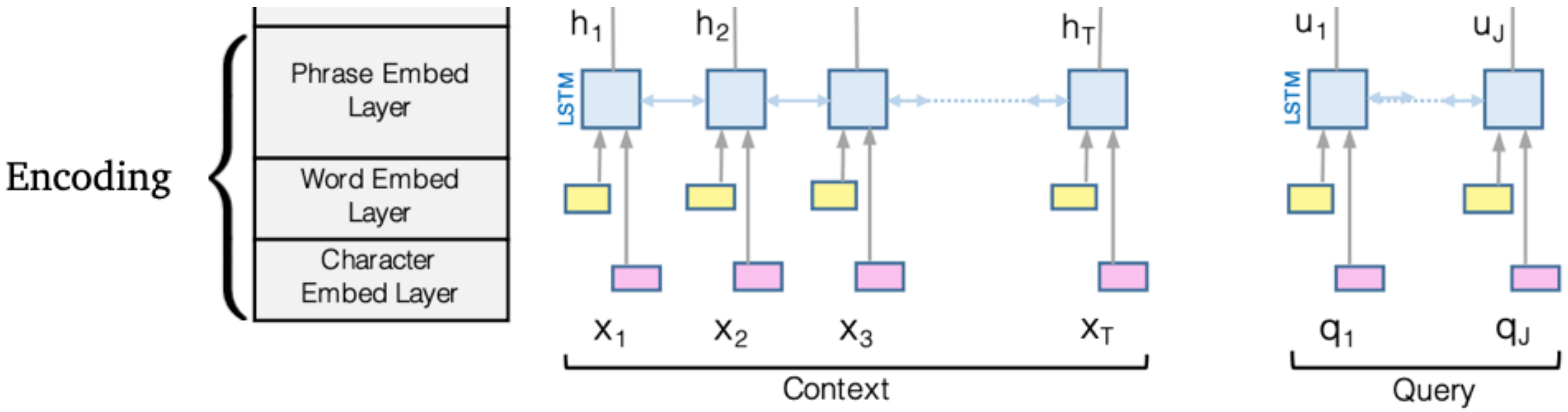


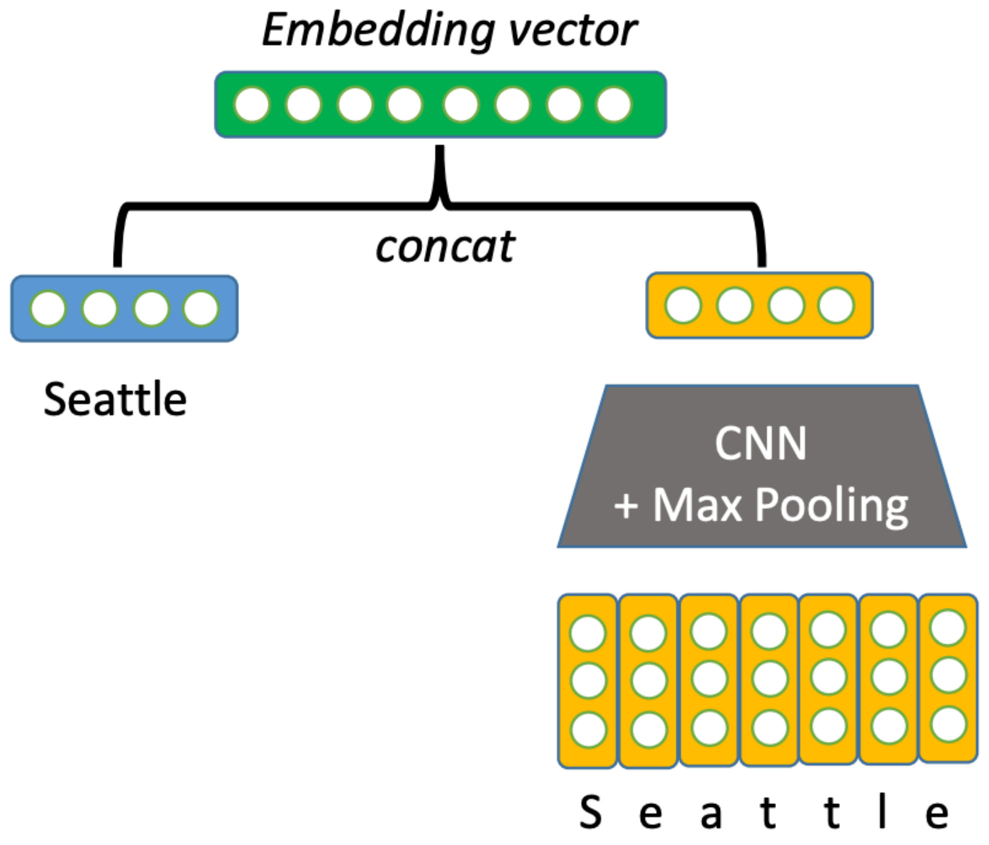
BiDAF: the Bidirectional Attention Flow model



(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

BiDAF: Encoding





- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query.

$$e(c_i) = f([\text{GloVe}(c_i); \text{charEmb}(c_i)]) \qquad e(q_i) = f([\text{GloVe}(q_i); \text{charEmb}(q_i)])$$

f: high-way networks omitted here

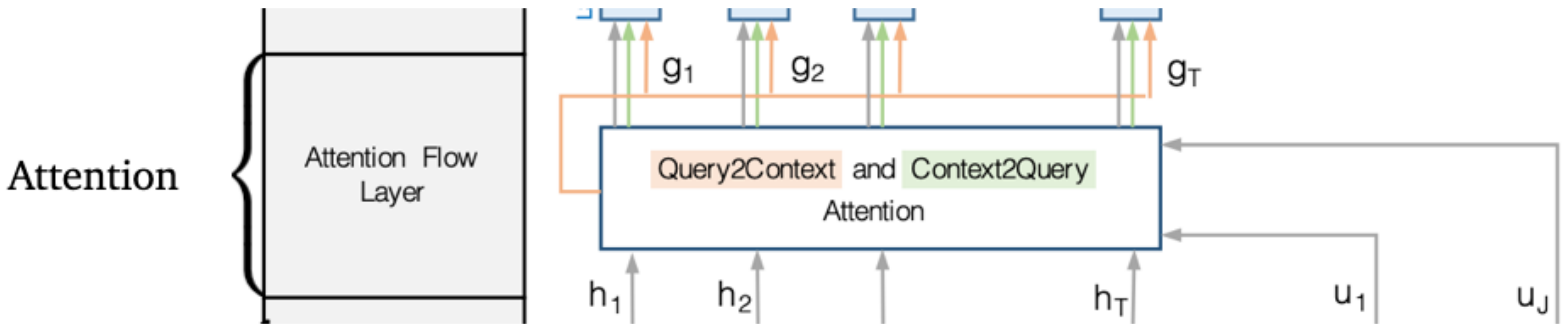
- Then, use two **bidirectional** LSTMs separately to produce contextual embeddings for both context and query.

$$\overrightarrow{\mathbf{c}}_i = \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H \qquad \overrightarrow{\mathbf{q}}_i = \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$

$$\overleftarrow{\mathbf{c}}_i = \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H \qquad \overleftarrow{\mathbf{q}}_i = \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$

$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H} \qquad \mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

BiDAF: Attention



- Context-to-query attention: For each context word, choose the most relevant words from the query words.

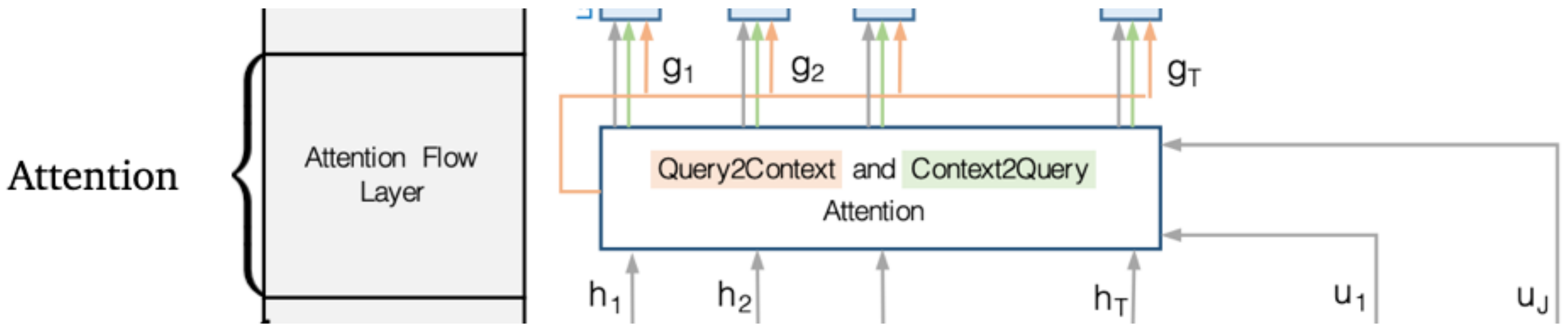
Q: *Who leads the United States?*

C: *Barak Obama is the president of the USA.*

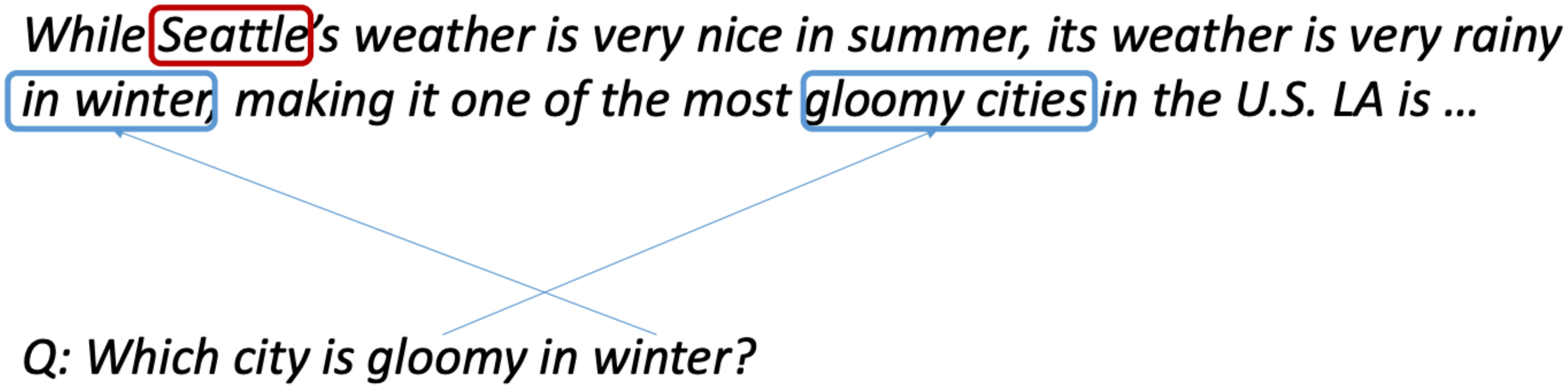
For each context word, find the most relevant query word.

(Slides adapted from Minjoon Seo)

BiDAF: Attention

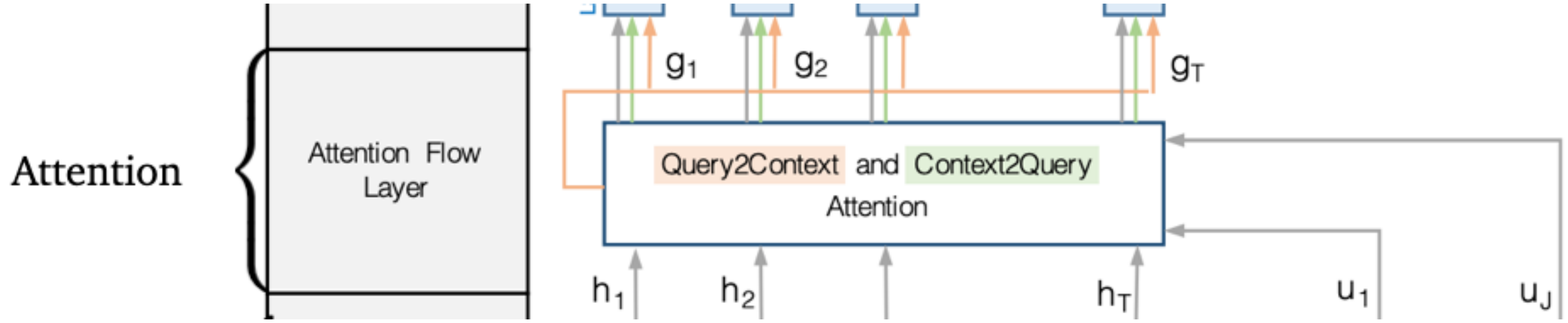


- Query-to-context attention: choose the context words that are most relevant to one of query words.



(Slides adapted from Minjoon Seo)

BiDAF: Attention



The final output is $\mathbf{g}_i = [\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}] \in \mathbb{R}^{8H}$

- First, compute a similarity score for every pair of $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{i,j} = \mathbf{w}_{\text{sim}}^{\intercal}[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j] \in \mathbb{R} \qquad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6H}$$

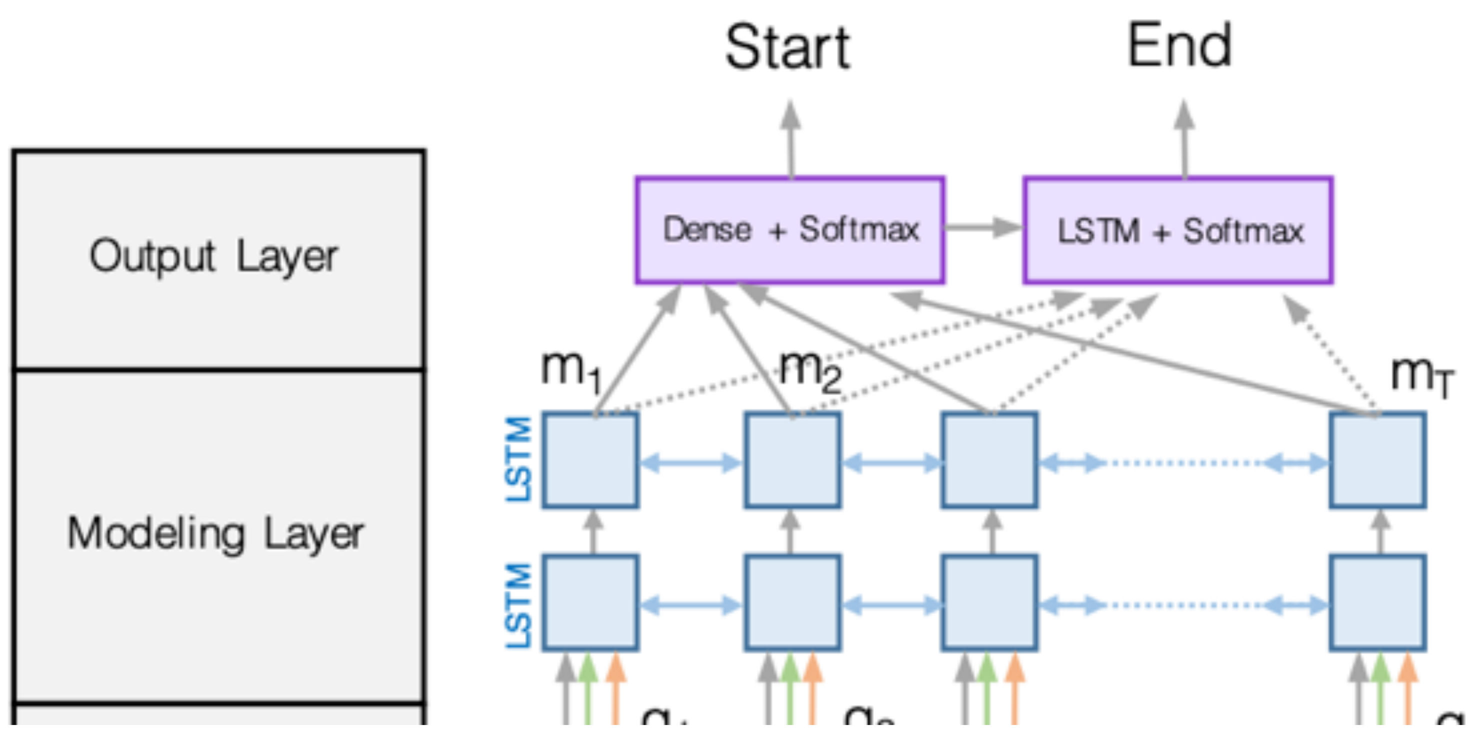
- Context-to-query attention (which question words are more relevant to c_i):

$$\alpha_{i,j} = \text{softmax}_j(S_{i,j}) \in \mathbb{R} \qquad \mathbf{a}_i = \sum_{j=1}^{M} \alpha_{i,j}\mathbf{q}_j \in \mathbb{R}^{2H}$$

- Query-to-context attention (which context words are relevant to some question words):

$$\beta_i = \text{softmax}_i(\max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N \qquad \mathbf{b} = \sum_{i=1}^{N} \beta_i\mathbf{c}_i \in \mathbb{R}^{2H}$$

BiDAF: Modeling and output layers



The final training loss is

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

Modeling layer: pass $\mathbf{g}_i$ to another two layers of **bi-directional** LSTMs.

- Attention layer is modeling interactions between query and context
- Modeling layer is modeling interactions within context words

$$\mathbf{m}_i = \text{BiLSTM}(\mathbf{g}_\text{i}) \in \mathbb{R}^{2H}$$

Output layer: two classifiers predicting the start and end positions:

$$p_{\text{start}} = \text{softmax}(\mathbf{w}_{\text{start}}^{\intercal}[\mathbf{g}_i; \mathbf{m}_i]) \qquad p_{\text{end}} = \text{softmax}(\mathbf{w}_{\text{end}}^{\intercal}[\mathbf{g}_i; \mathbf{m}'_i])$$

$$\mathbf{m}'_i = \text{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H} \quad \mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$$

BiDAF: Performance on SQuAD

This model achieved 77.3 F1 on SQuAD v1.1.

- Without context-to-query attention $\Longrightarrow$ 67.7 F1
- Without query-to-context attention $\Longrightarrow$ 73.7 F1
- Without character embeddings $\Longrightarrow$ 75.4 F1

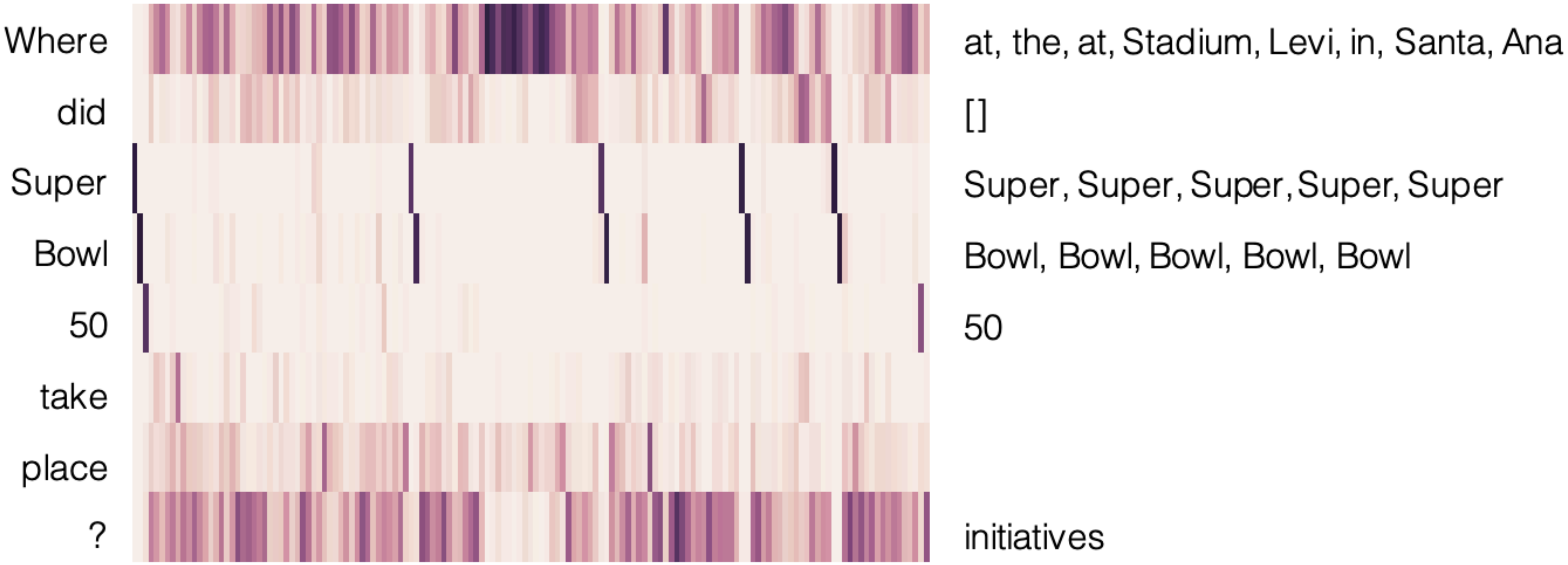
	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Human performance	91.2

64

(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

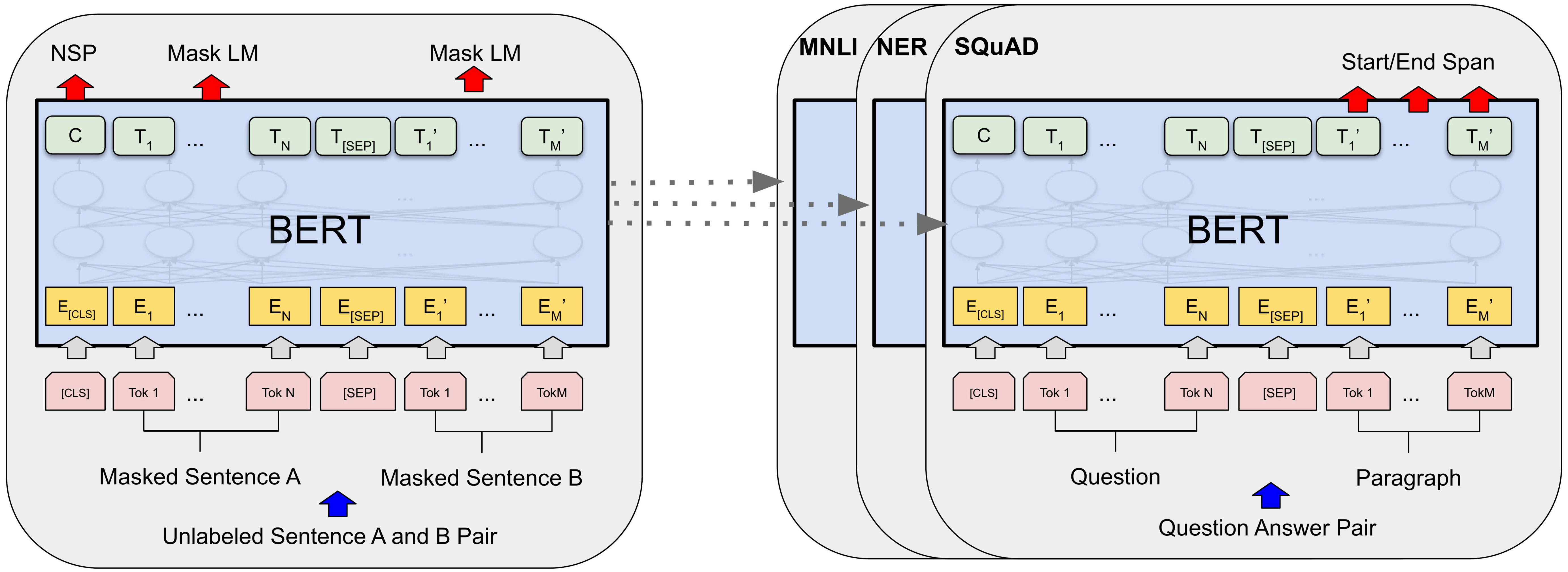
Attention visualization

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 .



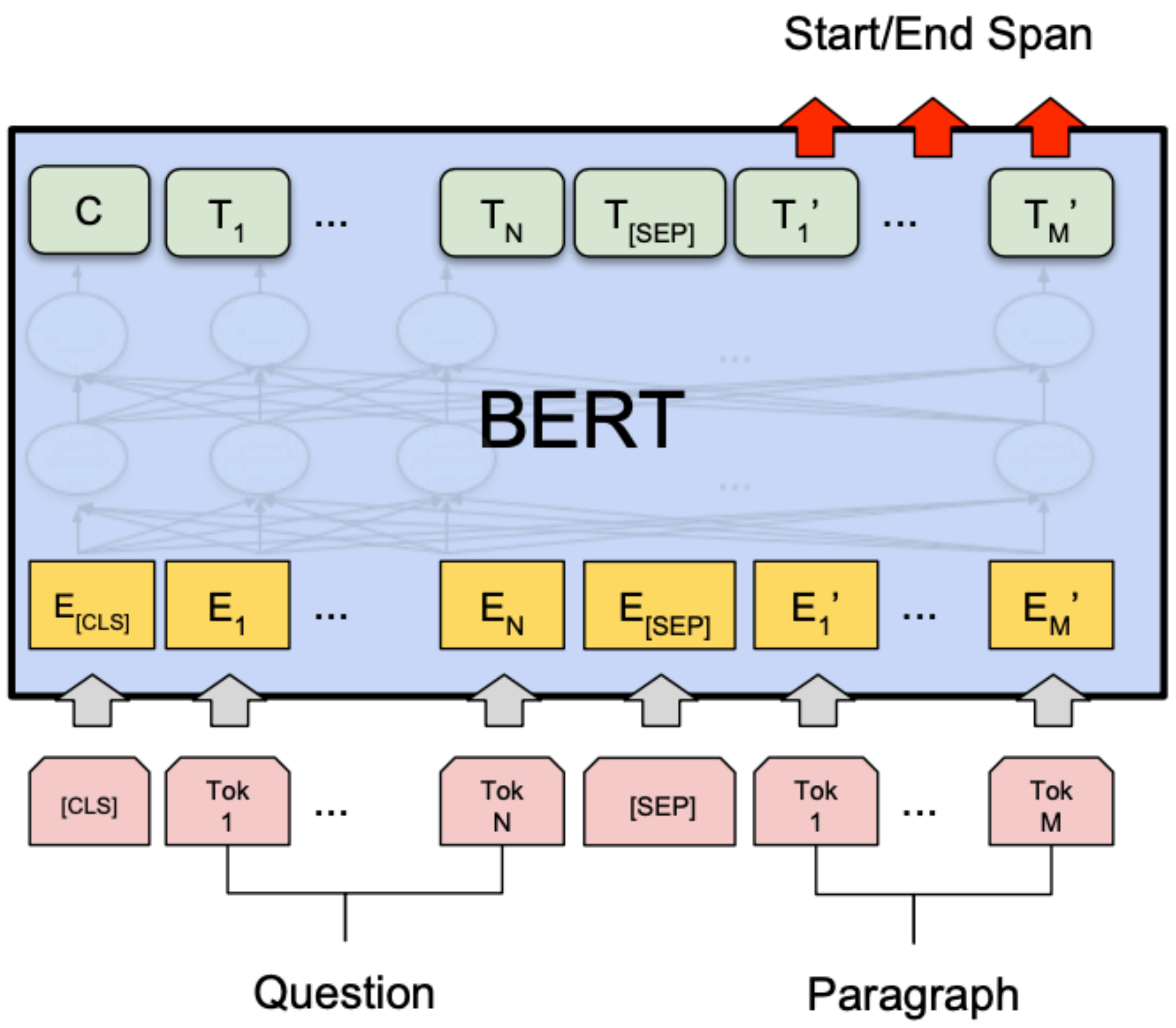
BERT for reading comprehension

- BERT is a deep bidirectional Transformer encoder pre-trained on large amounts of text (Wikipedia + BooksCorpus)
- BERT is pre-trained on two training objectives:
    - Masked language model (MLM)
    - Next sentence prediction (NSP)
- $\text{BERT}_{\text{base}}$ has 12 layers and 110M parameters, $\text{BERT}_{\text{large}}$ has 24 layers and 330M parameters



Pre-training Fine-Tuning

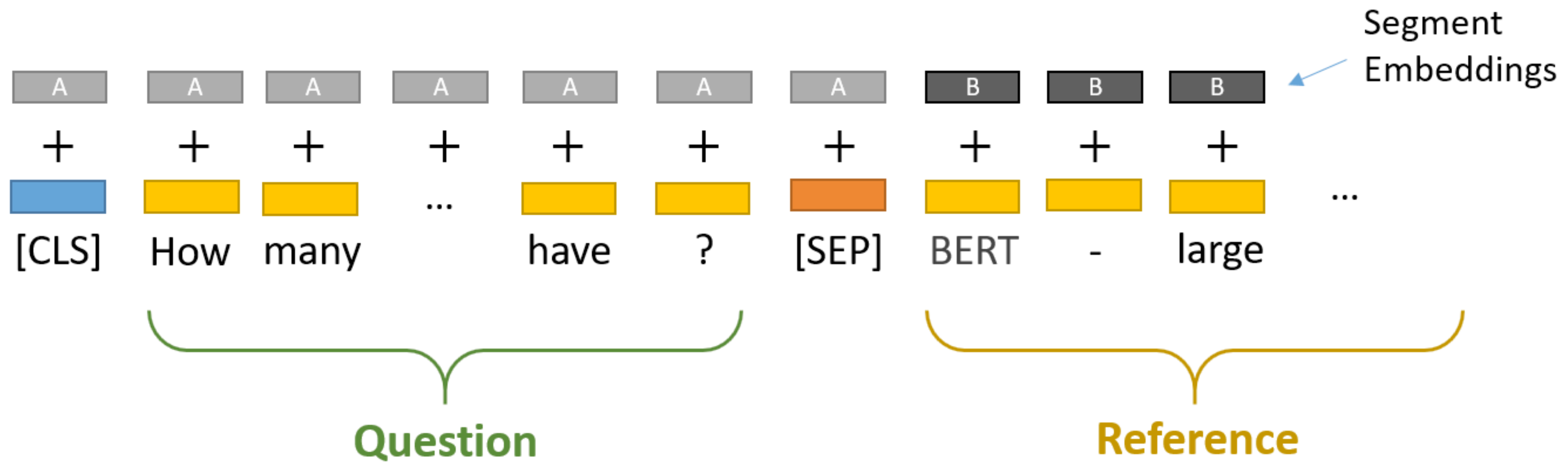
BERT for reading comprehension



Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B



Question: `How many parameters does BERT-large have?`

Reference Text: `BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.`

Image credit: https://mccormickml.com/

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\intercal}\mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\intercal}\mathbf{h}_i)$$

where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

BERT for reading comprehension

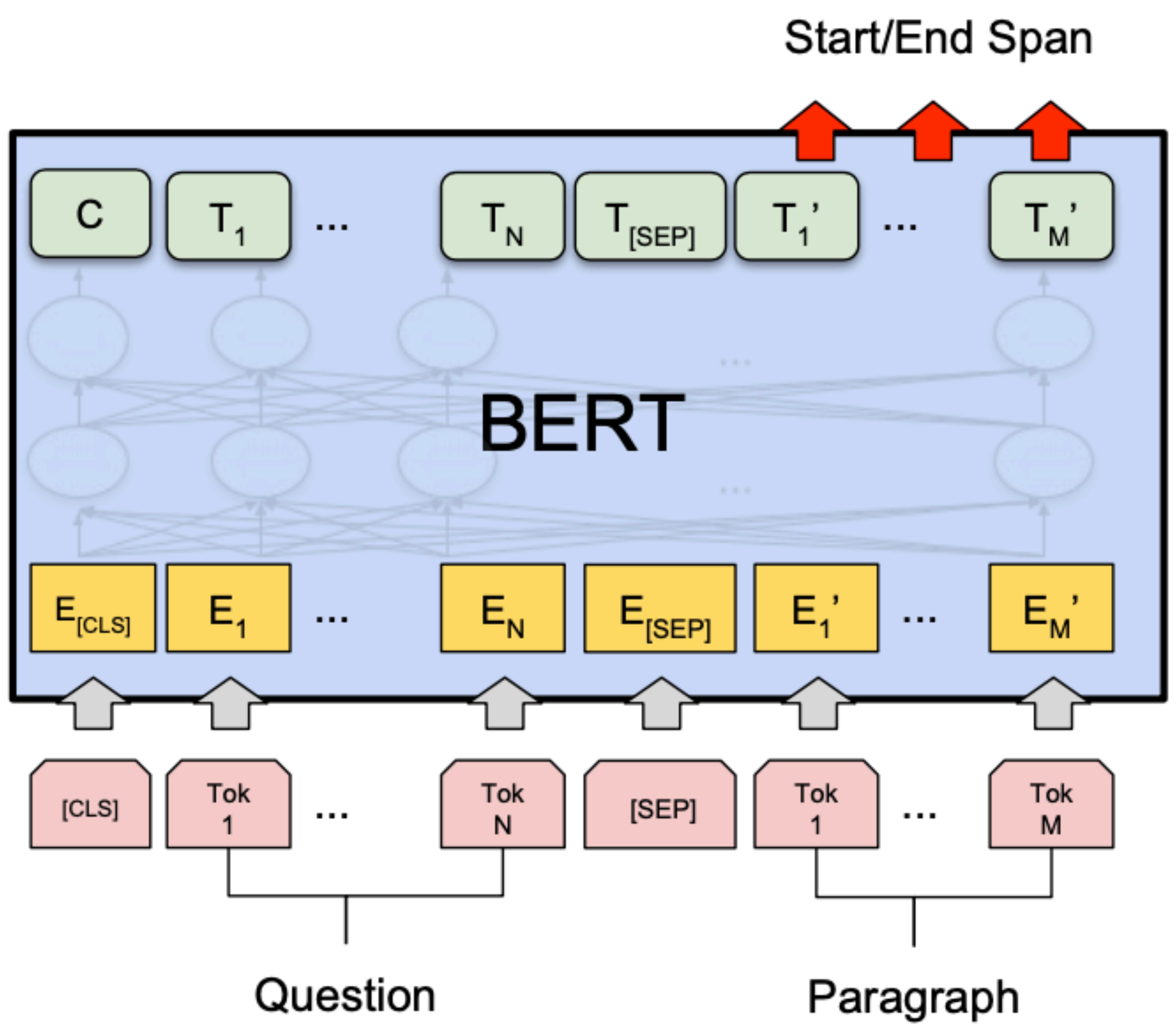
$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

- All the BERT parameters (e.g., 110M) as well as the newly introduced parameters $\mathbf{h}_{\text{start}}, \mathbf{h}_{\text{end}}$ (e.g., 768 x 2 = 1536) are optimized together for $\mathcal{L}$.

- It works amazing well. Stronger pre-trained language models can lead to even better performance and SQuAD becomes a standard dataset for testing pre-trained models.

	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
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

(dev set, except for human performance)



Comparisons between BiDAF and BERT models

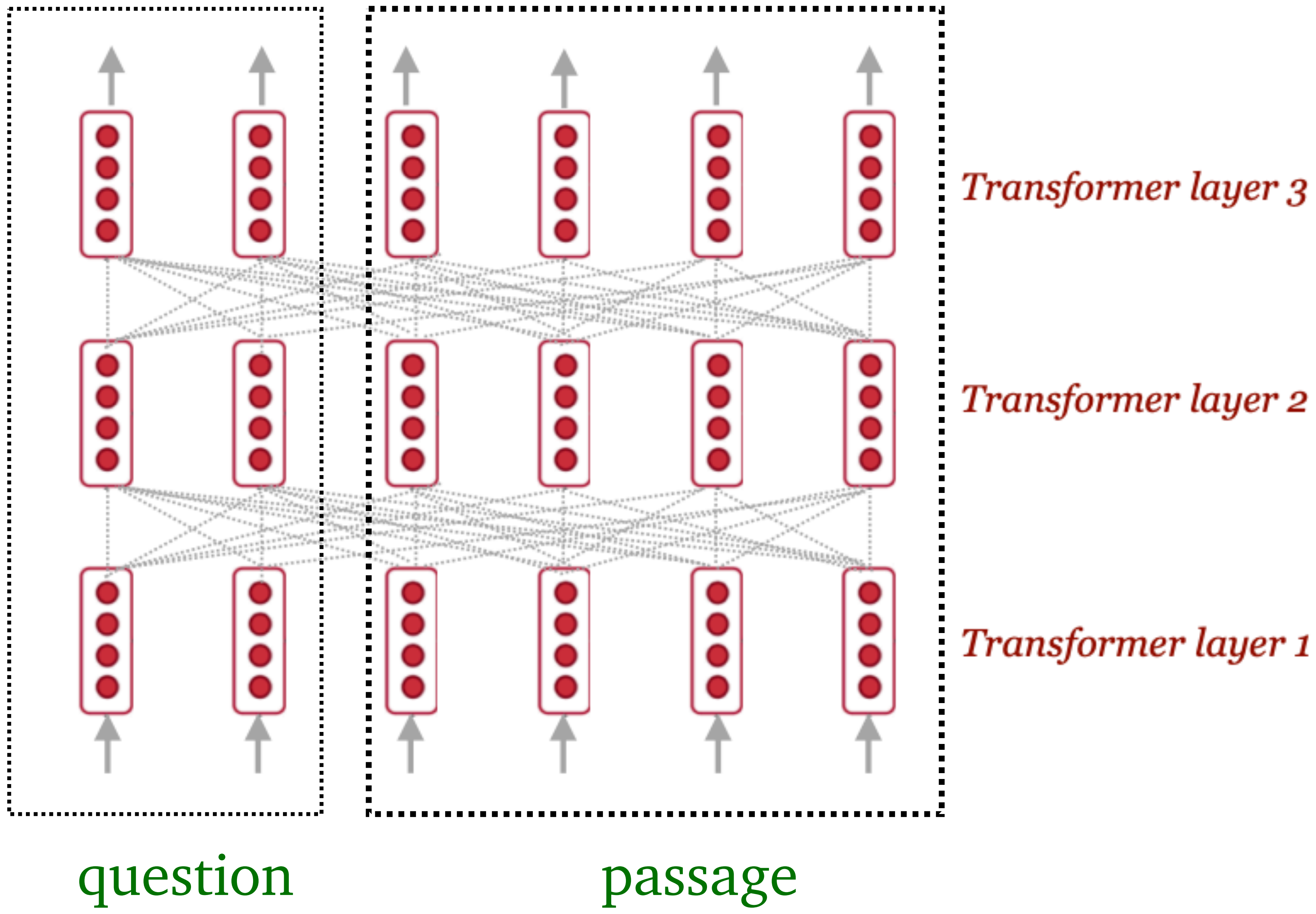
- BERT model has many many more parameters (110M or 330M)
  BiDAF has ~2.5M parameters.
- BiDAF is built on top of several bidirectional LSTMs while BERT is built on top of Transformers (no recurrence architecture and easier to parallelize).
- BERT is **pre-trained** while BiDAF is only built on top of GloVe (and all the remaining parameters need to be learned from the supervision datasets).

Pre-training is clearly a game changer but it is expensive..

Comparisons between BiDAF and BERT models

Are they really fundamentally different? Probably not.

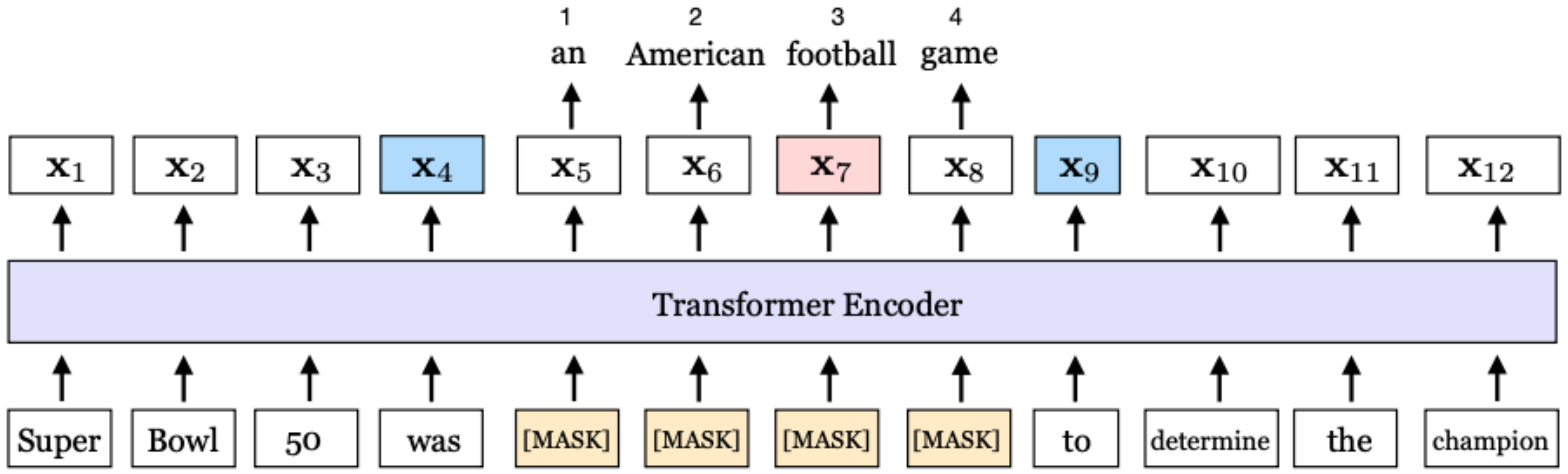
- BiDAF and other models aim to model the interactions between question and passage.
- BERT uses self-attention between the **concatenation** of question and passage = attention(P, P) + attention(P, Q) + attention(Q, P) + attention(Q, Q)
- (Clark and Gardner, 2018) shows that adding a self-attention layer for the passage attention(P, P) to BiDAF also improves performance.



Can we design better pre-training objectives?

The answer is yes!

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$



Two ideas:

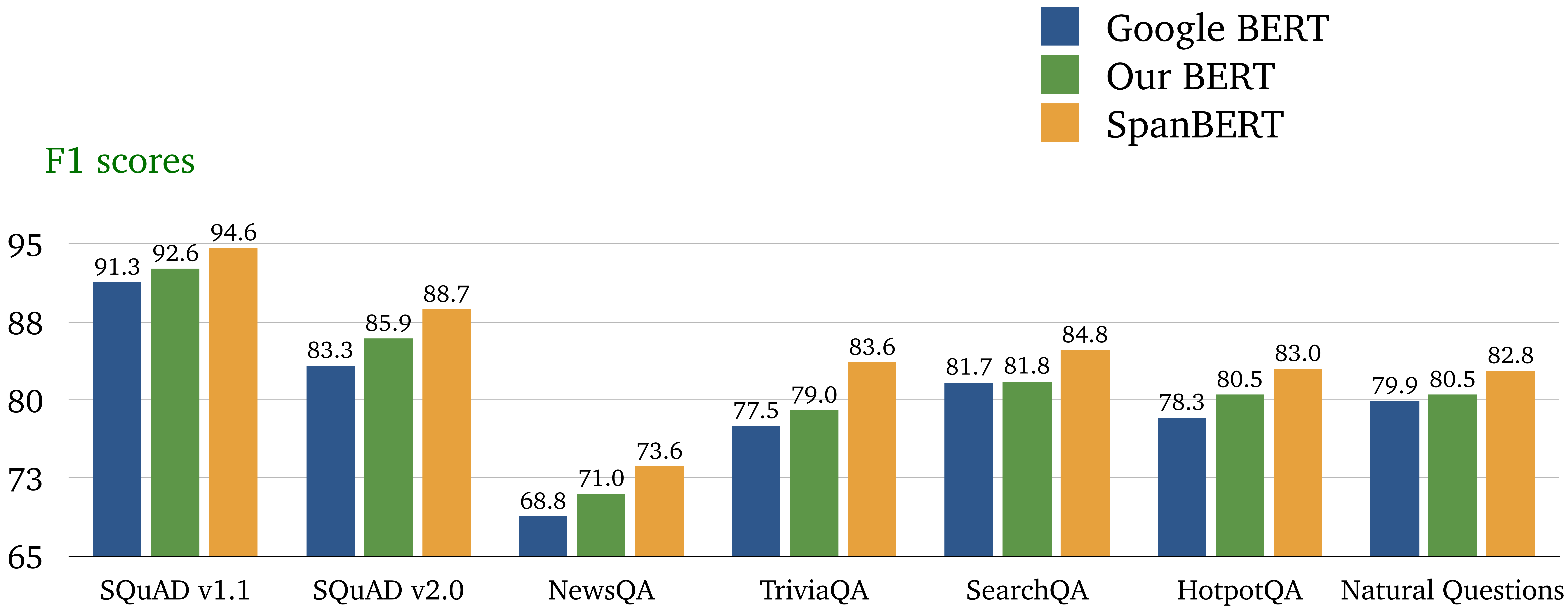
1) masking contiguous spans of words instead of 15% random words

2) using the two end points of span to predict all the masked words in between = compressing the information of a span into its two endpoints

$$\mathbf{y}_i = f(\mathbf{x}_{s-1}, \mathbf{x}_{e+1}, \mathbf{p}_{i-s+1})$$

(Joshi & Chen et al., 2020): SpanBERT: Improving Pre-training by Representing and Predicting Spans

SpanBERT performance



Is reading comprehension solved?

- We have already surpassed human performance on SQuAD. Does it mean that reading comprehension is already solved? **Of course not!**

- The current systems still perform poorly on adversarial examples or examples from out-of-domain distributions

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSENT	27.3	29.4	34.3	34.2
ADDONESENT	39.0	41.8	45.7	46.9
ADDANY	7.6	11.7	4.8	2.7
ADDCOMMON	38.9	51.0	41.7	52.6

(Jia and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

Is reading comprehension solved?

Systems trained on one dataset can’t generalize to other datasets:

Fine-tuned on		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

(Sen and Saffari, 2020): What do Models Learn from Question Answering Datasets?

Is reading comprehension solved?

BERT-large model trained on SQuAD

	Test *TYPE* and Description	Failure Rate (🤖)	Example Test cases (with expected behavior and 🤖 prediction)
Vocab	***MFT:*** comparisons	20.0	**C:** Victoria is younger than Dylan. **Q:** Who is less young? **A:** Dylan 🤖: Victoria
Vocab	***MFT:*** intensifiers to superlative: most/least	91.3	**C:** Anna is worried about the project. Matthew is extremely worried about the project. **Q:** Who is least worried about the project? **A:** Anna 🤖: Matthew
Taxonomy	***MFT:*** match properties to categories	82.4	**C:** There is a tiny purple box in the room. **Q:** What size is the box? **A:** tiny 🤖: purple
Taxonomy	***MFT:*** nationality vs job	49.4	**C:** Stephanie is an Indian accountant. **Q:** What is Stephanie's job? **A:** accountant 🤖: Indian accountant
Taxonomy	***MFT:*** animal vs vehicles	26.2	**C:** Jonathan bought a truck. Isabella bought a hamster. **Q:** Who bought an animal? **A:** Isabella 🤖: Jonathan
Taxonomy	***MFT:*** comparison to antonym	67.3	**C:** Jacob is shorter than Kimberly. **Q:** Who is taller? **A:** Kimberly 🤖: Jacob
Taxonomy	***MFT:*** more/less in context, more/less antonym in question	100.0	**C:** Jeremy is more optimistic than Taylor. **Q:** Who is more pessimistic? **A:** Taylor 🤖: Jeremy
Robust.	***INV:*** Swap adjacent characters in **Q** (typo)	11.6	**C:** ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... **Q:** What was the ideal duty → udty of a Newcomen engine? **A:** INV 🤖: 7 million → 5 million
Robust.	***INV:*** add irrelevant sentence to **C**	9.8	(no example)

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

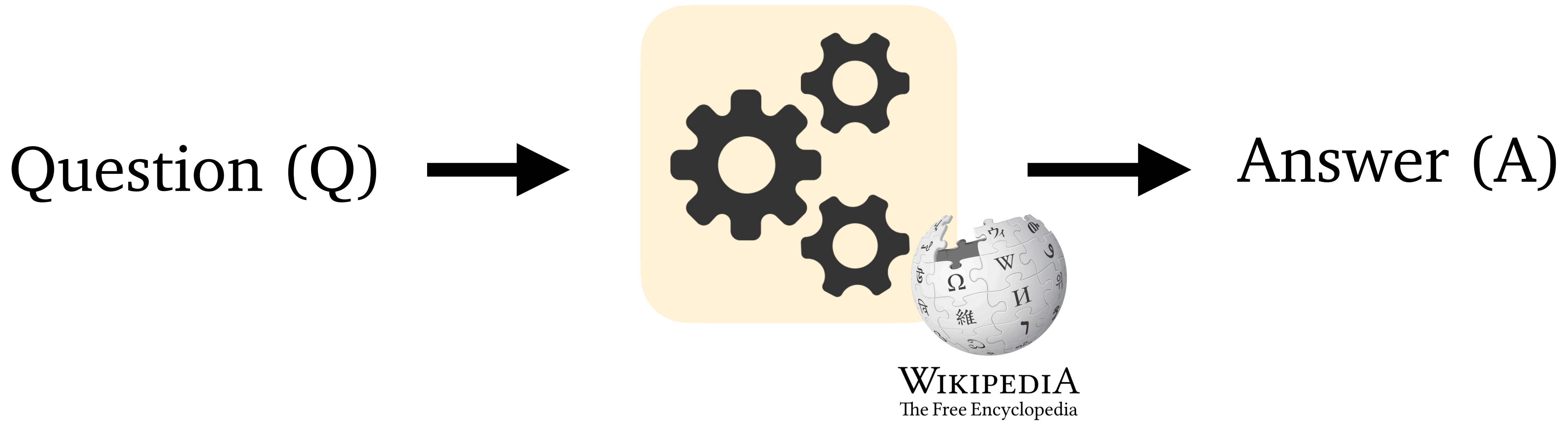
Is reading comprehension solved?

BERT-large model trained on SQuAD

Temporal	***MFT:*** change in one person only	41.5	**C:** Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail : Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle : Logan
Neg.	***MFT:*** Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca : Aaron
	MFT: **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron : Mark
Coref.	***MFT:*** Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio : Melissa
	MFT: Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria : Alex
	MFT: former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly : Jennifer
SRL	***MFT:*** subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth : Richard
	MFT: subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa : Jose

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

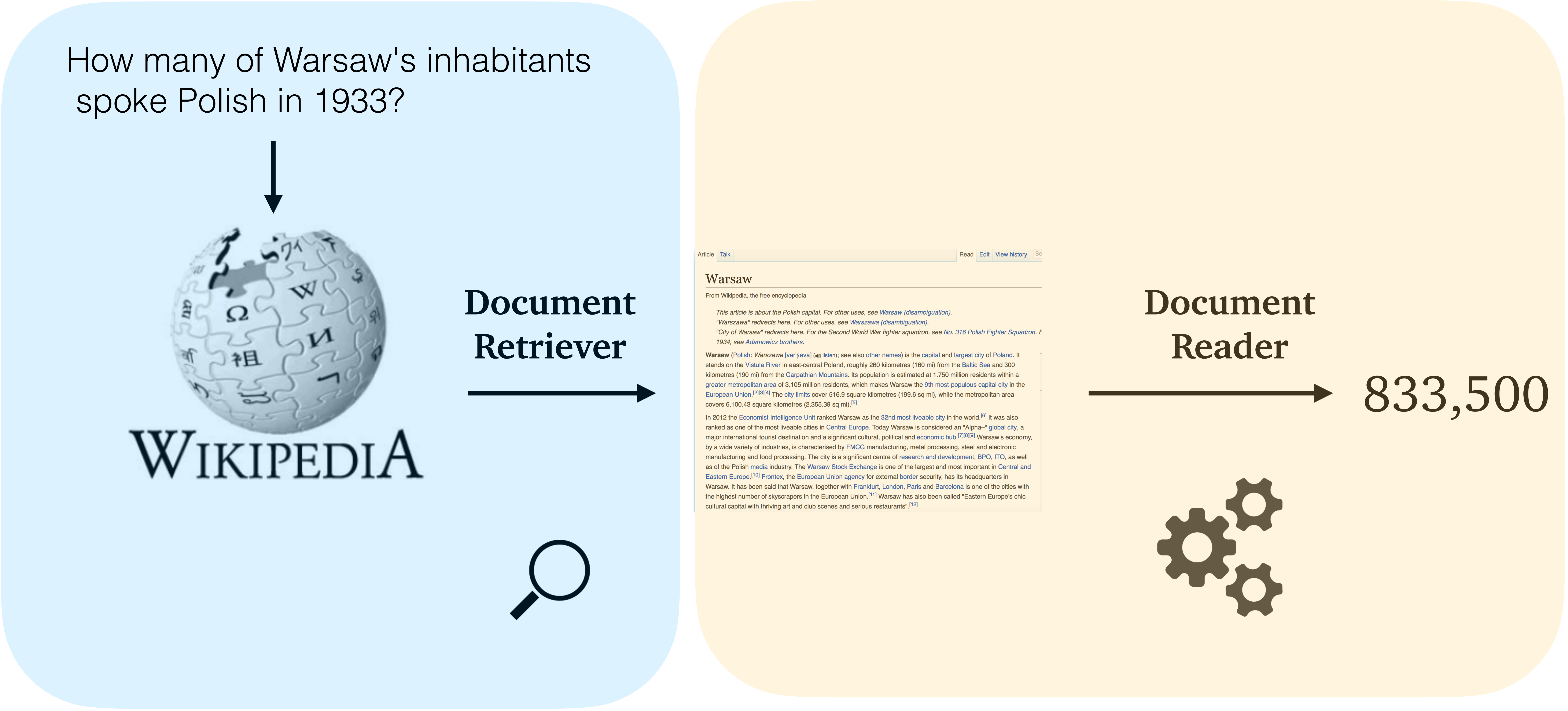
3. Open-domain question answering



- Different from reading comprehension, we don't assume a given passage.
- Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don't know where the answer is located, and the goal is to return the answer for any open-domain questions.
- Much more challenging and a more practical problem!

In contrast to ***closed-domain*** *systems that deal with questions under a specific domain (medicine, technical support).*

Retriever-reader framework



https://github.com/facebookresearch/DrQA

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Retriever-reader framework

- Input: a large collection of documents $\mathscr{D} = D_1, D_2, \ldots, D_N$ and Q
- Output: an answer string A

- Retriever: $f(\mathscr{D}, Q) \longrightarrow P_1, \ldots, P_K$ K is pre-defined (e.g., 100)
- Reader: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$ A reading comprehension problem!

In DrQA,

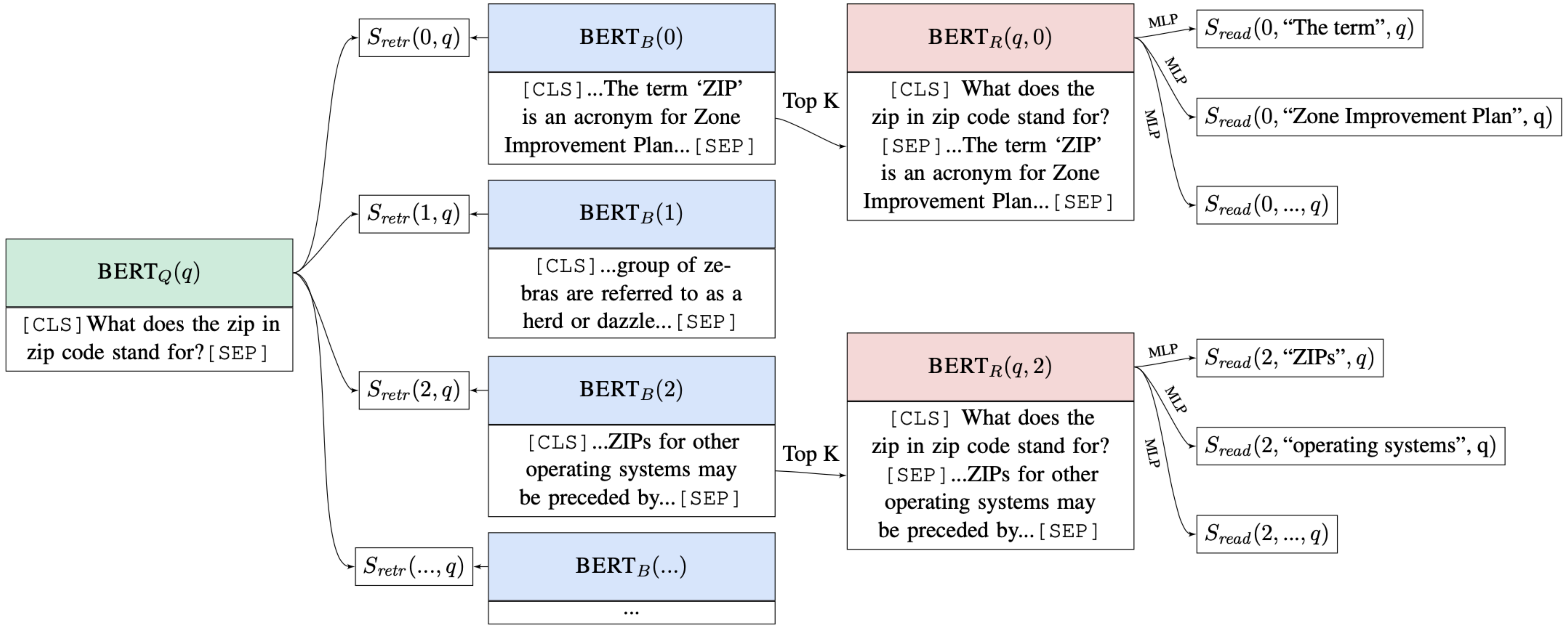
- Retriever = A standard TF-IDF information-retrieval sparse model (a fixed module)
- Reader = a neural reading comprehension model that we just learned
  - Trained on SQuAD and other distantly-supervised QA datasets

Distantly-supervised examples: (Q, A) $\longrightarrow$ (P, Q, A)

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

We can train the retriever too

- **Joint training** of retriever and reader



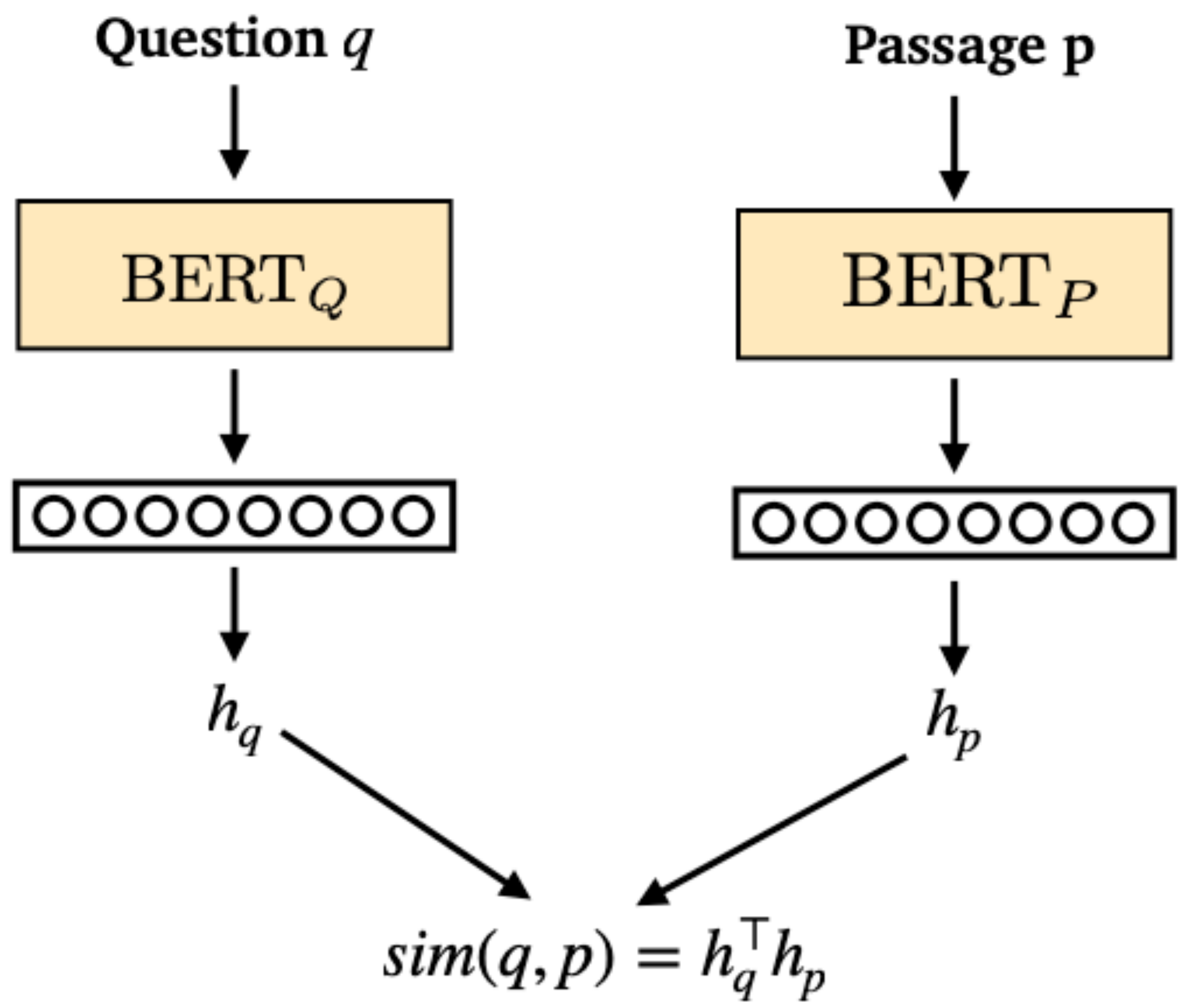
- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.

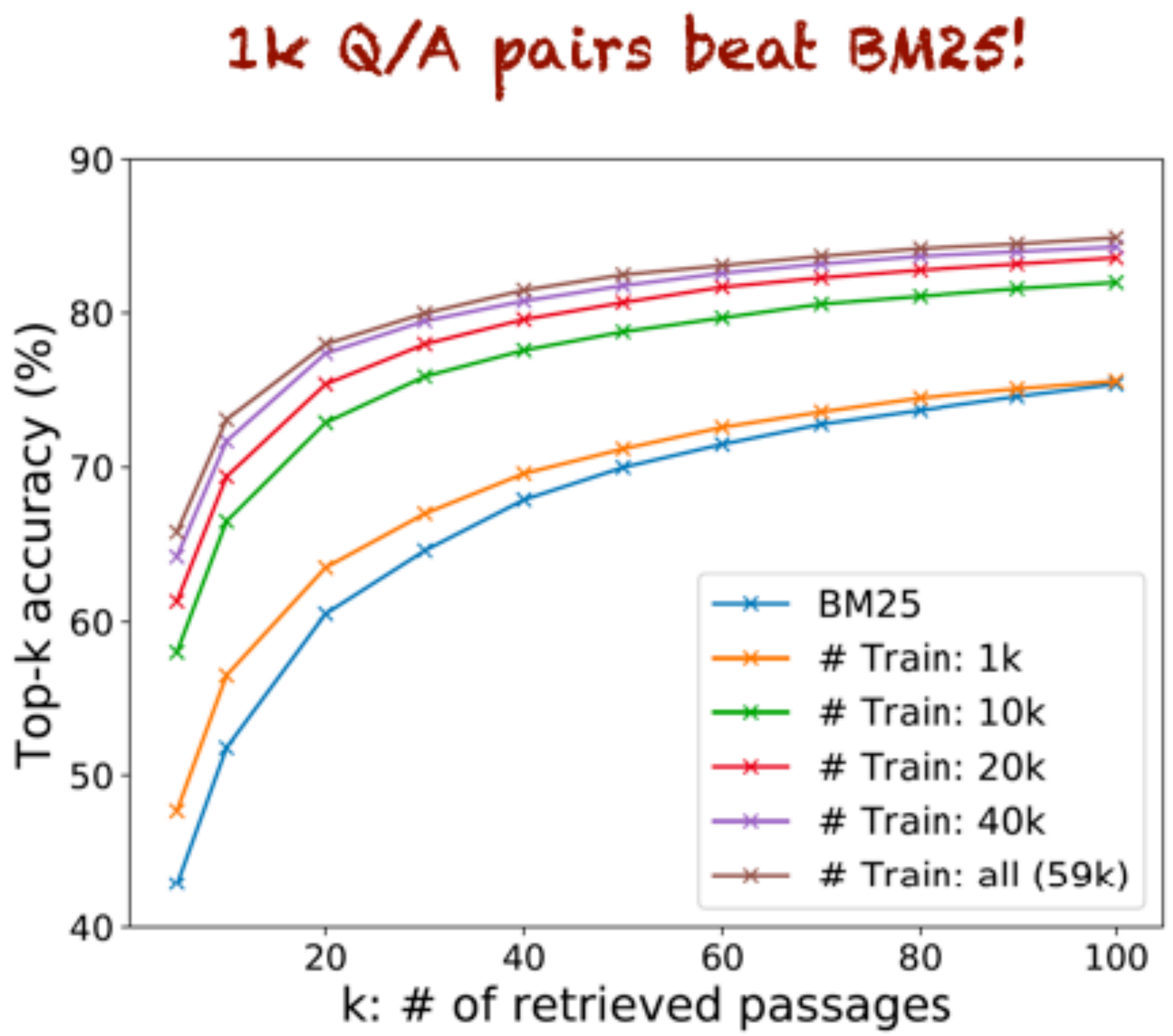
- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

We can train the retriever too

- Dense passage retrieval (DPR) - We can also just train the retriever using question-answer pairs!





- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

We can train the retriever too

Who tells harry potter that he is a wizard in the harry potter series? | Run

Title: *Harry Potter (film series)* Retrieval ranking: #90 $P(p|q)$=0.85 $P(a|p,q)$=1.00 $P(a,p|q)$=0.84

... and uncle. At the age of eleven, half-giant **Rubeus Hagrid** informs him that he is actually a wizard and that his parents were murdered by an evil wizard named Lord Voldemort. Voldemort also attempted to kill one-year-old Harry on the same night, but his killing curse mysteriously rebounded and reduced him to a weak and helpless form. Harry became extremely famous in the Wizarding World as a result. Harry begins his first year at Hogwarts School of Witchcraft and Wizardry and learns about magic. During the year, Harry and his friends Ron Weasley and Hermione Granger become entangled in the ...

Title: *Harry Potter (character)* Retrieval ranking: #1 $P(p|q)$=0.04 $P(a|p,q)$=0.97 $P(a,p|q)$=0.04

... Harry Potter (character) Harry James Potter is the titular protagonist of J. K. Rowling's "Harry Potter" series. The majority of the books' plot covers seven years in the life of the orphan Potter, who, on his eleventh birthday, learns he is a wizard. Thus, he attends Hogwarts School of Witchcraft and Wizardry to practice magic under the guidance of the kindly headmaster Albus Dumbledore and other school professors along with his best friends Ron Weasley and **Hermione Granger**. Harry also discovers that he is already famous throughout the novel's magical community, and that his fate is tied with that of ...

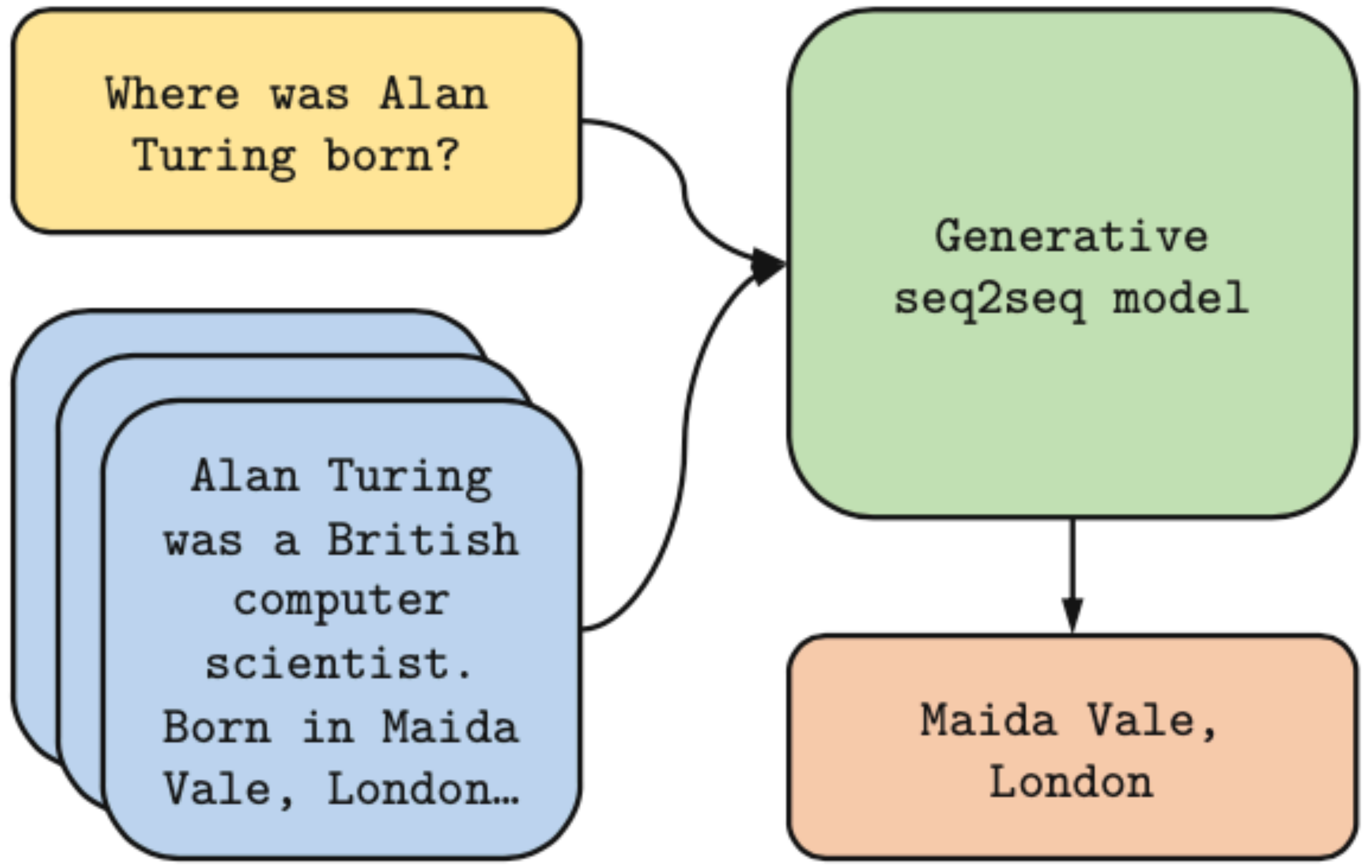
http://qa.cs.washington.edu:2020/

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Dense retrieval + generative models

Recent work shows that it is beneficial to generate answers instead of to extract answers.

Fusion-in-decoder (FID) = DPR + T5

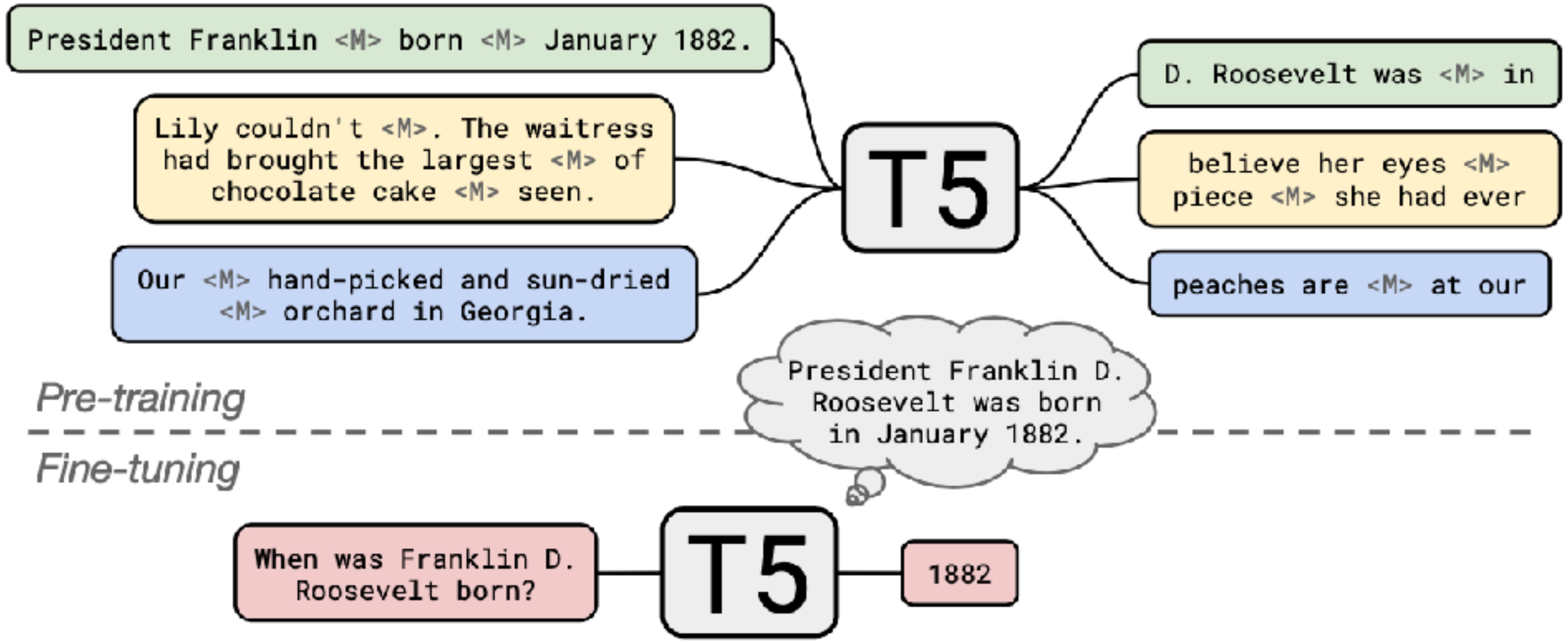


Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

Izacard and Grave 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

Large language models can do open-domain QA well

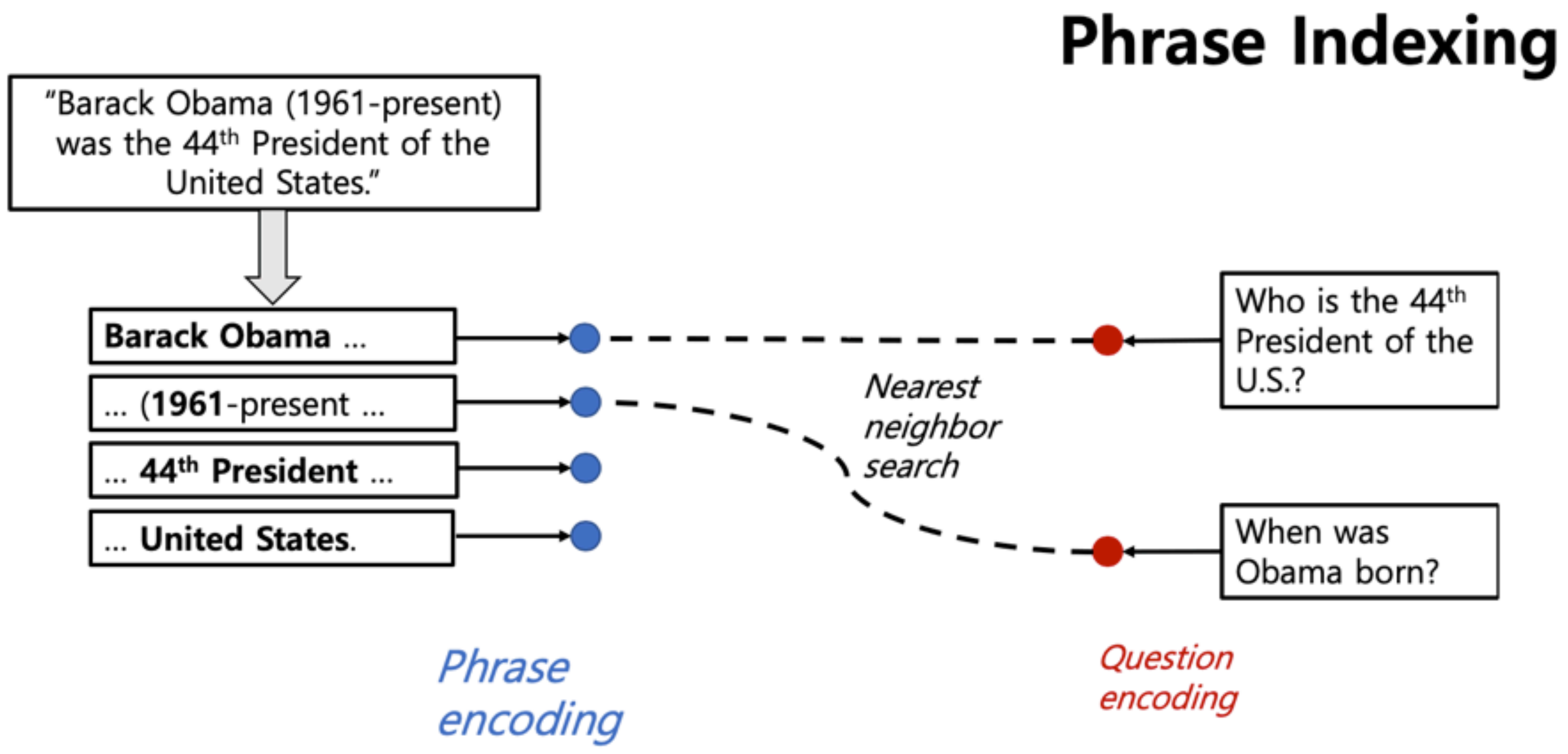
- … without an explicit retriever stage



Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Maybe the reader model is not necessary too!

It is possible to encode all the phrases (60 billion phrases in Wikipedia) using **dense** vectors and only do nearest neighbor search without a BERT model at inference time!



Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index

Lee et al., 2020. Learning Dense Representations of Phrases at Scale