SFU NatLangLab

CMPT 413/713: Natural Language Processing

NLP applications

Spring 2024
2024-03-18

Adapted from slides from Danqi Chen and Karthik Narasimhan
(with some content from slides from Chris Manning and Anoop Sarkar)

NLP applications

- Information extraction and question answering
- Text generation
- Dialogue and chatbots
- …

Question answering

Question Answering

- Goal: build computer systems to answer questions

Question	Answer
When were the first pyramids built?	2630 BC
What's the weather like in Vancouver?	42 F
Where is Einstein's house?	112 Mercer St, Princeton, NJ 08540
Why do we yawn?	When we're bored or tired we don't breathe as deeply as we normally do. This causes a drop in our blood-oxygen levels and yawning helps us counter-balance that.

Question Answering

- You can easily find these answers in google today!

Google

When were the first pyramids built?

All | Images | News | Shopping | Maps | More | Tools

About 17,000,000 results (0.76 seconds)

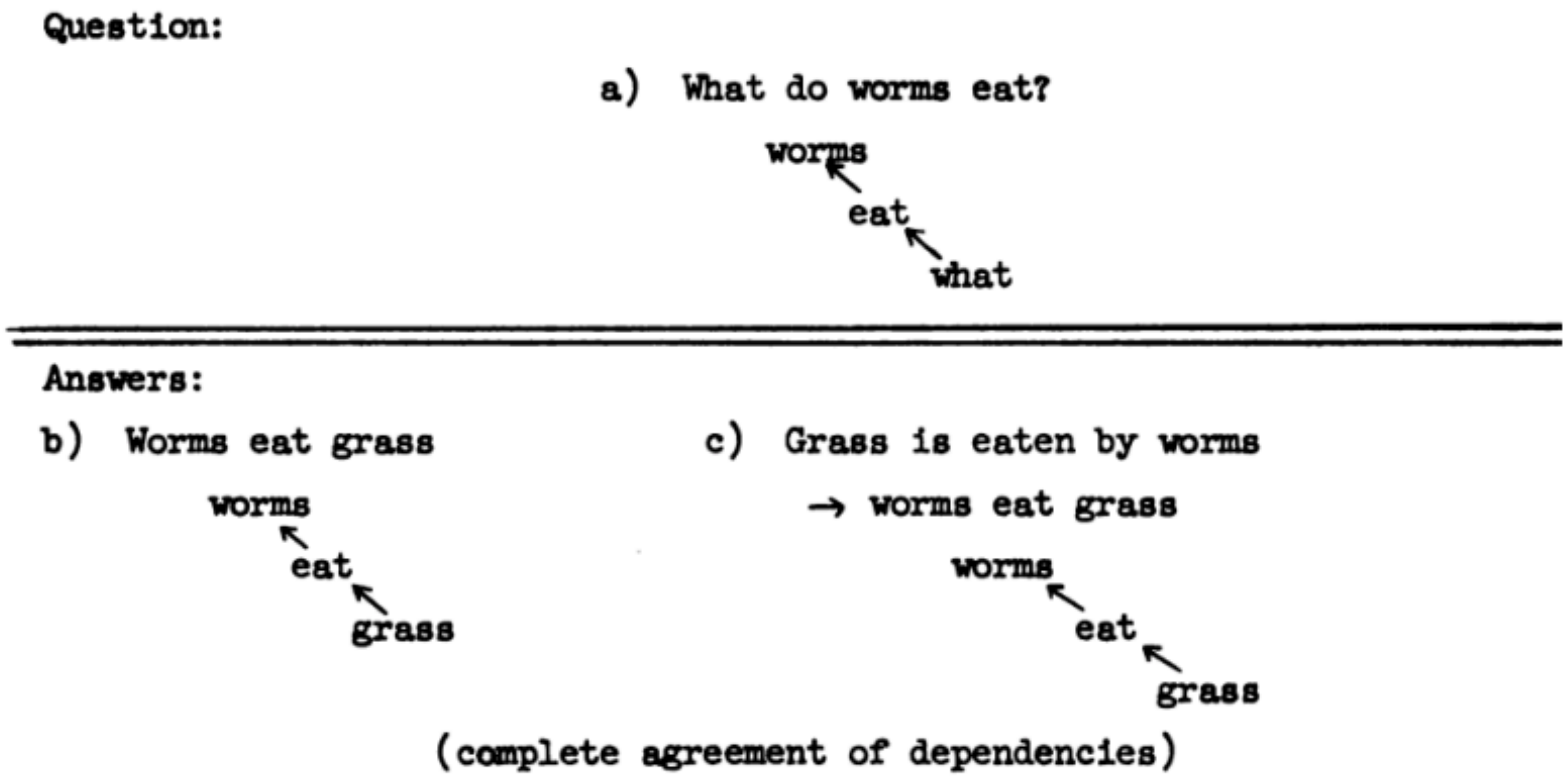
Around 2780 BCE, King Djoser's architect, Imhotep, built the first pyramid by placing six mastabas, each smaller than the one beneath, in a stack to form a pyramid rising in steps. This Step Pyramid stands on the west bank of the Nile River at Sakkara near Memphis.

https://www.si.edu › spotlight › ancient-egypt › pyramid

The Egyptian Pyramid | Smithsonian Institution

Question answer has a long history

Earliest QA system dated back to the 1960s!



Indexing and dependency logic for answering english questions
(Simmons et al, 1964)

Why care about question answering?

- Lots of immediate applications: search engines, dialogue systems
- Question answering is an important testbed for evaluating how well compute systems understand human language

THE PROCESS OF QUESTION ANSWERING

May 1977

Research Report #88

Wendy Lehnert

When a person understands a story, he can demonstrate his understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding. Question answering is therefore a task criterion for evaluating reading skills.

If a computer is said to understand a story, we must demand of the computer the same demonstrations of understanding that we require of people. Until such demands are met, we have no way of evaluating text understanding programs. Any computer programmer can write a program which inputs text. If the programmer assures us that his program 'understands' text, it is a bit like being reassured by a used car salesman about a suspiciously low speedometer reading. Only when we can ask a program to answer questions about what it reads will we be able to begin to assess that program's comprehension.

"Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**."

QA Taxonomy

- Context (and available information sources)
  - A passage, a document, a large collection of documents, all web documents
  - Knowledge base
  - Semi-structured tables
  - Images

- Question type
  - Factoid vs non-factoid
  - Open-domain vs closed-domain
  - Simple vs compositional

- Answer type
  - A short span of text
  - A paragraph
  - Yes/No
  - A database entry
  - A list

Textual Question Answering

Also called “Reading Comprehension”

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

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

Textual Question Answering

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

A) Fries

B) Pudding

C) James

D) Jane

2) What did James pull off of the shelves in the grocery store?

A) pudding

B) fries

C) food

D) splinters

(Richardson et al, 2013): MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Conversational Question Answering

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q: What are the candidates **running** for? A: Governor

Q: **Where?** A: Virginia

Q: Who is the democratic candidate? A: Terry McAuliffe

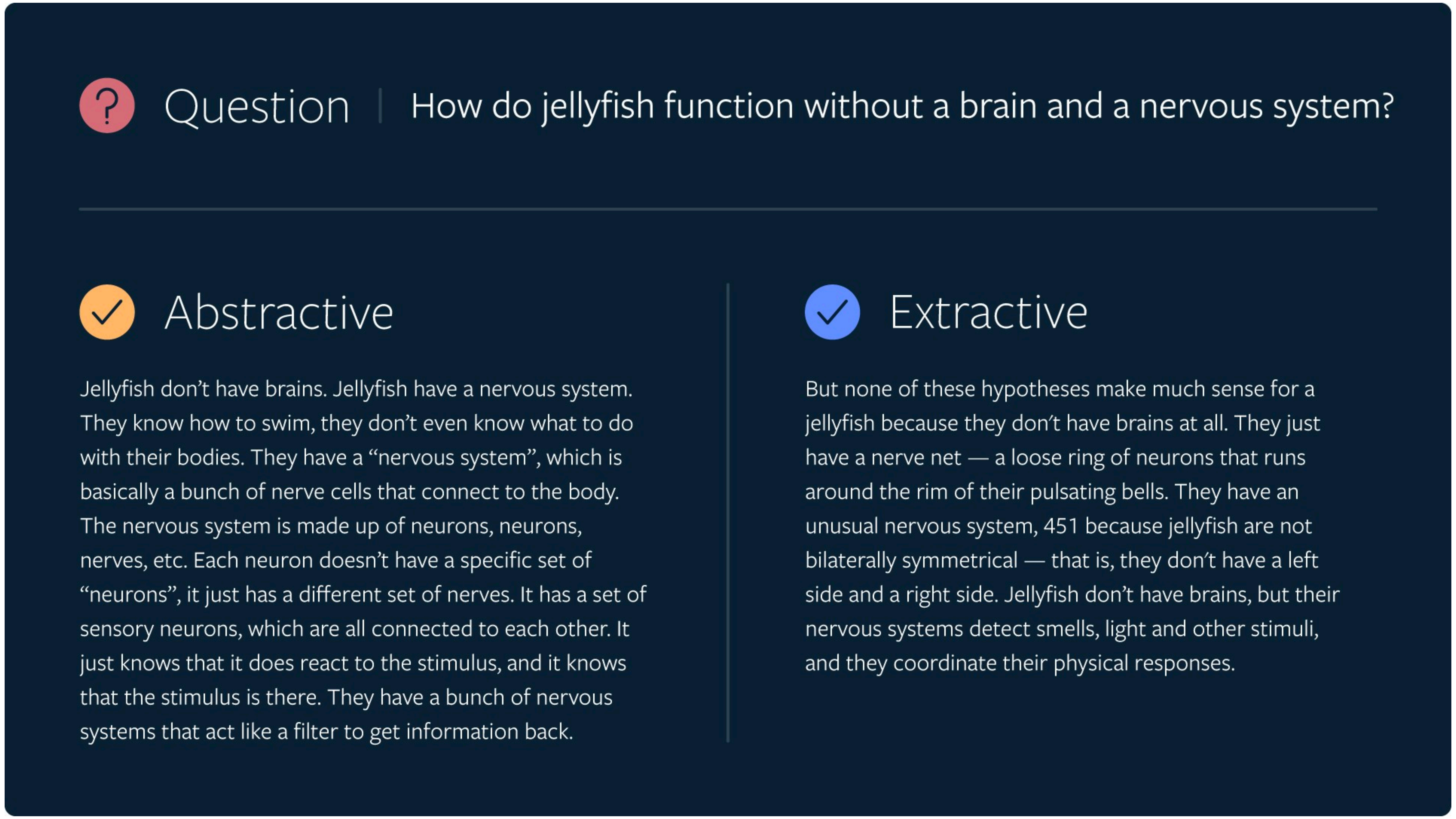
Q: Who is **his** opponent? A: Ken Cuccinelli

Q: What party does **he** belong to? A: Republican

Q: Which of **them** is winning?

(Reddy & Chen et al, 2019): CoQA: A Conversational Question Answering Challenge

Long-form Question Answering



Abstractive:
Answer made up of novel words and sentences composed through paraphrasing

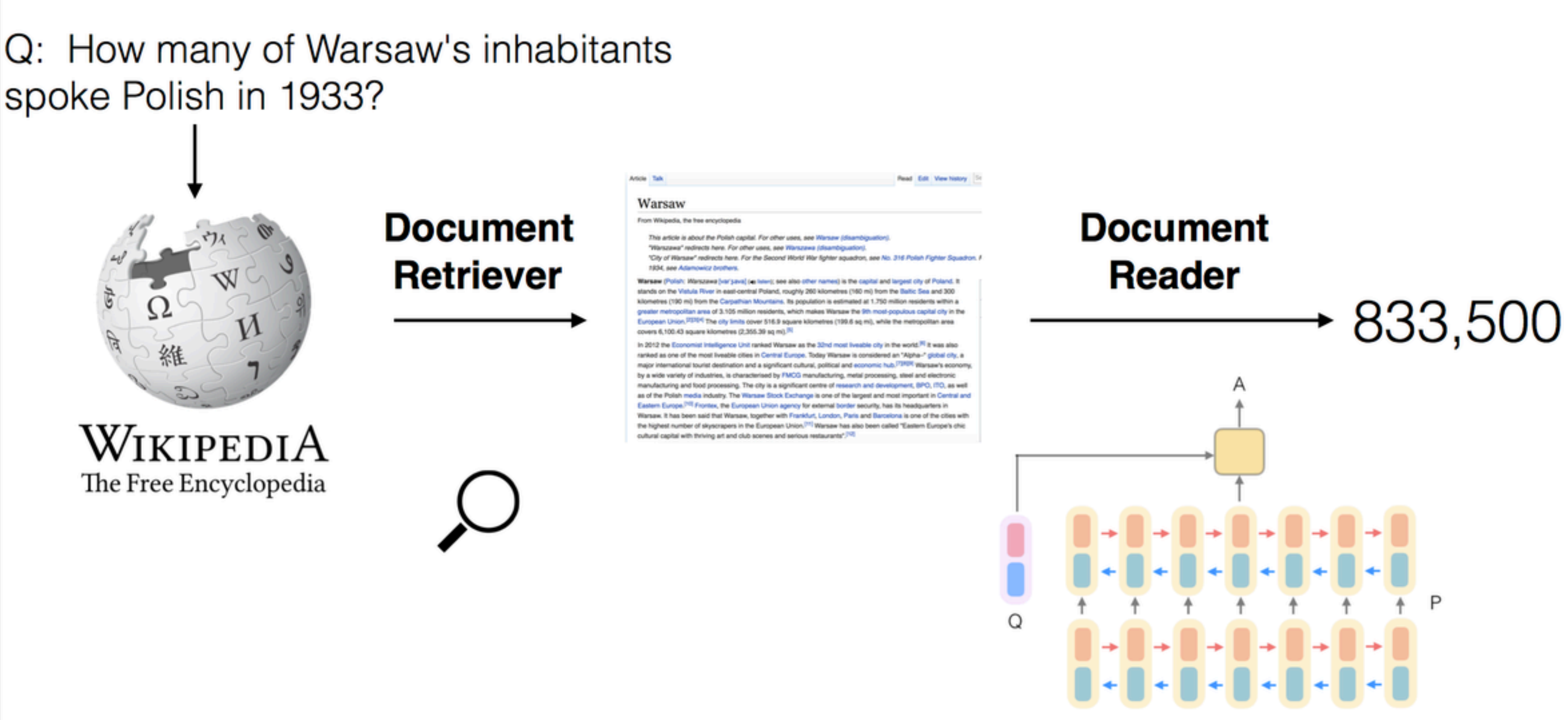
Extractive:
Select excerpts (extracts) and concatenate them to form the answer.

https://ai.facebook.com/blog/longform-qa/
(Fan et al, 2019): ELI5: Long Form Question Answering

Open-domain Question Answering

DrQA

- Factored into two parts:
  - Find documents that might contain an answer (handled with traditional information retrieval)
  - Finding an answer in a paragraph or a document (reading comprehension)



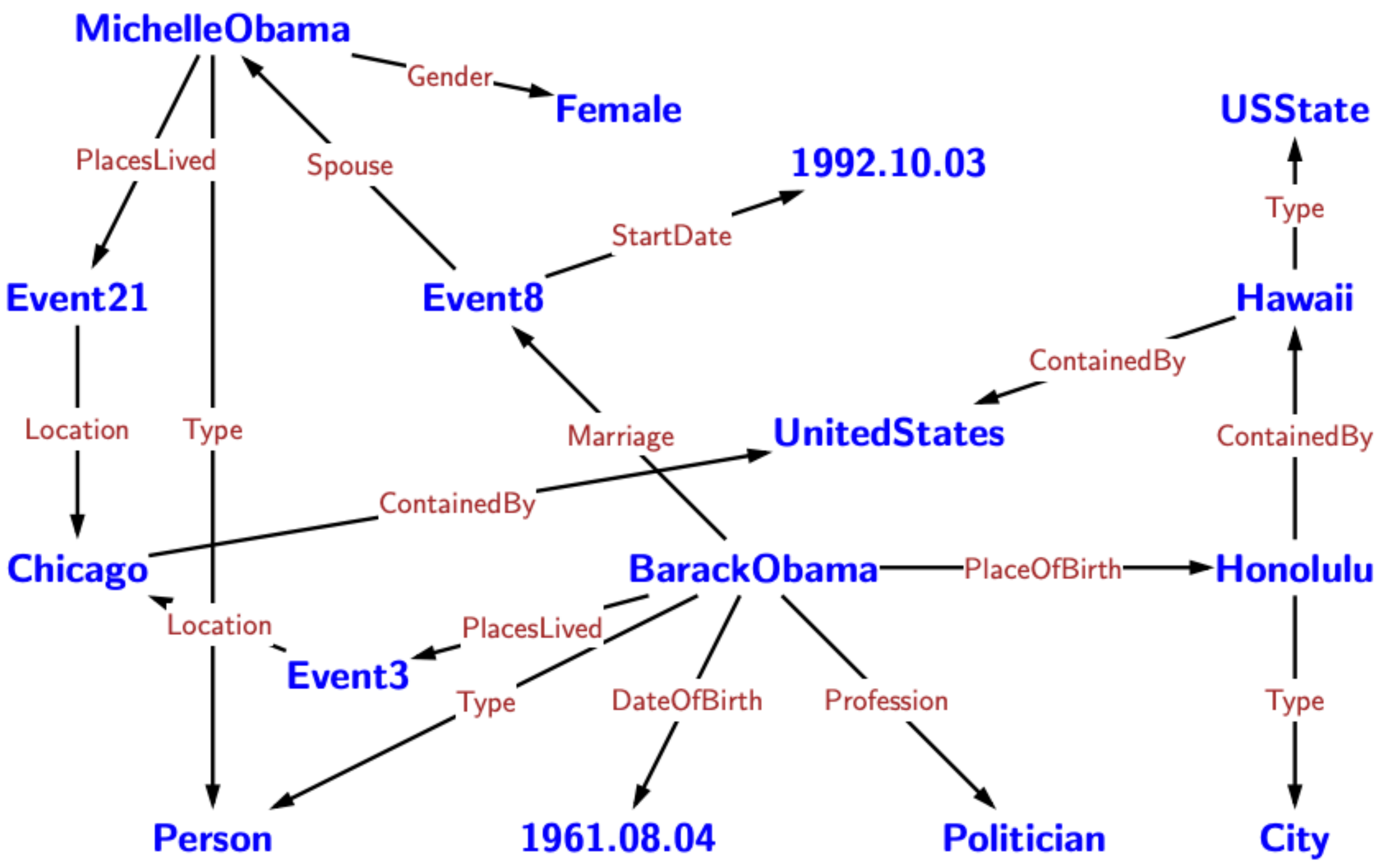

```
>>> process('What is the answer to life, the universe, and everything?')

Top Predictions:
+------+--------+--------------------------------------------------+--------------+-----------+
| Rank | Answer |                       Doc                        | Answer Score | Doc Score |
+------+--------+--------------------------------------------------+--------------+-----------+
|  1   |   42   | Phrases from The Hitchhiker's Guide to the Galaxy |    47242     |   141.26  |
+------+--------+--------------------------------------------------+--------------+-----------+
```

(Chen et al, 2017): Reading Wikipedia to Answer Open-Domain Questions

Knowledge Base Question Answering

Freebase

100M entities (nodes)   1B assertions (edges)



Structured knowledge representation

Which states' capitals are also their largest cities by area?

semantic parsing

$\mu x.\text{Type.USState} \sqcap \text{Capital.argmax}(\text{Type.City} \sqcap \text{ContainedBy}.x, \text{Area})$

execute

Arizona, Hawaii, Idaho, Indiana, Iowa, Oklahoma, Utah

QA via semantic parsing

(Berant et al, 2013): Semantic Parsing on Freebase from Question-Answer Pairs

Table-based Question Answering

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004

(Pasupat and Liang, 2015): Compositional Semantic Parsing on Semi-Structured Tables.

Visual Question Answering

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

Reading Comprehension

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

UCD ***finished*** the 2006 championship as Dublin champions , by ***beating*** St Vincents in the final .

finished

Who finished something? - UCD
What did someone finish? - the 2006 championship
What did someone finish something as? - Dublin champions
How did someone finish something? - by beating St Vincents in the final

beating

Who beat someone? - UCD
When did someone beat someone? - in the final
Who did someone beat? - St Vincents

(He et al. 2015)

Slide credit: John Hewitt

Stanford Question Answering Dataset (SQuAD)

Passage

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.

Question: Which NFL team won Super Bowl 50?
Answer: Denver Broncos

Question: What does AFC stand for?
Answer: American Football Conference

Question: What year was Super Bowl 50?
Answer: 2016

SQuAD 2.0:
Have classifier/threshold to decide whether to take the most likely prediction as answer

- (passage, question, answer) triples
- Passage is from Wikipedia (~100-500 words), question is crowd-sourced
- Answer must be a span of text in the passage (aka. "extractive question answering")
- SQuAD 1.1: 100k answerable questions, SQuAD 2.0: another 50k unanswerable questions

https://stanford-qa.com
(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

3 gold answers are collected for each question

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

Slide credit: Chris Manning

Stanford Question Answering Dataset (SQuAD)

SQuAD 1.1 evaluation:

- Two metrics: exact match (EM) and F1
  - Exact match: 1/0 accuracy on whether you match one of the three answers
  - F1: take each gold answer and system output as bag of words, compute precision, recall and harmonic mean. Take the max of the three scores.
- Final exact match and F1 are average of instance exact and F1 scores
- Estimated human performance: EM = 82.3, F1 = 91.2

Example

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

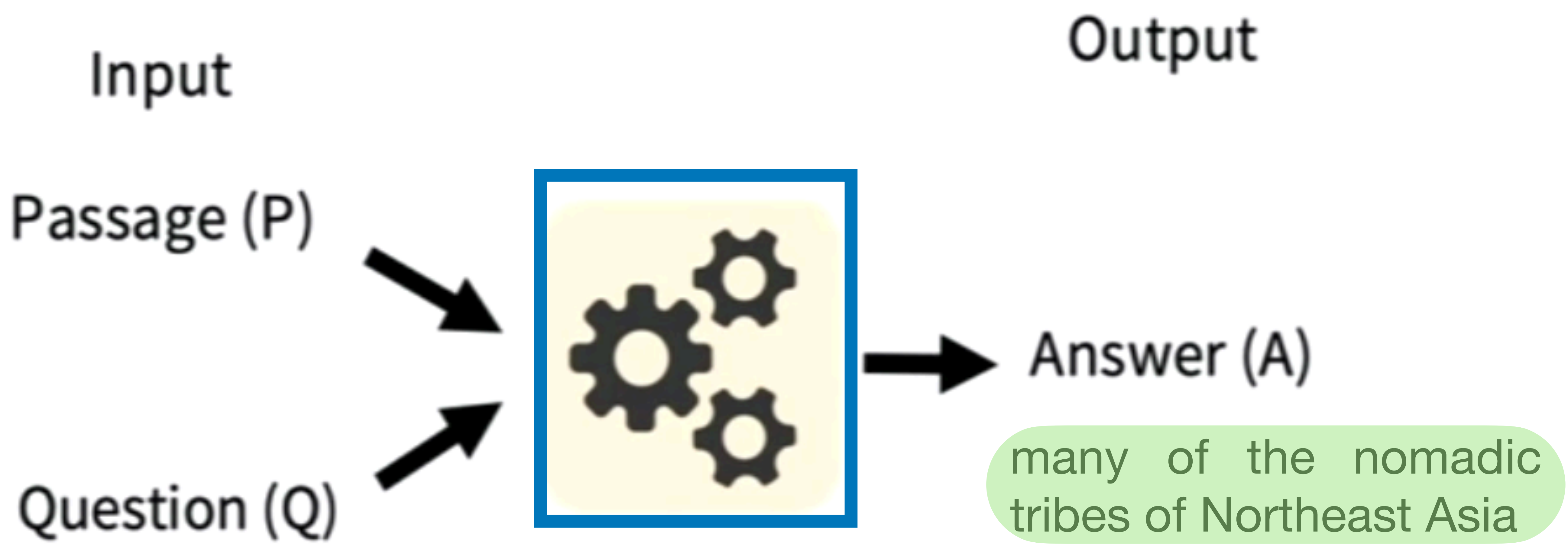
F1: max{0.67, 0.67, 0.61} = 0.67

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

Models for Reading Comprehension

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

Who did **Genghis Khan unite before he** began **conquering** the rest of **Eurasia**?



Feature-based models (2016)

- Generate a list of candidate answers $\{a_1, a_2, \ldots, a_M\}$
  - Considered only the constituents in parse trees

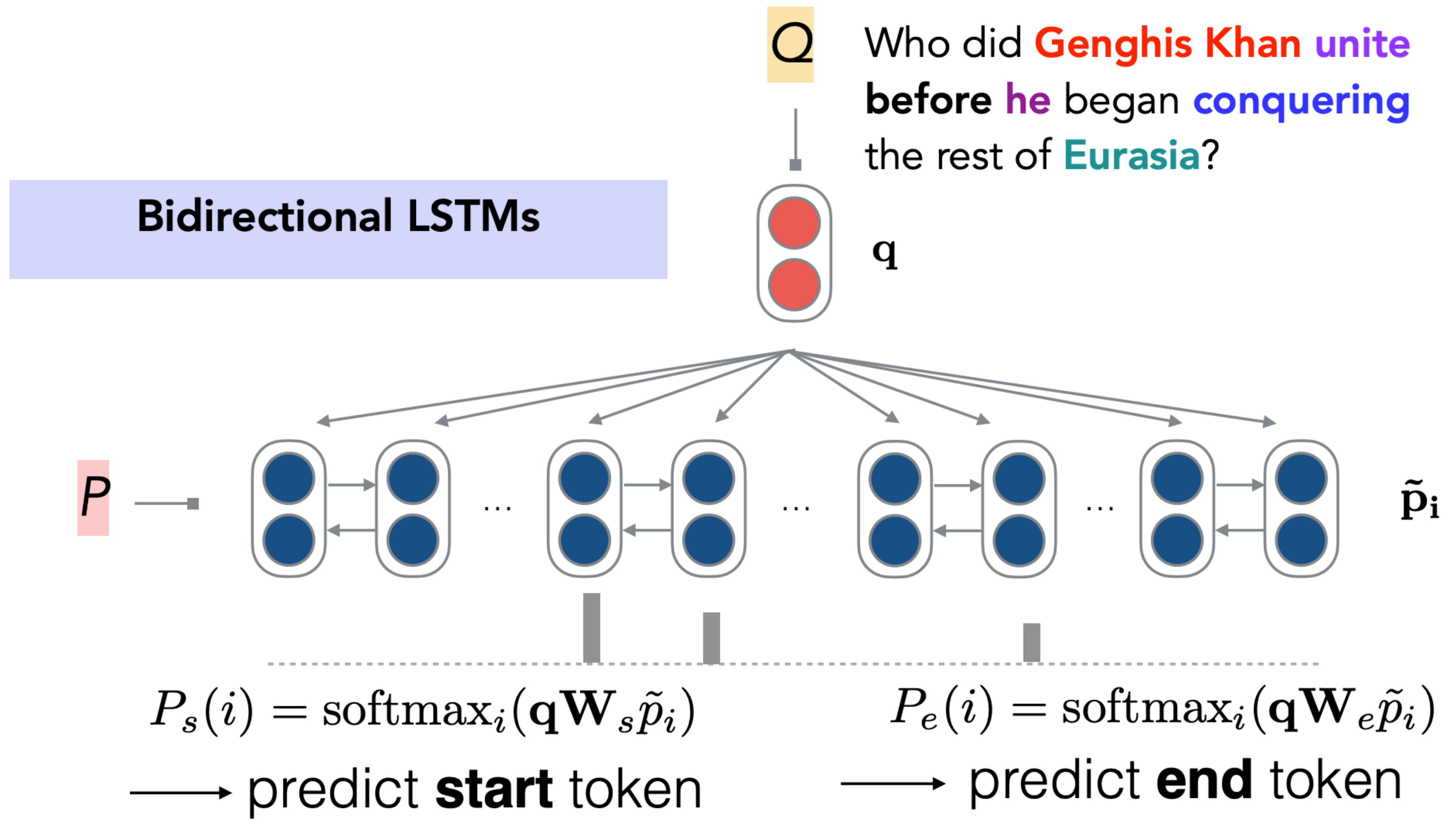
- Define a feature vector $\phi(p, q, a_i) \in \mathbb{R}^d$:
  - Word/bigram frequencies
  - Parse tree matches
  - Dependency labels, length, part-of-speech tags

- Apply a (multi-class) logistic regression model

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

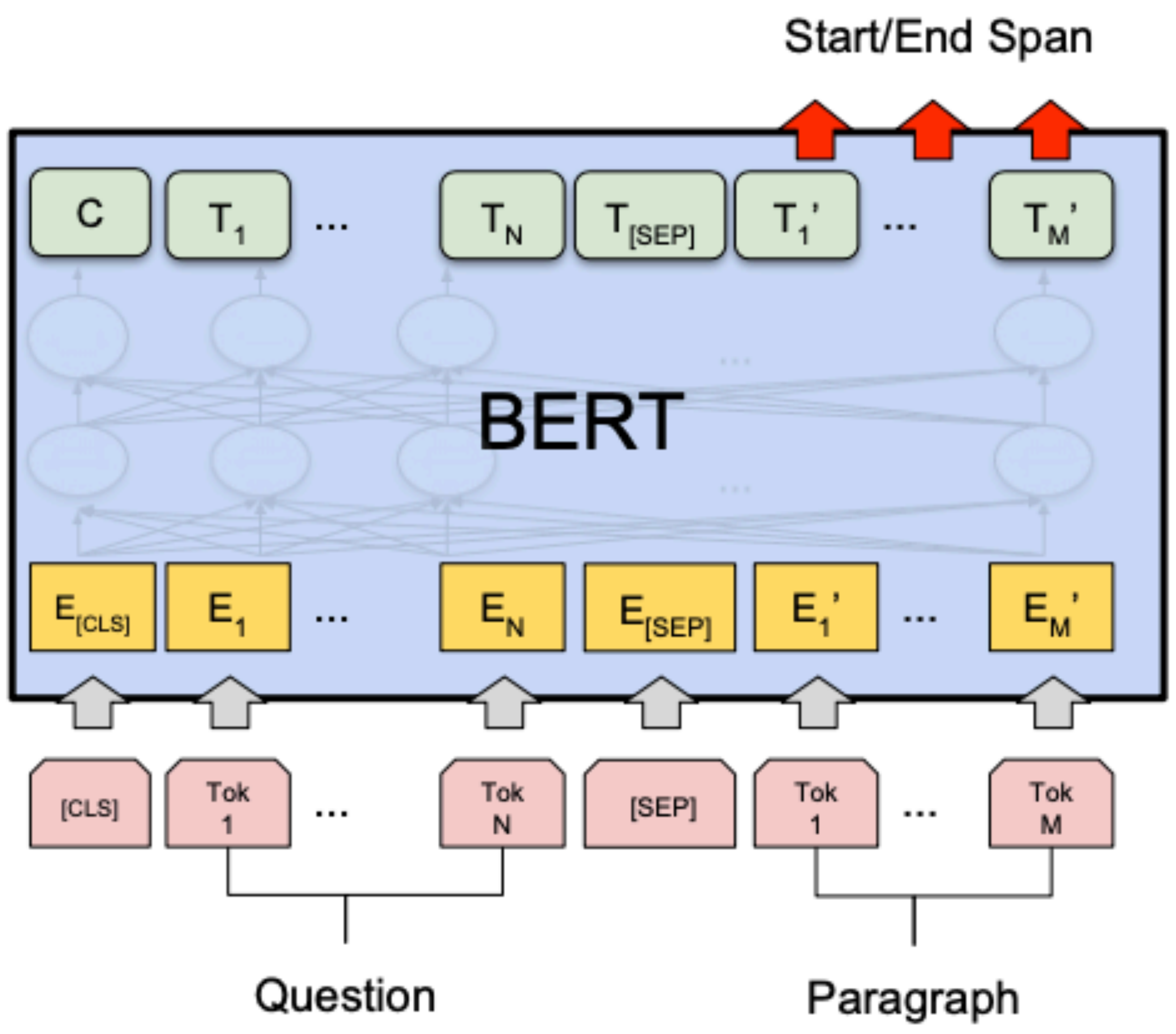
Neural models for reading comprehension (after 2016)

- LSTM-based models with attention (2016-2018)



Chen et al, 2016

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

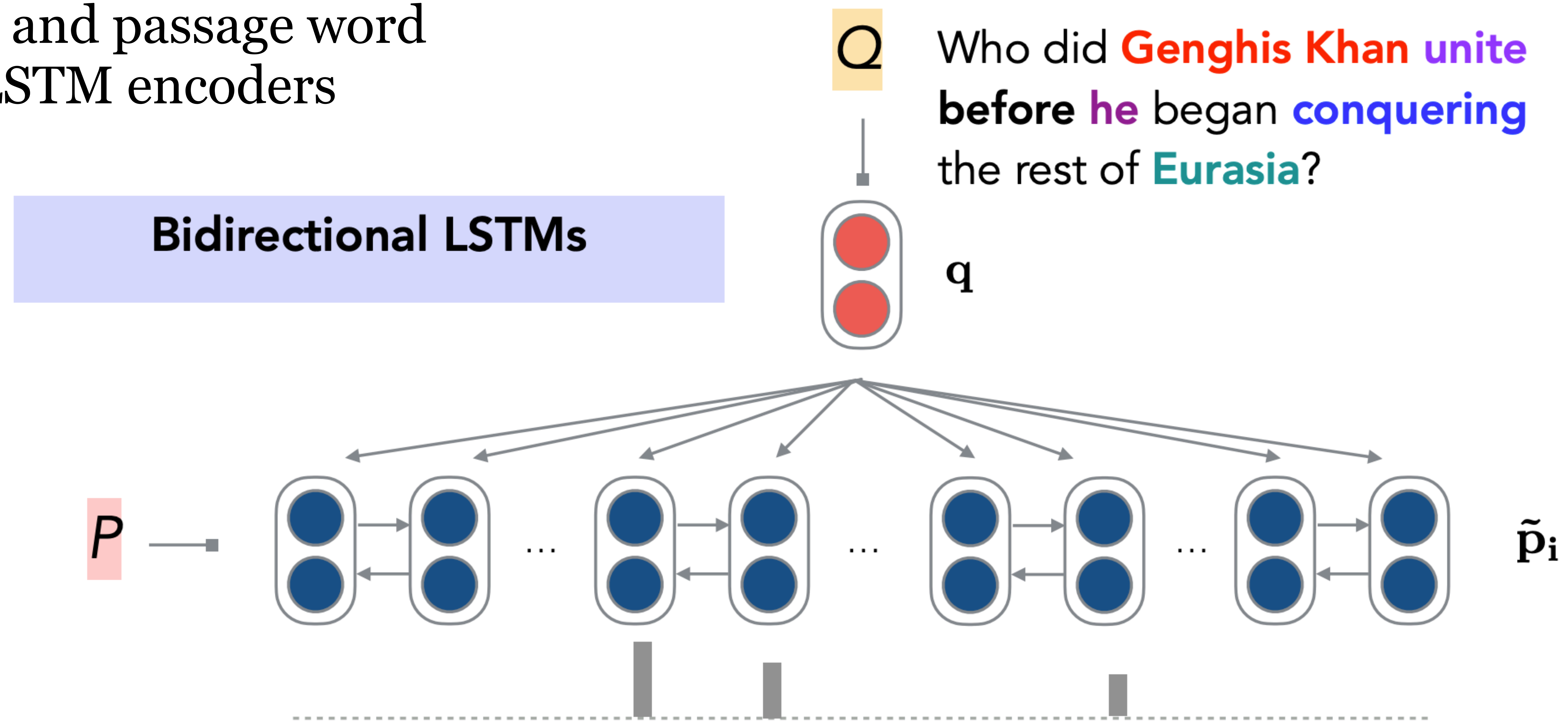


Devlin et al, 2018

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)...

Stanford Attentive Reader (Chen, Bolten, and Manning, 2016)

- Simple model with good performance
- Encode the question and passage word embeddings and BiLSTM encoders



Use **attention** to predict start and end span

$$P_s(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$$

→ predict **start** token

$$P_e(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$$

→ predict **end** token

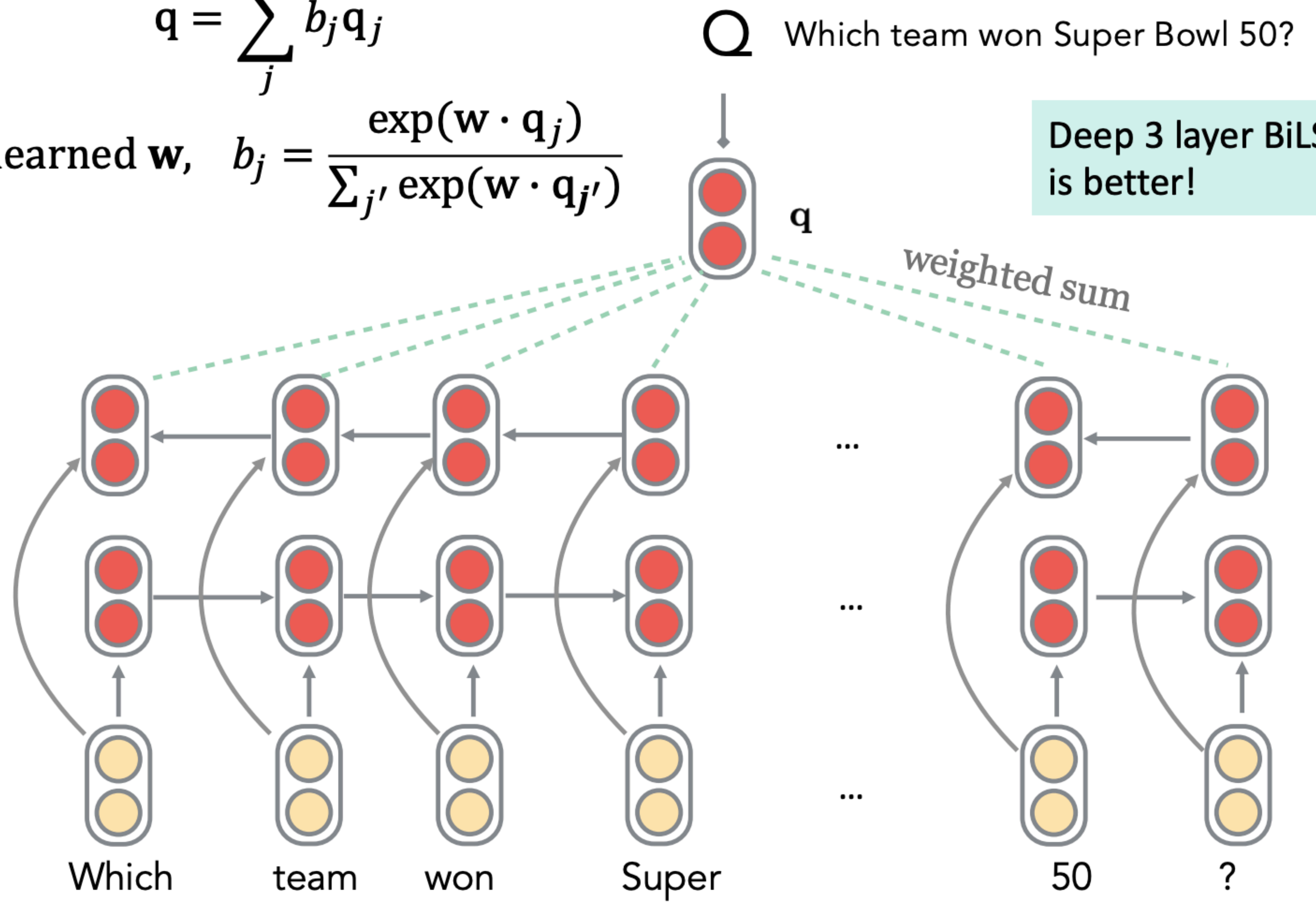
Stanford Attentive Reader++ (from DrQA)

$$\mathbf{q} = \sum_j b_j \mathbf{q}_j$$

For learned $\mathbf{w}$, $\quad b_j = \frac{\exp(\mathbf{w} \cdot \mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w} \cdot \mathbf{q}_{j'})}$

Take weighted sum of hidden states at all time steps of LSTM!

Deep 3 layer BiLSTM is better!



Slide credit: Chris Manning

Stanford Attentive Reader++ (from DrQA)

Reading Wikipedia to Answer Open-Domain Questions [Chen et al. 2017]

- Vector representation $\mathbf{p}_i$ of each token p_i in passage made from concatenation of
  - Word embedding (GloVe 300d): $\mathbf{E}(p_i)$
  - Exact match (whether the word appeared in the question)
    - 3 binary features: exact, uncased, lemma
  - Linguistic features: POS & NER tags (one-hot encoded)
  - Term frequency (unigram probability)
  - Aligned question embedding ("car" vs "vehicle")
    - Weighted sum over embedded question tokens $\mathbf{E}(q_j)$ with attention score $a_{i,j}$ (α is a single dense layer with ReLU nonlinear)

Improved passage word/position representations

Matching of words in the question to words in the passage

$$f_{\text{align}} = \sum_j a_{i,j} \mathbf{E}(q_j)$$

$$a_{i,j} = \frac{\exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j)))}{\sum_{j'} \exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q'_j)))}$$

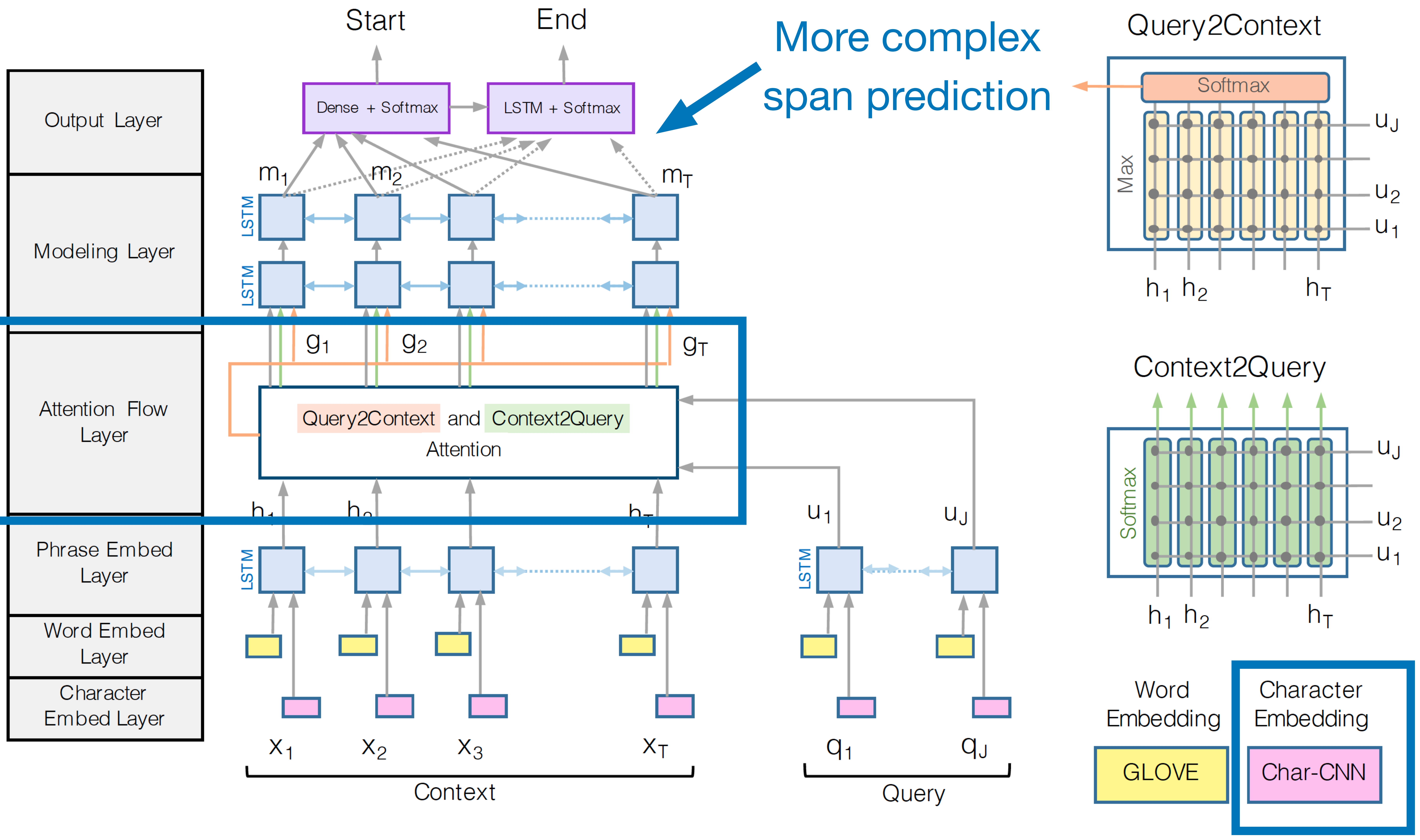
Slide credit: Chris Manning

BiDAF

Bidirectional Attention Flow for Machine Comprehension [Seo et al, 2017]

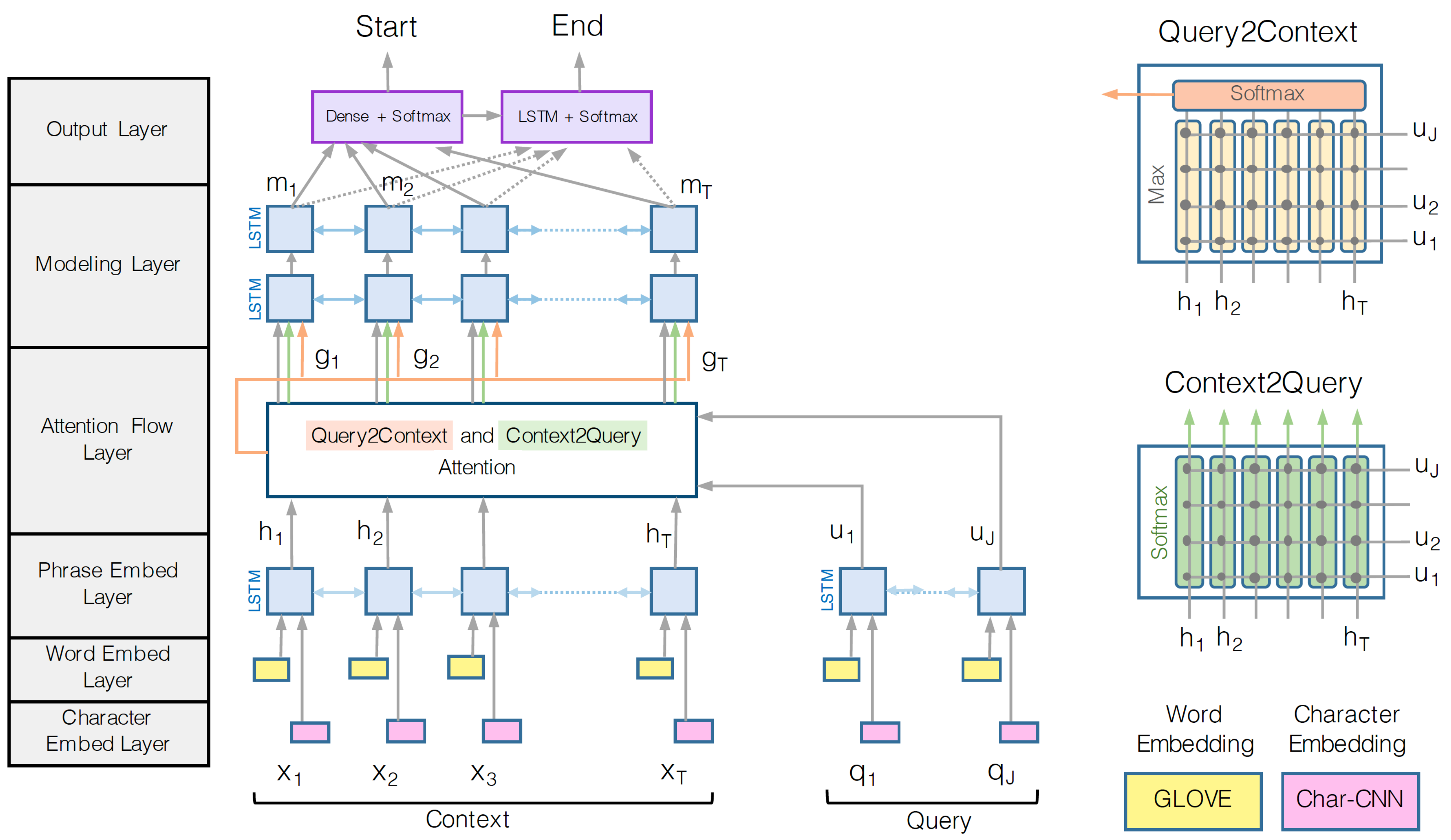
Attention flowing between question (query) and passage (context)

More complex span prediction



BiDAF

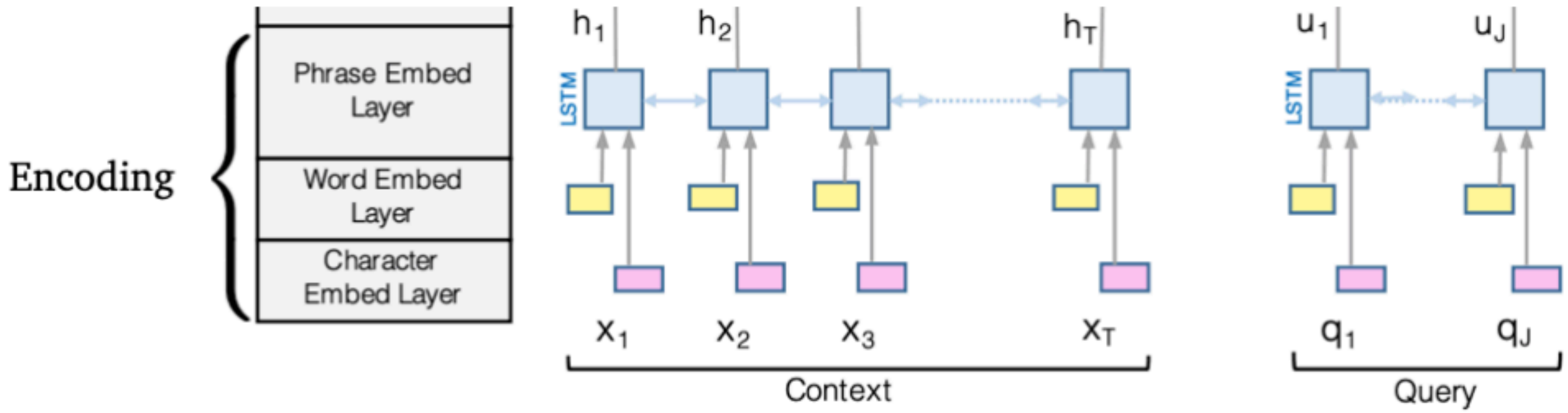
- Encode the question using word/character embeddings; pass to an biLSTM encoder
- Encode the passage similarly
- Passage-to-question and question-to-passage attention

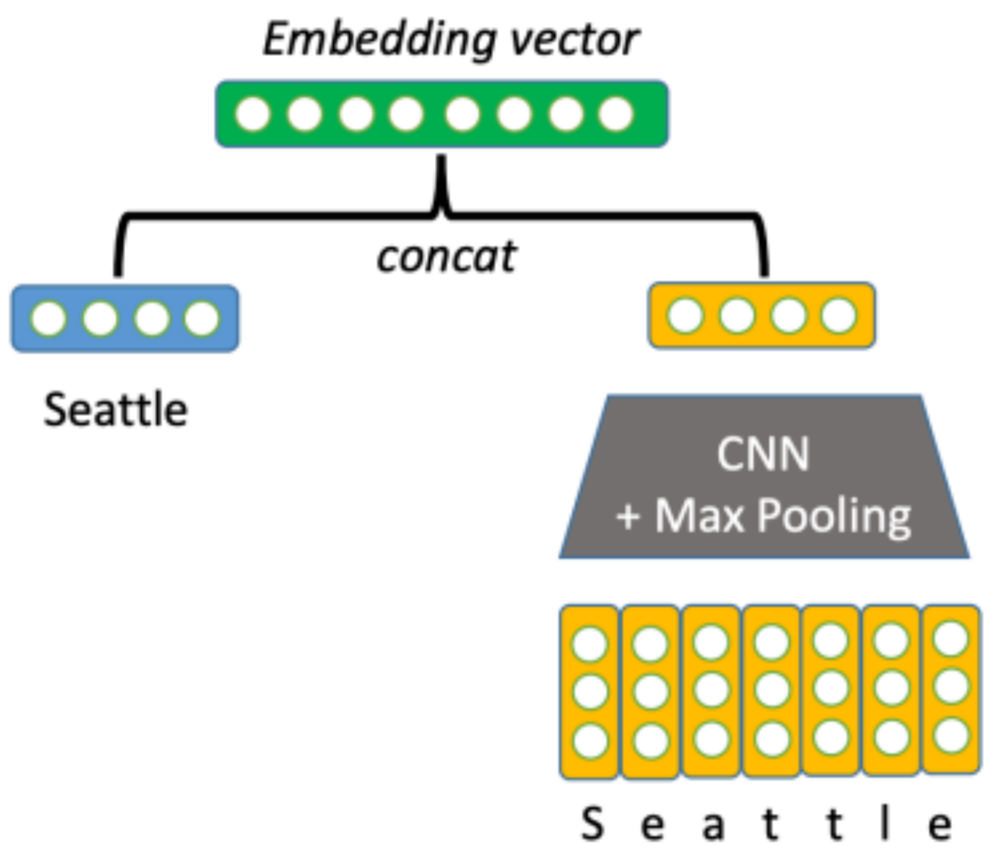


- Modeling layer: another BiLSTM layer
- Output layer: two classifiers for predicting start and end points
- The entire model can be trained in an end-to-end way

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

BiDAF: Encoding



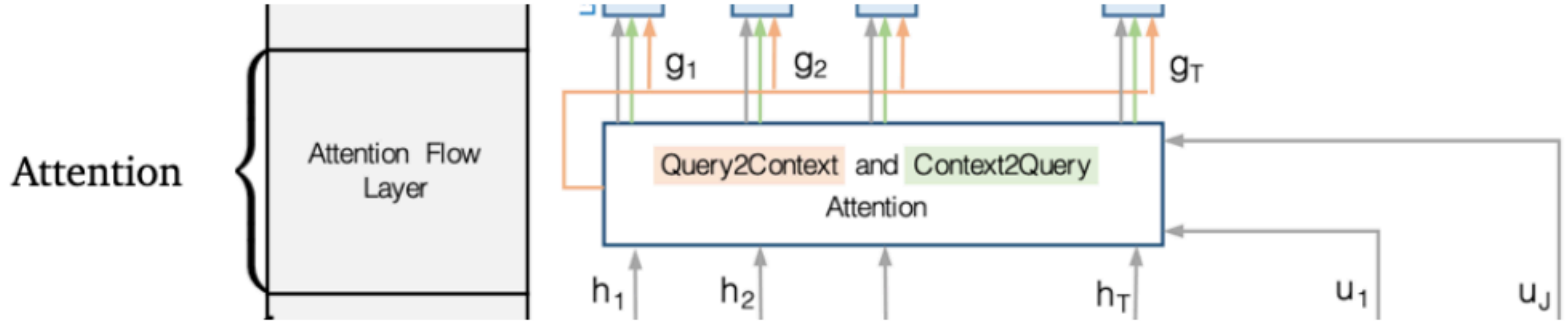


- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query
- Then, use two bidirectional LSTMs separately to produce contextual embeddings for both context and query

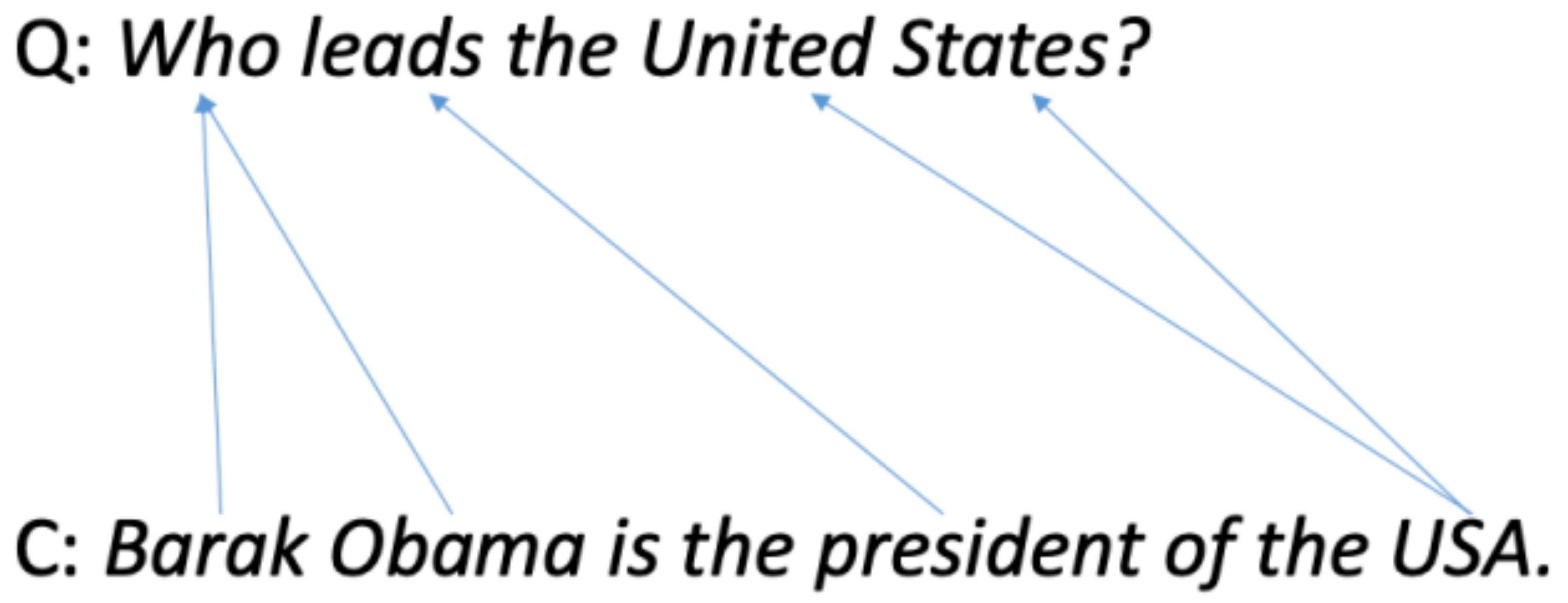
$$\overrightarrow{\mathbf{c}}_i = \mathrm{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H$$
$$\overleftarrow{\mathbf{c}}_i = \mathrm{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H$$
$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H}$$

$$\overrightarrow{\mathbf{q}}_i = \mathrm{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$
$$\overleftarrow{\mathbf{q}}_i = \mathrm{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$
$$\mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

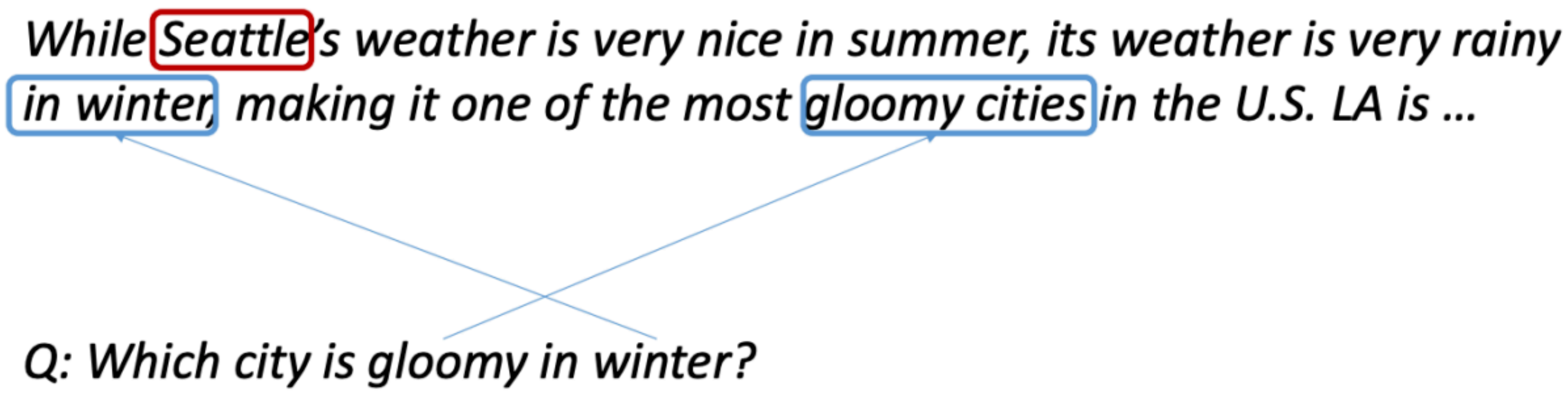
BiDAF: Attention



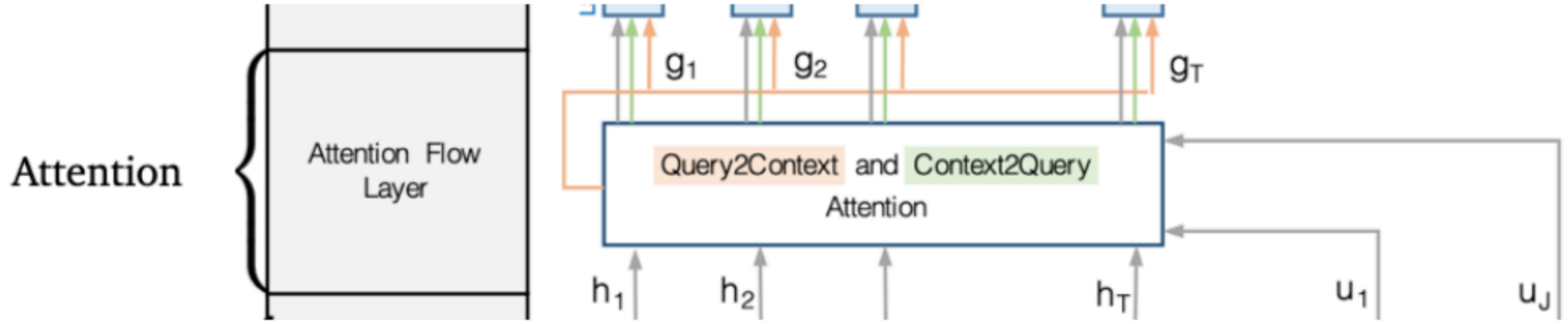
- Context-to-query attention: For each context word, choose the most relevant words from the query words.



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



BiDAF: Attention



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

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

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

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

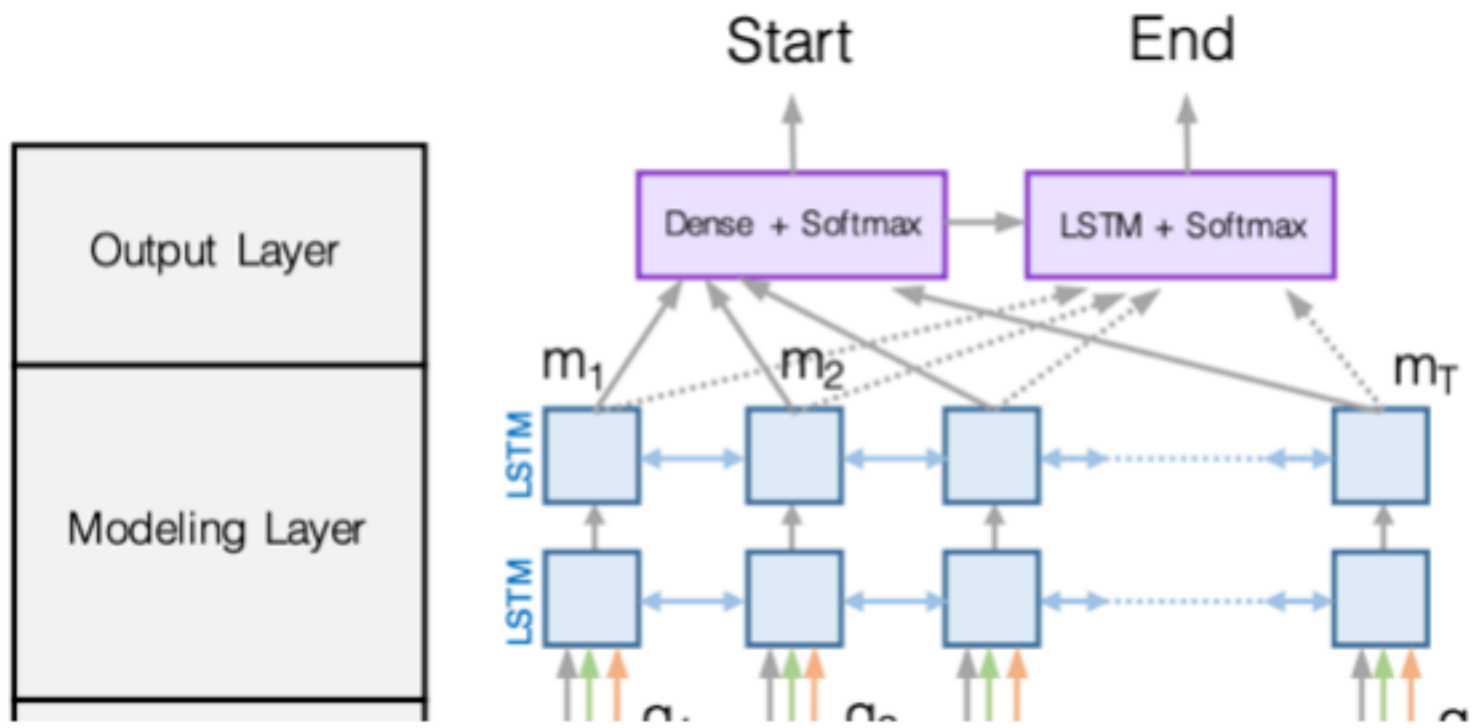
- Query-to-context attention (which context words are relevant to some question words):

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

The final output is

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

BiDAF: Modeling and output layers



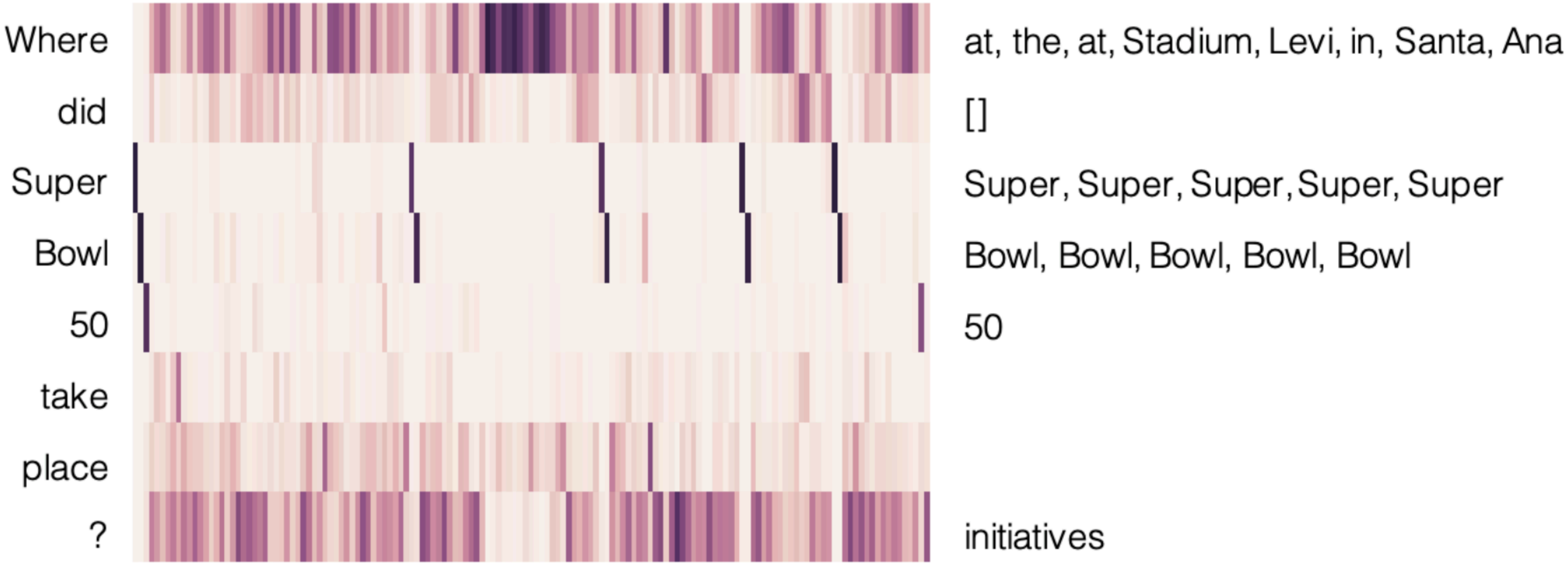
- Modeling layer: pass $\mathbf{g}_i$ to another two layers of bi-directional LSTMs.
  - Attention layer is modeling interactions between query and context
  - Modeling layer is modeling interactions within context words
- Output layer: two classifiers predicting the start and end positions

The final training loss is

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

Visualizing attention

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 .



SQuAD v1.1 performance (2017)

	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

Slide credit: Chris Manning

LSTM vs BERT based models

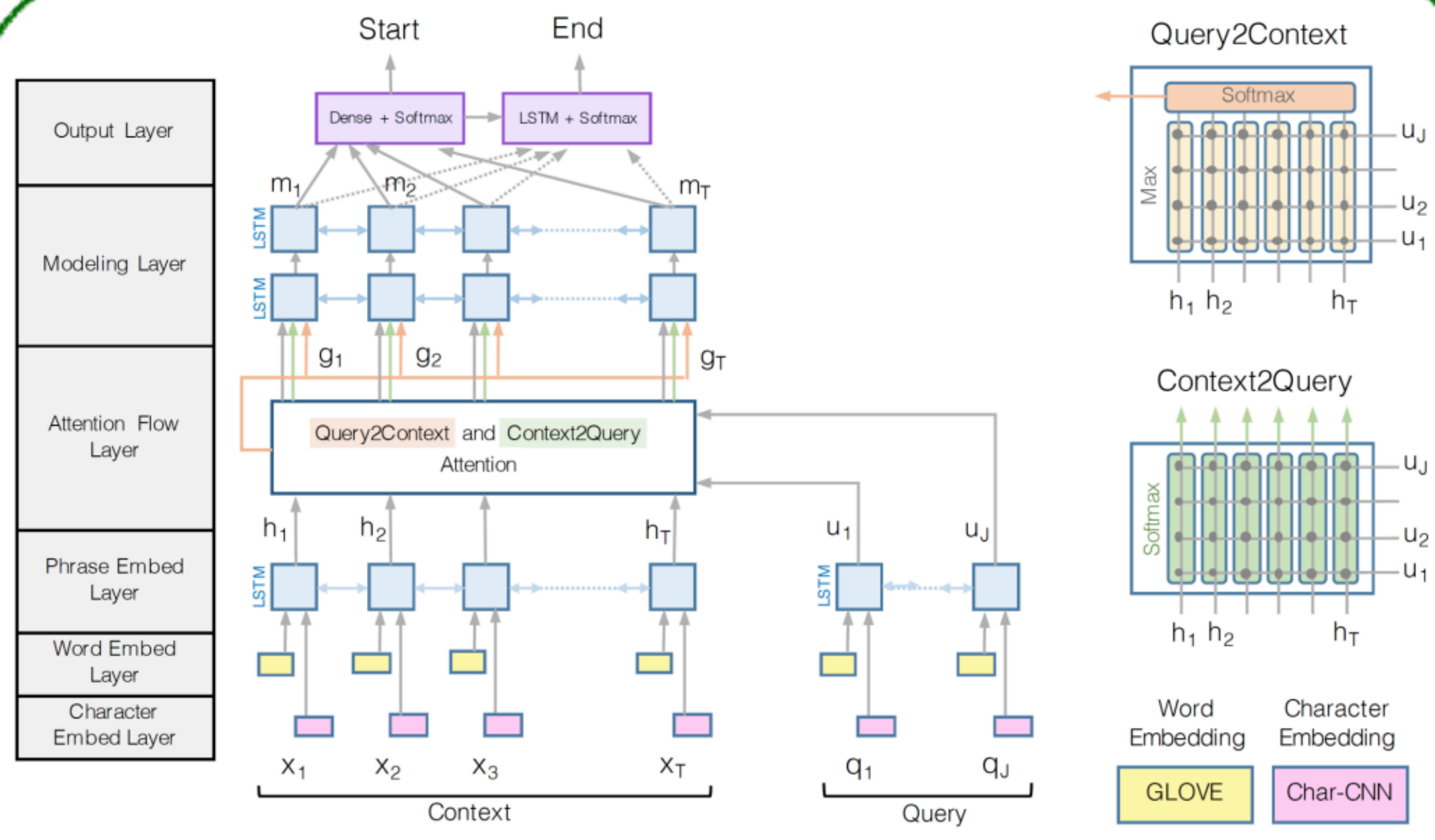


Image credit: (Seo et al, 2017)

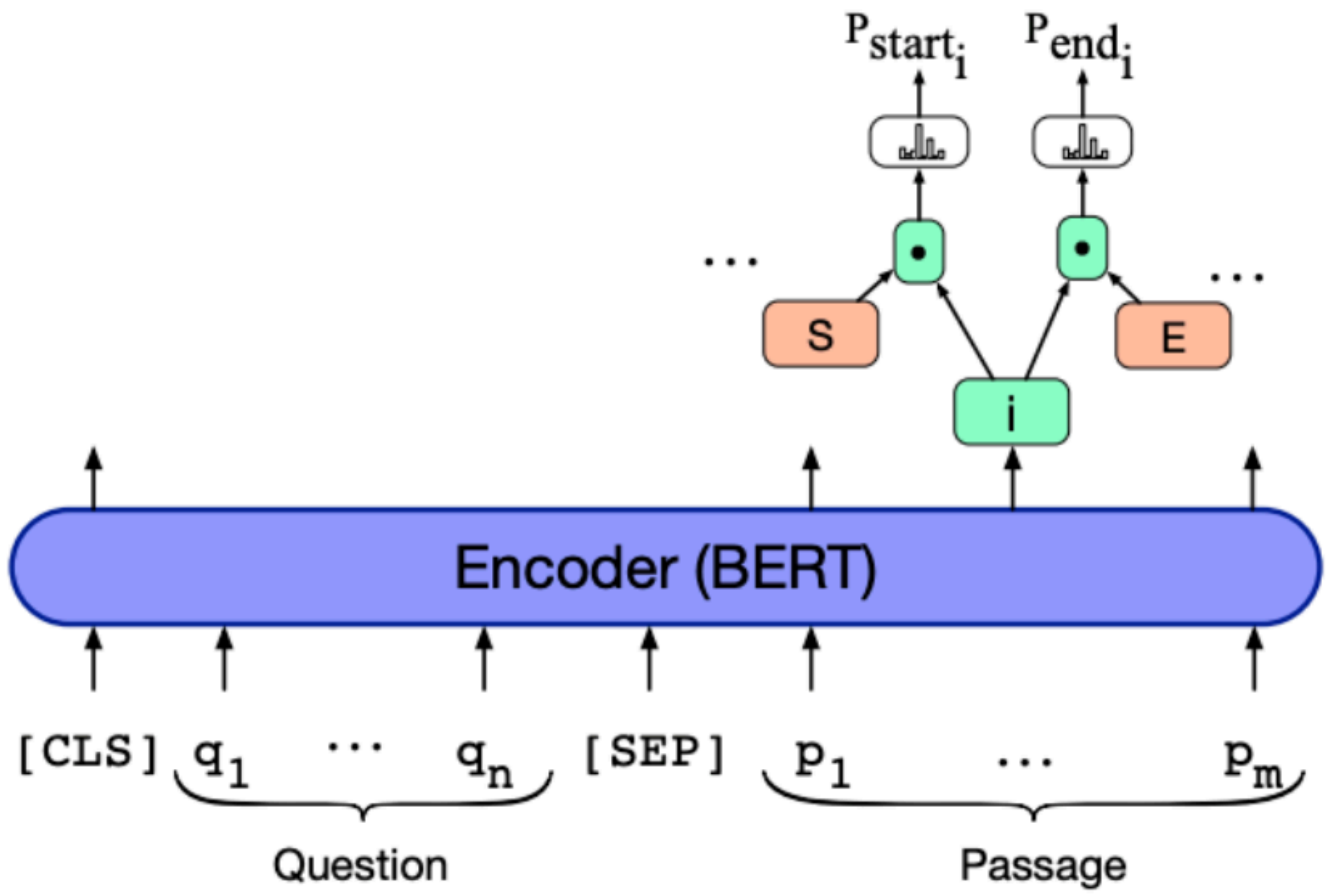
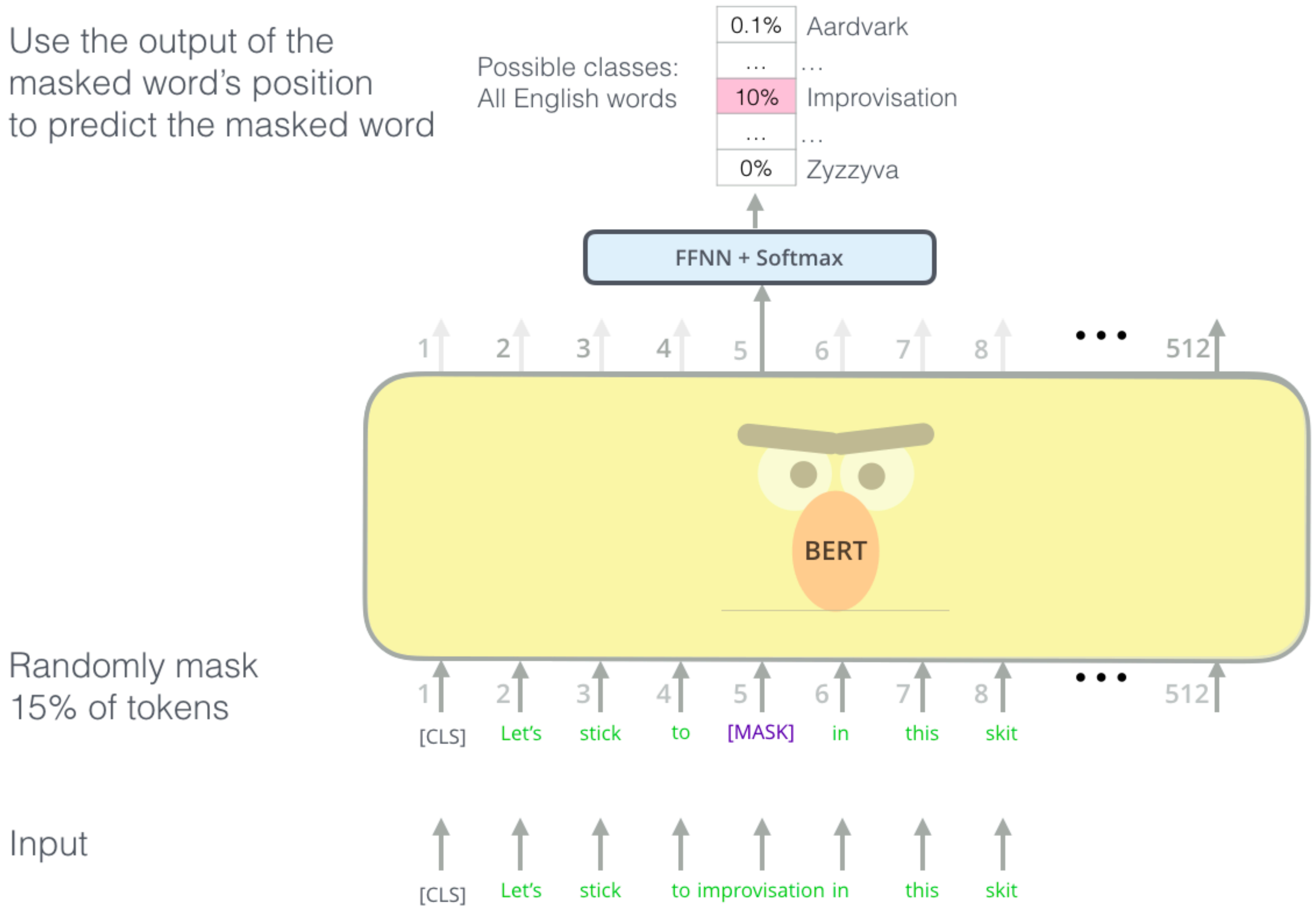


Image credit: J & M, edition 3

BERT-based models

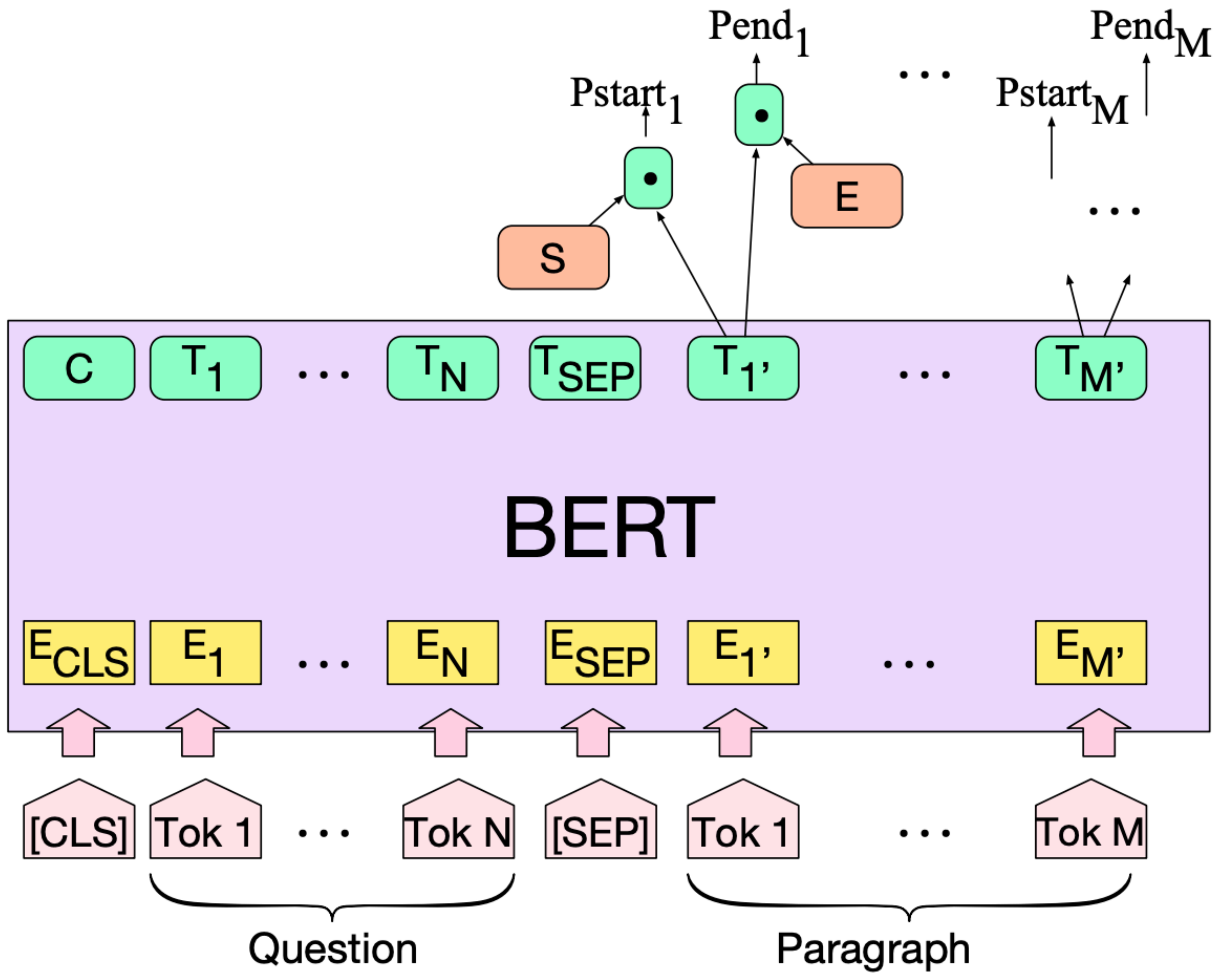


Pre-training

BERT-based models

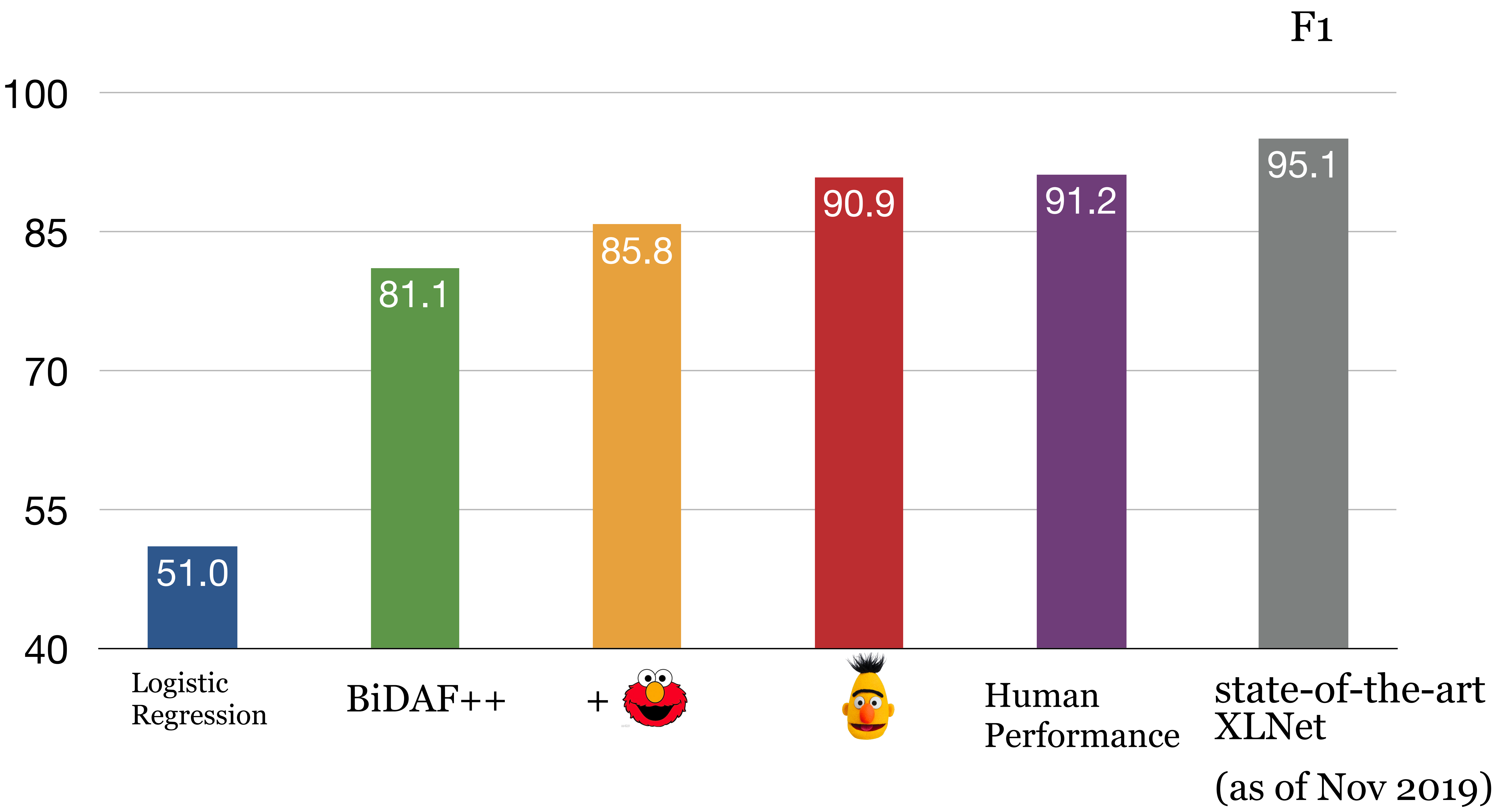
$$Pstart_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

$$Pend_i = \frac{e^{E \cdot T_i}}{\sum_j e^{E \cdot T_j}}$$



- Concatenate question and passage as one single sequence separated with a [SEP] token, then pass it to the BERT encoder
- Train two classifiers on top of the passage tokens

Experiments on SQuAD v1.1



*: single model only

Comparison 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.

Comparison between BIDAF and BERT models

- BERT model has many many more parameters (110M or 330M) and 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).

SQuAD Limitations

- SQuAD has a number of limitations:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between question and answer span
  - Barely any multi-fact/sentence inference beyond coreference

- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - The most used and competed QA dataset
  - A useful starting point for building systems in industry (although in-domain data always really helps!)

Slide credit: Chris Manning

Beyond SQUAD 1.1

- SQuAD 2.0 (Rajparkar et al, 2018)
    - unanswerable questions
- HotPotQA (Yang et al, 2018)
    - multi-hop reasoning
- QuAC(Choi et al, 2018) and CoQA (Reddy et al, 2018)
    - conversational QA
- Natural Questions (Kwiatkowski et al, 2019)
    - Real world questions issued to Google
- BooIQ (Clark et al, 2019)
    - Hard yes/no questions from Google queries

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q_1: What are the candidates **running** for?
A_1: Governor
R_1: The Virginia governor's race

Q_2: **Where**?
A_2: Virginia
R_2: The Virginia governor's race

Q_3: Who is the democratic candidate?
A_3: **Terry McAuliffe**
R_3: Democrat Terry McAuliffe

Q_4: Who is **his** opponent?
A_4: **Ken Cuccinelli**
R_4 Republican Ken Cuccinelli

Q_5: What party does **he** belong to?
A_5: Republican
R_5: Republican Ken Cuccinelli

Q_6: Which of **them** is winning?
A_6: Terry McAuliffe
R_6: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

CoQA (Reddy et al, 2018)

Beyond SQUAD 1.1

Natural Questions

Real world queries to Google

Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

Example 3
Question: why does queen elizabeth sign her name elizabeth r
Wikipedia Page: Royal_sign-manual
Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).

Short answer: NULL

(Kwiatkowski et al, 2019)

BoolQ

Hard yes/no questions from Google queries

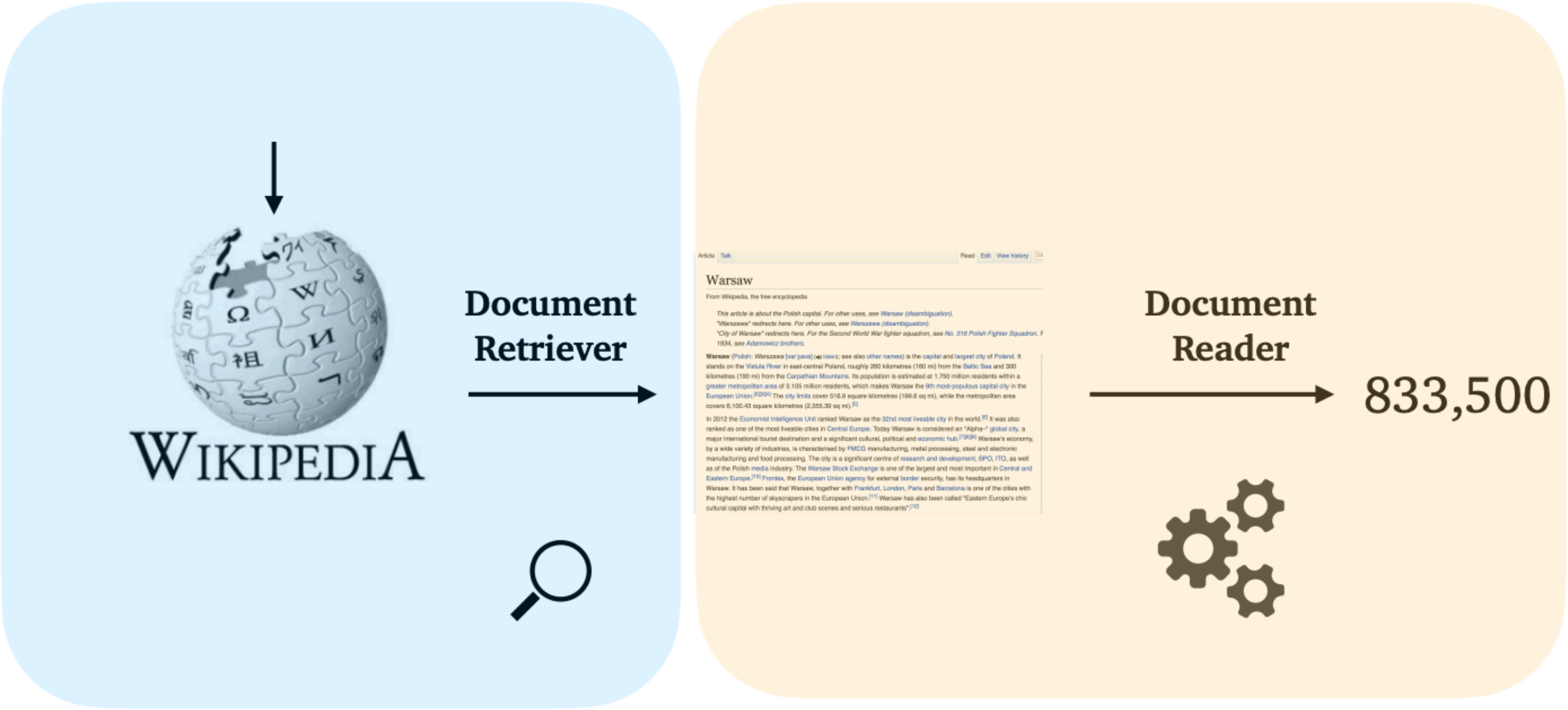
Q:	Has the UK been hit by a hurricane?
P:	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands . . .
A:	Yes. [An example event is given.]
Q:	Does France have a Prime Minister and a President?
P:	. . . The extent to which those decisions lie with the Prime Minister or President depends upon . . .
A:	Yes. [Both are mentioned, so it can be inferred both exist.]
Q:	Have the San Jose Sharks won a Stanley Cup?
P:	. . . The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 . . .
A:	No. [They were in the finals once, and lost.]

(Clark et al, 2019)

Open domain question answering

- Different from reading comprehension, we don't assume a given passage. Question (Q) Answer (A)
- 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 but a more practical problem!

Retrieve and read



https://github.com/facebookresearch/DrQA

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

DrQA: Document Retrieval

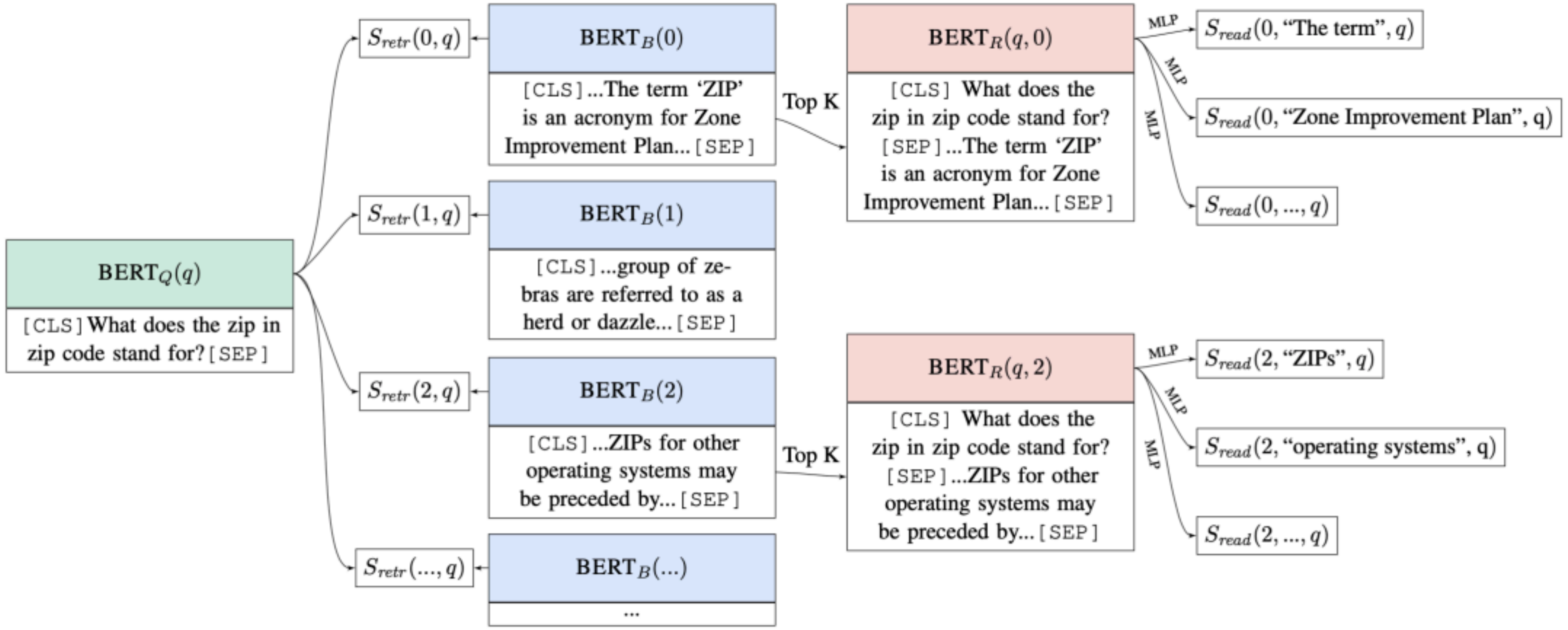
Dataset	Wiki Search	Doc. Retriever plain	Doc. Retriever +bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

Traditional tf.idf inverted index + efficient bigram hash

For **70–86%** of questions, the answer segment appears in the top 5 articles

Slide credit: Chris Manning

Joint training of retriever and reader

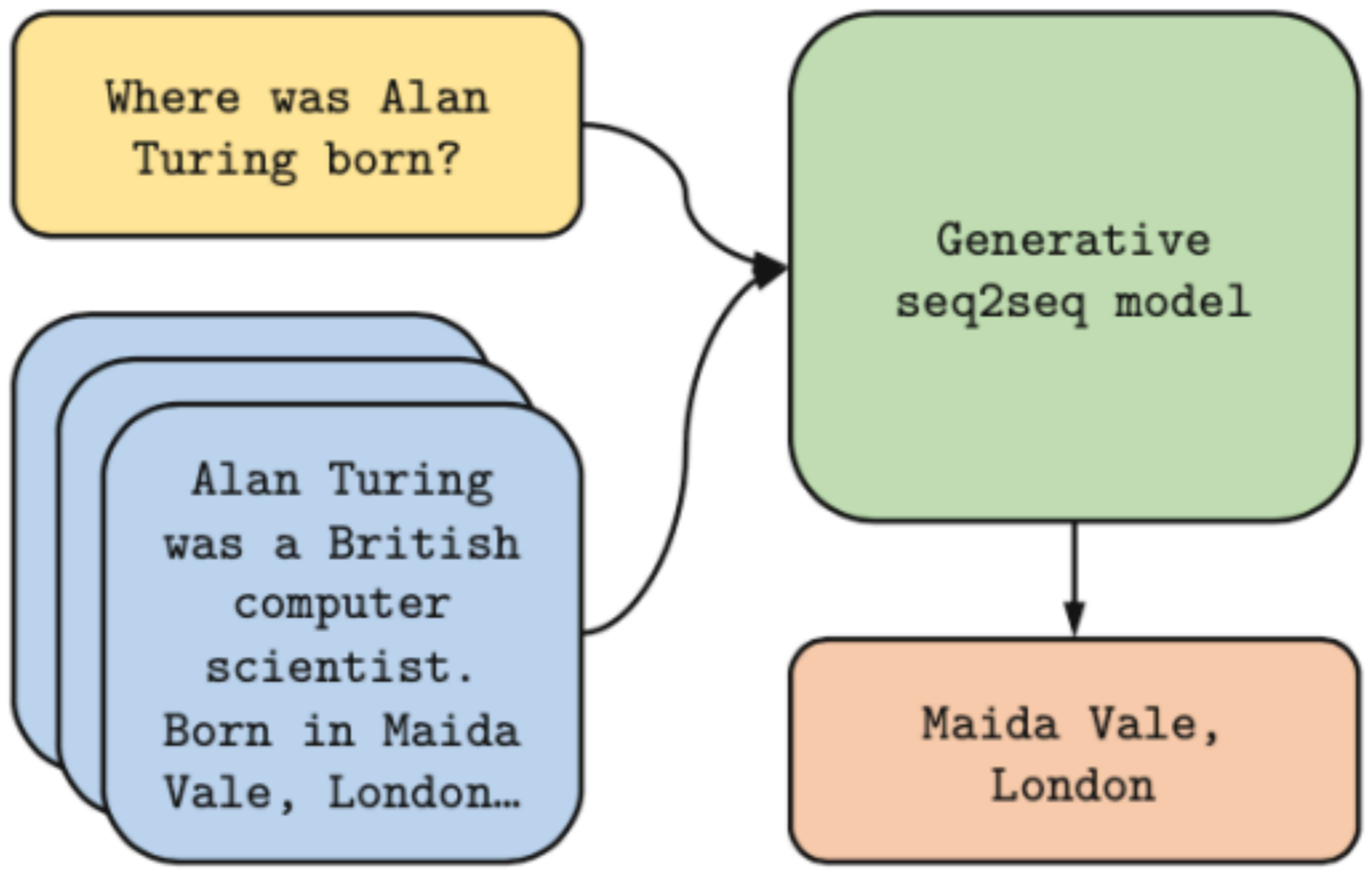


- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.
- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

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

Dense retrieval + generate answers

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

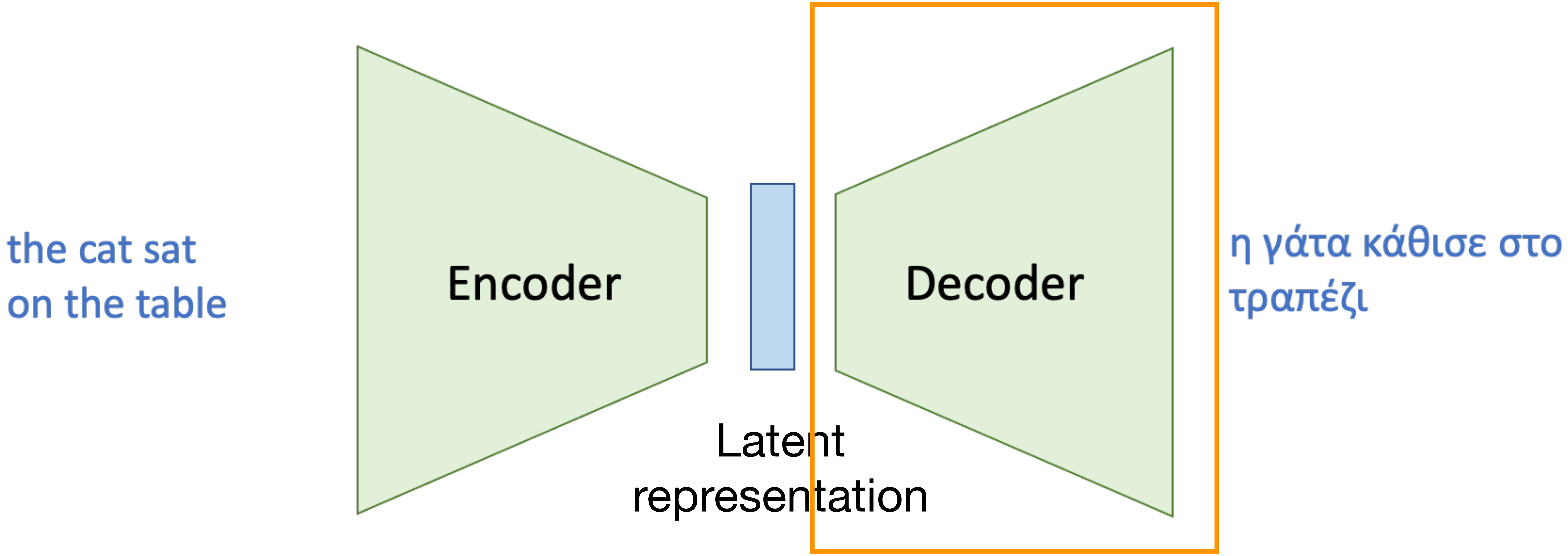


Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	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

Text generation

Encoder-Decoder Model



Understanding what is said
(encoding, parsing, feature extraction)

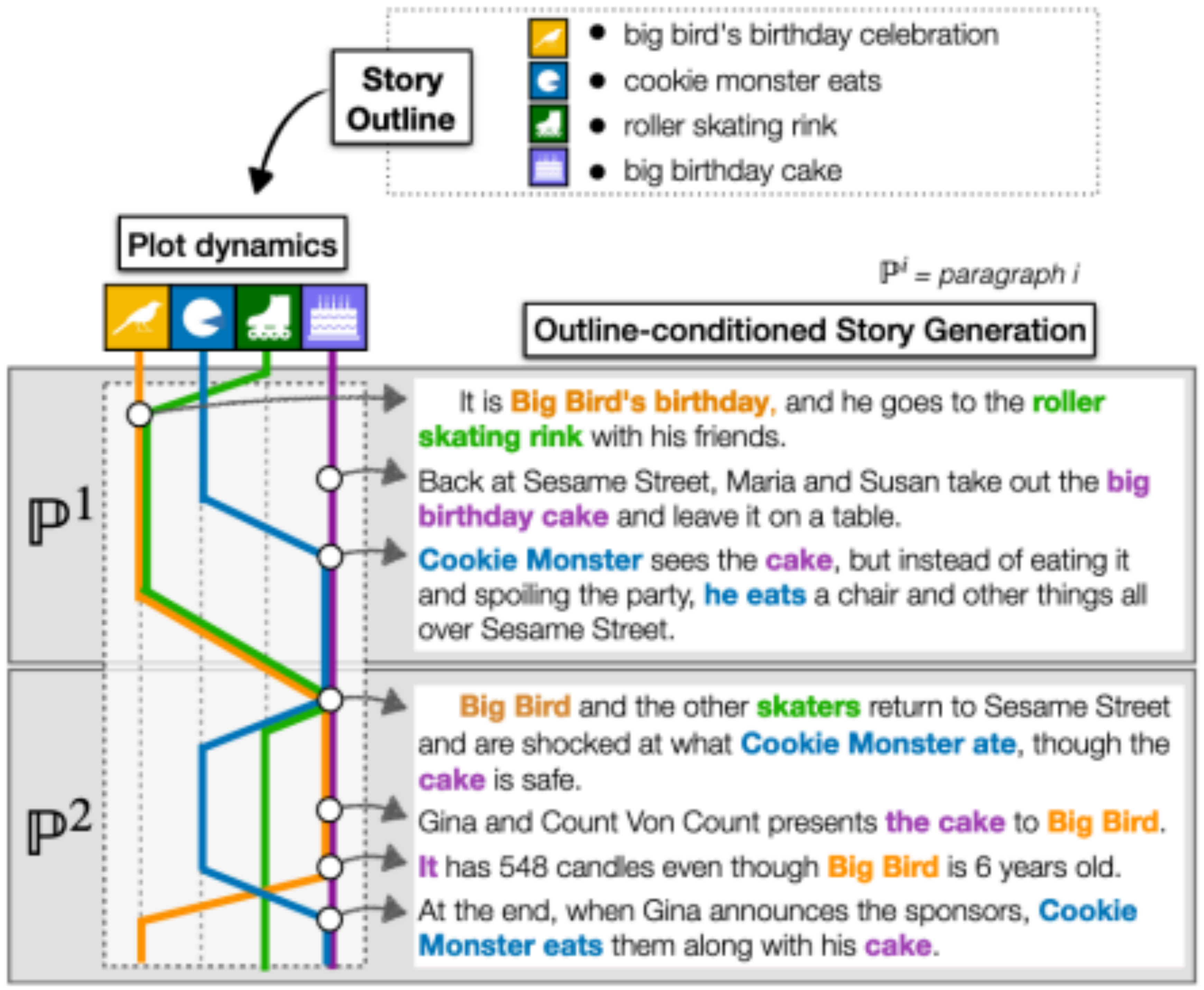
Deciding what to say
(decoding, generating)

Many tasks and applications for natural language generation (NLG)

Task/Application	Input	Output
Machine Translation	French	English
Summarization	Document	Short Summary
Dialogue	Utterance Dialog history	Response
Image Captioning	Image	Caption
Story Generation	Prompt	Story

Examples of NLG

Creative stories



(Rashkin et al., EMNLP 2020)

Data-to-text

Table Title: Robert Craig (American football)
Section Title: National Football League statistics
Table Description:None

		RUSHING					RECEIVING				
YEAR	TEAM	ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1050	4.9	62	9	92	1016	11	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	**-**	**1991**	**8189**	**4.1**	**71**	**56**	**566**	**4911**	**8.7**	**73**	**17**

Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

(Parikh et al., EMNLP 2020)

Visual description

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

(Krause et al. CVPR 2017)

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

ChatGPT: a SoTA NLG system

Idea generation

XL Got any creative ideas for a 10 year old's birthday?

Here are some creative birthday ideas for a 10 year old:

1. Scavenger Hunt: A birthday scavenger hunt can be a fun and interactive activity for kids. You can create a series of clues that lead the birthday child to the next one, until they reach the final prize.
2. DIY Crafts: Crafts are always a hit with kids. You can arrange a crafting session where the birthday child and their friends can make their own personalized gifts or decorations.
3. Movie Night: Have a movie night with friends and family at home or at a local theater. You can also make it a sleepover by setting up sleeping bags, pillows, and blankets.

Poem generation

XL Write me a poem about sorting algorithms

Sorting algorithms, oh so wise,
Bringing order to chaotic ties.
Bubble sort, quick sort, merge sort too,
Each one unique, with their own special view.

Bubble sort, gentle as can be,
Compares and swaps, so efficiently.
It moves like waves, across the sea,
Bringing order, so beautifully.

Quick sort, so fast and strong,
Divides and conquers all day long.
With each partition, it grows so wide,
Order is found, on every side.

Merge sort, so neat and trim,
Divides and merges, one by one.
It builds a bridge, to conquer all,
Order is its overall call.

Slide adapted from Stanford CS224N (Xiang Lisa Li)

Spectrum of NLG tasks

Machine Translation

Source Sentence: 当局已经宣布今天是节假日。

Reference Translations:

1. Authorities have announced a national holiday today.
2. Authorities have announced that today is a national holiday.
3. Today is a national holiday, announced by the authorities.

Output space not very diverse

Spectrum of NLG tasks

ChitChat Dialog

Input: Hey, how are you?

Outputs:

1. Good! You?
2. I just heard an exciting news, do you want to hear it?
3. Thx for asking! Barely surviving my hws.

More possible “correct” generations

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Spectrum of NLG tasks

Story Generation

Input: Write a story about three little pigs?

Outputs:
… (lots of different options!)…

Very open-ended!

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Spectrum of NLG tasks

Less open-ended — More open-ended

Less diverse — More diverse

Output is mostly determined by the input

Lots of freedom in the output, output distribution should be varied and diverse

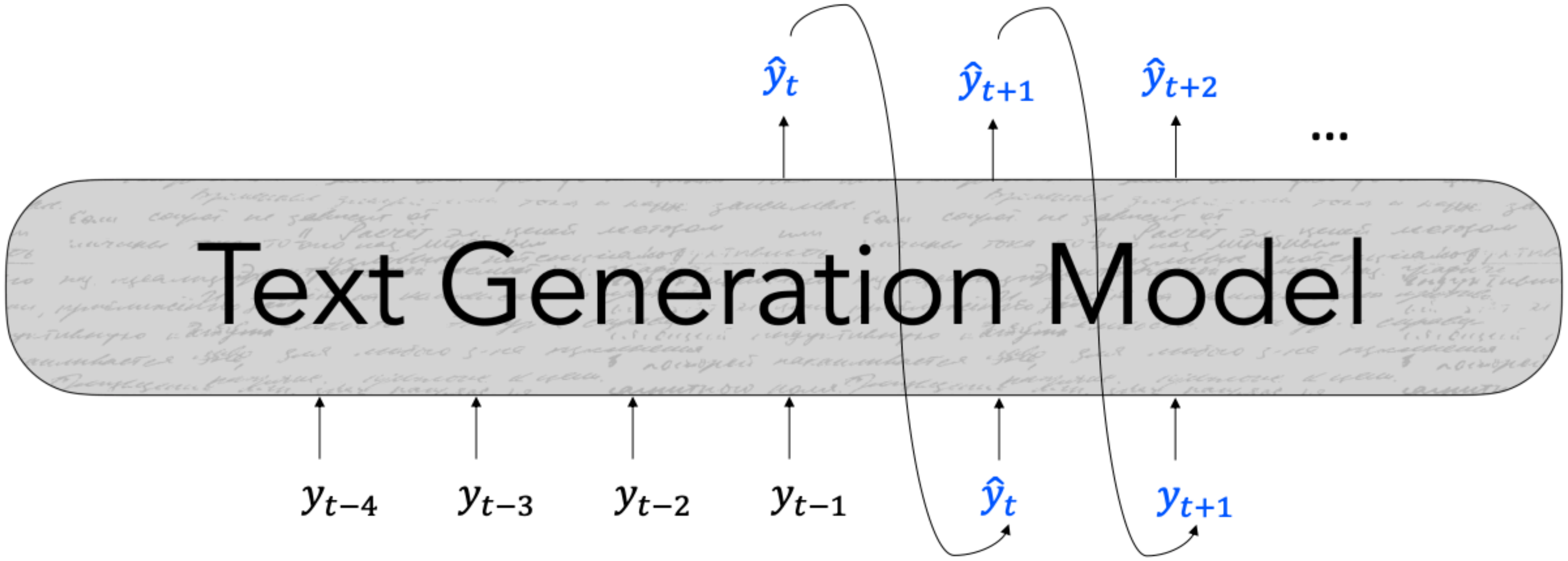
Can characterize the spectrum of tasks using entropy.
Can use different decoding and training strategies for each.

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Review of autoregressive text generation

- Autoregressive text models generate future words based on past words
- At each time step t, the model is given sequence of tokens as input $\{y\}_{<t}$ and predicts next token $\hat{y}_t$
- For model $f(\,.\,)$ and vocabulary V, the model estimate the probability of the next token by taking the softmax of the scores: $S = f(\{y_{<t}, \theta\} \in \mathbb{R}^V$

$$P_t(y_t = w|\{y_{<t}\}) = \frac{\exp(S_w)}{\sum_{w' \in V} \exp(S_{w'})}$$



Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Causal LMs: Common Pitfalls

- **Generated output is too short/long**: LM may require further tuning, also asking for more tokens can help
- **Incorrect generation mode**: greedy decoding or sampling? Which is better depends on your task
- **Wrong padding side**: you may need to pad the prompt text on the left to ensure that the input is the same size as the training phase of the LM.
- **Wrong prompt**: this is tricky and has produced a whole industry of "prompt engineering"

https://huggingface.co/docs/transformers/llm_tutorial

c.f. for code samples

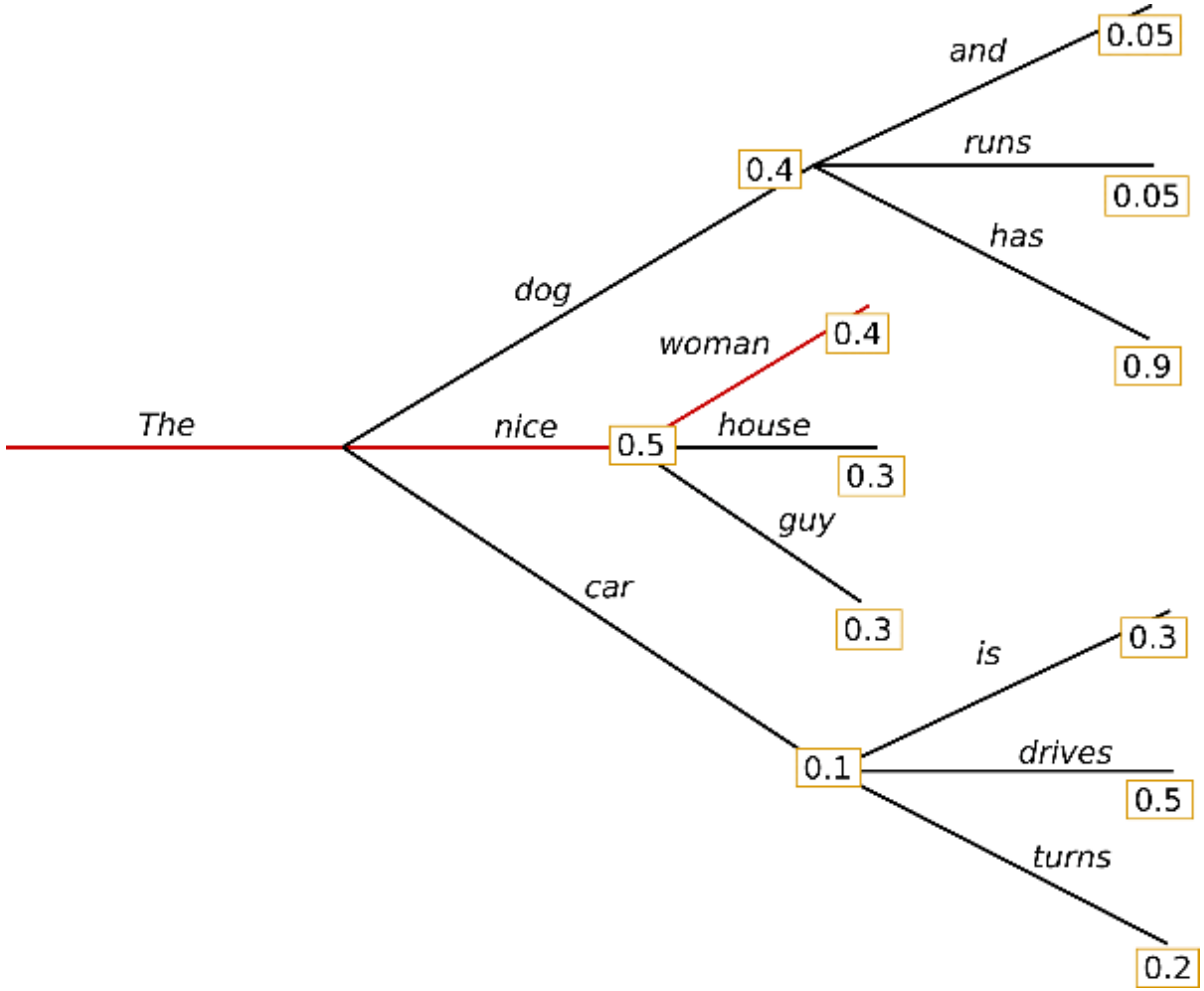
Decoding methods

https://huggingface.co/blog/how-to-generate

$$P(w_{1:T}|W_0) = \prod_{t=1}^{T} P(w_t|w_{1:t-1}, W_0) \text{ ,with } w_{1:0} = \emptyset,$$

- W_0 is the initial context word sequence (aka the "prompt")
- The length T of the word sequence is determined on-the-fly
- T is determined by the generation of the end-of-sentence EOS also known as the <|endoftext|> token
- The EOS token is produced like the other tokens from $P(w_t \mid w_{1:t-1}, W_0)$

Greedy Decoding



("The","nice","woman")
having an overall probability of
$0.5 \times 0.4 = 0.2$

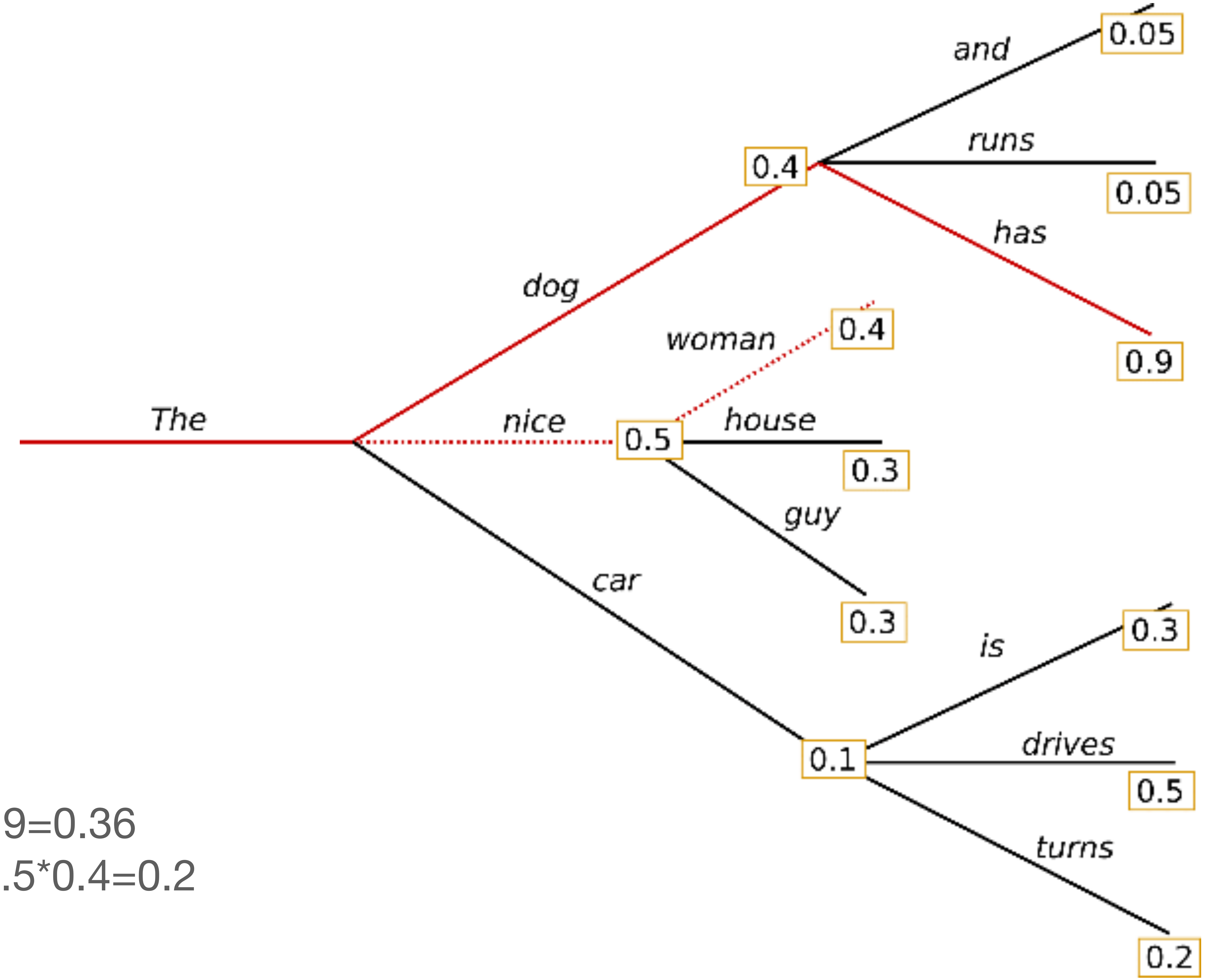
Beam Search

Let us assume a beam size of 2

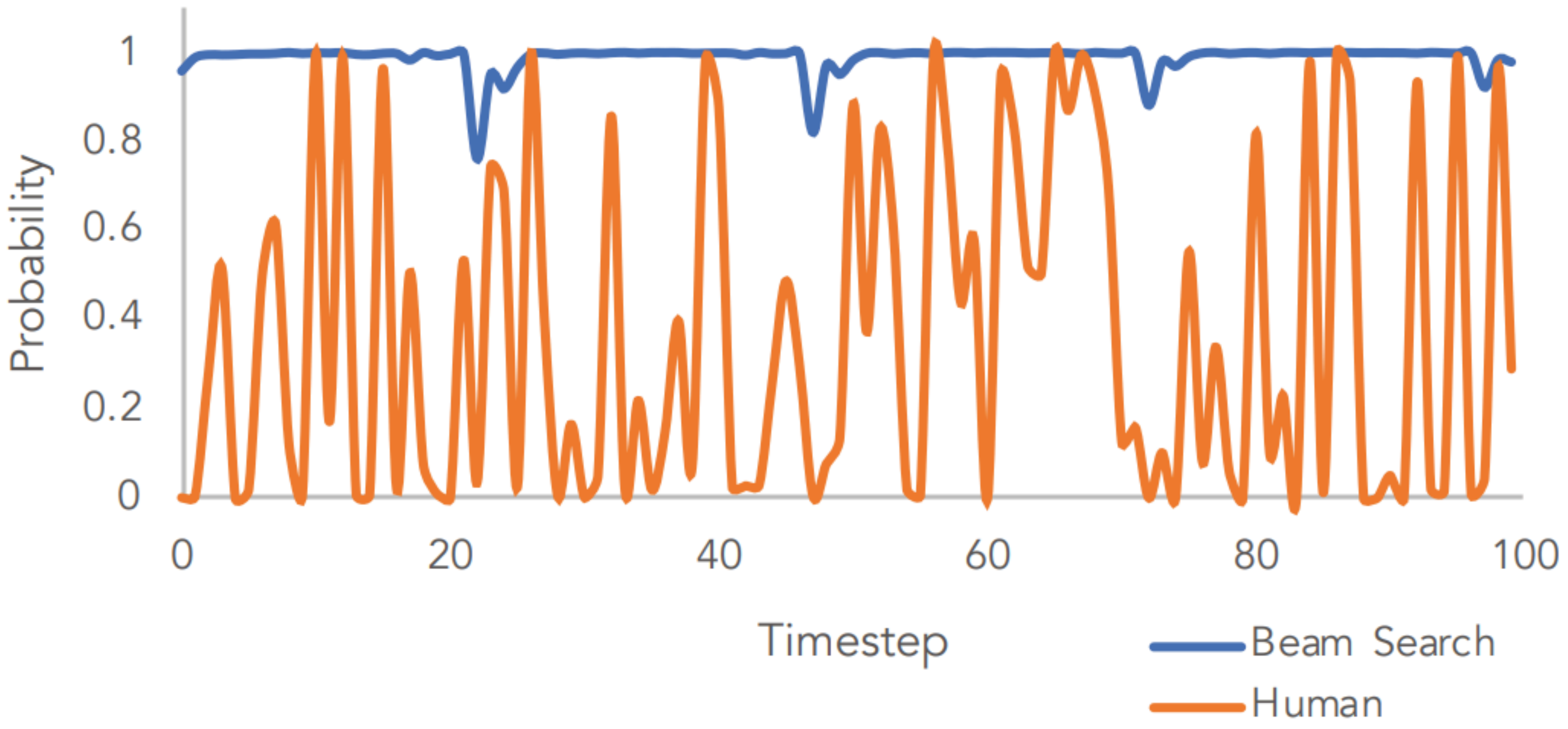
Keep the 2 best outcomes at each time step

In this example:
("The", "nice") 0.5
("The", "dog") 0.4

Next time step:
("The", "dog", "has") 0.5*0.9=0.36
("The", "nice", "woman") 0.5*0.4=0.2



Human generation has lots of diversity!



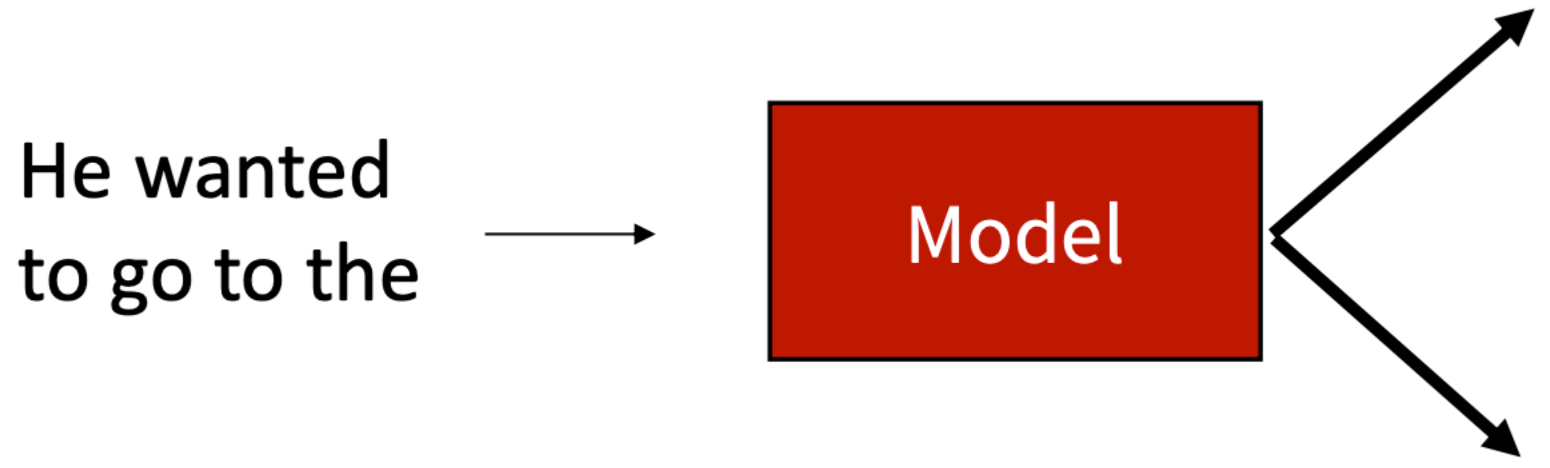
The Curious Case of Neural Text Degeneration
https://openreview.net/pdf?id=rygGQyrFvH
[Holtzman et al, ICLR 2020]

Different ways to sample during decoding

- Basic/vanilla sampling over entire distribution
- Top-k sampling
- Top-p (nucleus) sampling
- Temperature based sampling

Issues with vanilla sampling

- Sample from entire probability distribution

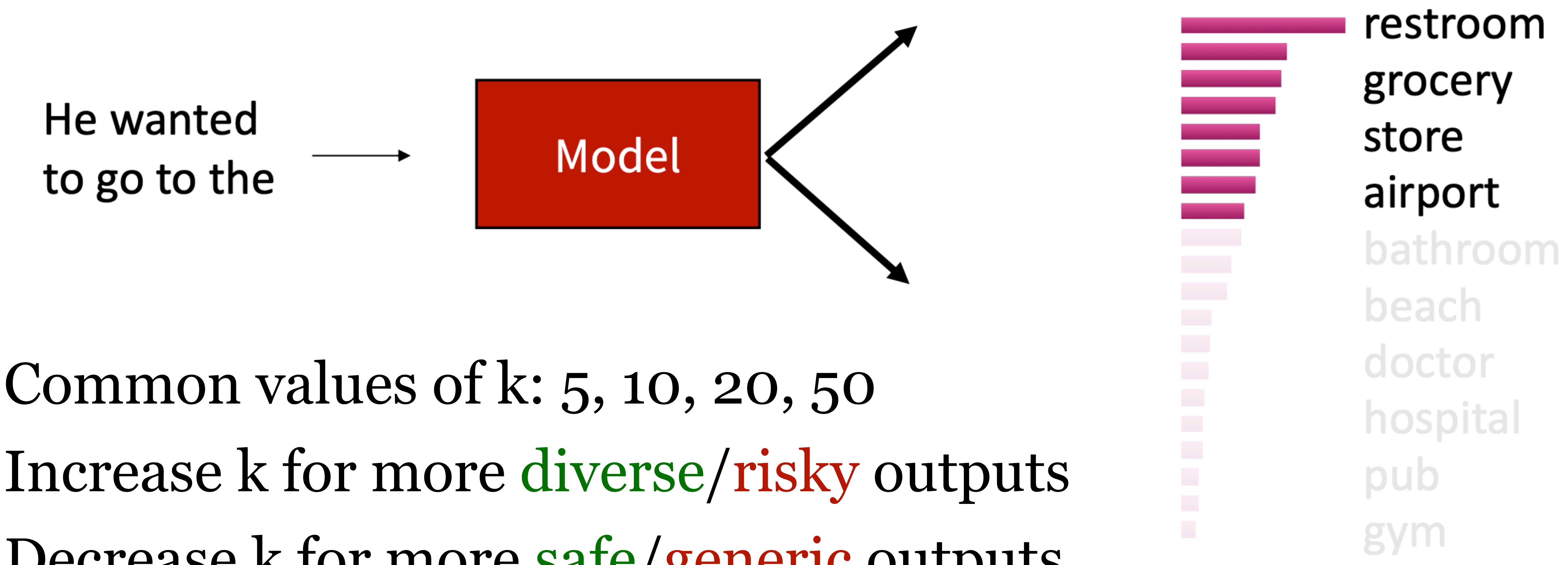


- Long tail could have enough mass unlikely words are still selected

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Decoding: Top-k sampling

- Only sample from top **k tokens** in the probability distribution



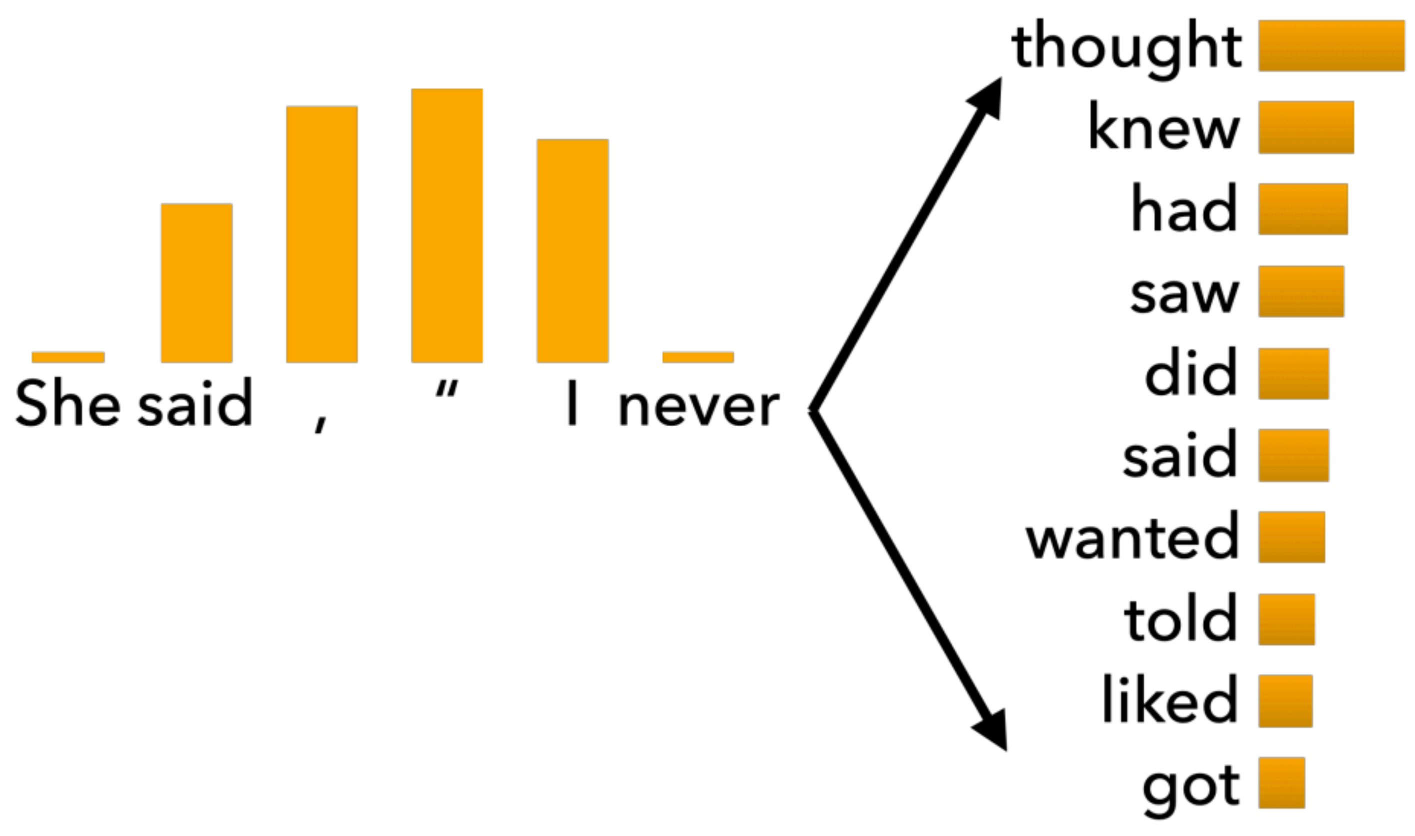
- Common values of k: 5, 10, 20, 50
- Increase k for more diverse/risky outputs
- Decrease k for more safe/generic outputs

- Greedy search: $k = 1$, Pure sampling: $k = |V|$

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

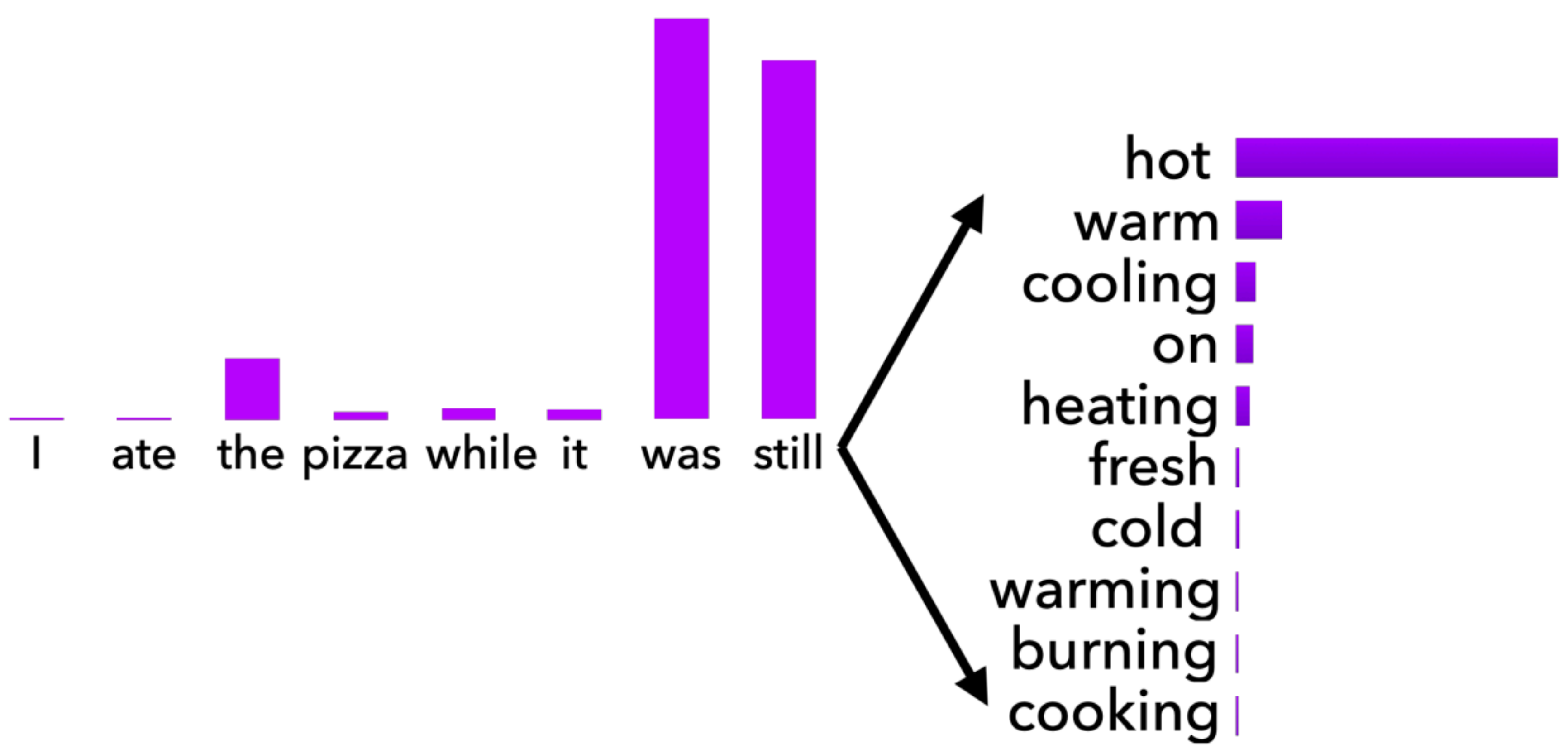
Decoding: Top-k sampling

Cuts off too slowly!



Flat distribution

Cuts off too quickly!



Peaky distribution

The Curious Case of Neural Text Degeneration
https://openreview.net/pdf?id=rygGQyrFvH
[Holtzman et al, ICLR 2020]

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Decoding: Top-p (nucleus) sampling

- Sample from all tokens in the **top p** cumulative probability mass
- This allows **k to vary** depending on the peakiness of the distribution P_t

$P_t^1(y_t = w \mid \{y\}_{<t})$ $\quad$ $P_t^2(y_t = w \mid \{y\}_{<t})$ $\quad$ $P_t^3(y_t = w \mid \{y\}_{<t})$

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Decoding: Other variants

- Typical Sampling [Meister et al. 2022]
  - Reweights the score based on the entropy of the distribution
- Epsilon Sampling [Hewitt et al. 2022]
  - Set threshold for lower bounding valid probabilities

$P_t^1(y_t = w \mid \{y\}_{<t})$ $\quad$ $P_t^2(y_t = w \mid \{y\}_{<t})$ $\quad$ $P_t^3(y_t = w \mid \{y\}_{<t})$

Improving decoding: Temperature scaled softmax

- Recall: On timestep t, the model samples from the distribution P_t which is computed by taking the softmax of a vector of scores $S \in \mathbb{R}^{|V|}$

$$P_t(y_t = w) = \frac{\exp(S_w)}{\sum_{w' \in V} \exp(S_{w'})}$$

- We can apply a temperature hyperparameter τ to the softmax to rebalance the distribution

$$P_t(y_t = w) = \frac{\exp(S_w/\tau)}{\sum_{w' \in V} \exp(S_{w'}/\tau)}$$

- Raise the temperature $\tau > 1$: P_t becomes more uniform
  - More diverse output (probability is spread around vocabulary)
- Lower the temperature $\tau < 1$: P_t becomes more spiky
  - Less diverse output (probability is concentrated on top words)

Note: temperature scaled softmax is not a decoding algorithm!
It's a decoding hyperparameter you can apply at test time, in conjunction with a decoding algorithm (such as beam search or sampling)

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Improving Decoding: Re-ranking

- Decode a bunch of sequences (say 10) and re-rank with a score that measure the quality of the sequences
- Have a separate scoring function to approximate the quality of the sequences
  - Simplest is to use low perplexity
    - But repetitive sequences can have low perplexity…
  - Re-rankers can score a variety of properties
    - style, discourse, logical consistency, factuality, etc
  - Can combine these different rankers (but beware of poorly-calibrated re-rankers)

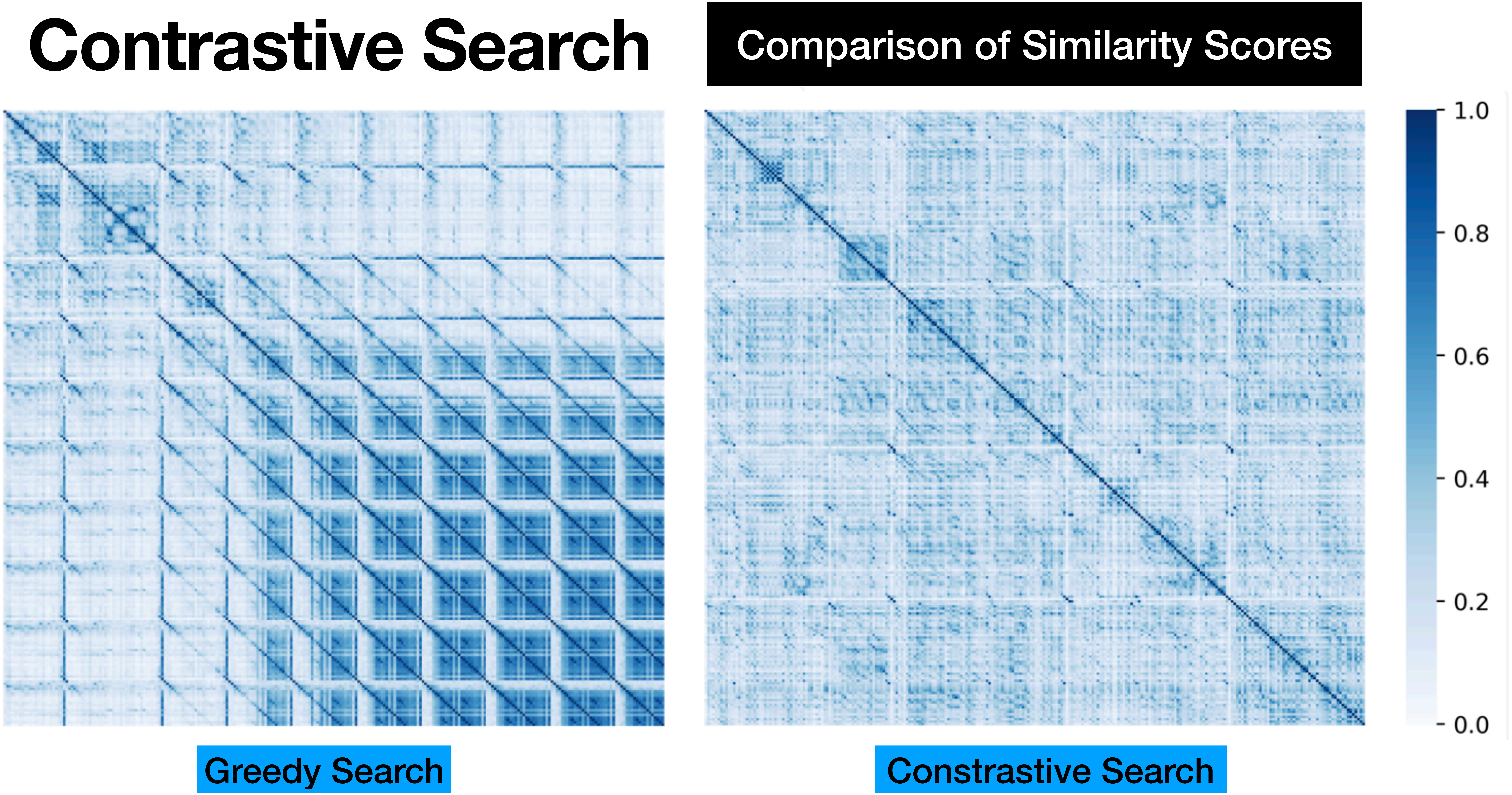
Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Contrastive Search

- Given a prefix text $\mathbf{x}_{<t}$ select the output next token x_t
- $V^{(k)}$ is the set of top-k predictions from the LM's probability distribution $p_\theta(v \mid \mathbf{x}_{<t})$ called the **model confidence**
- $\mathbf{s}(\cdot, \cdot)$ is the cosine similarity between two token representations is used to compute the **degeneration penalty**
- The more similar v is to the context the more we see **model degeneration.**
- Combine the two terms using a linear mixture.

$$x_t = \underset{v \in V^{(k)}}{\arg\max} \left\{ (1-\alpha) \times \underbrace{p_\theta(v|\boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha \times \underbrace{(\max\{s(h_v, h_{x_j}) : 1 \le j \le t-1\})}_{\text{degeneration penalty}} \right\},$$

Contrastive Search



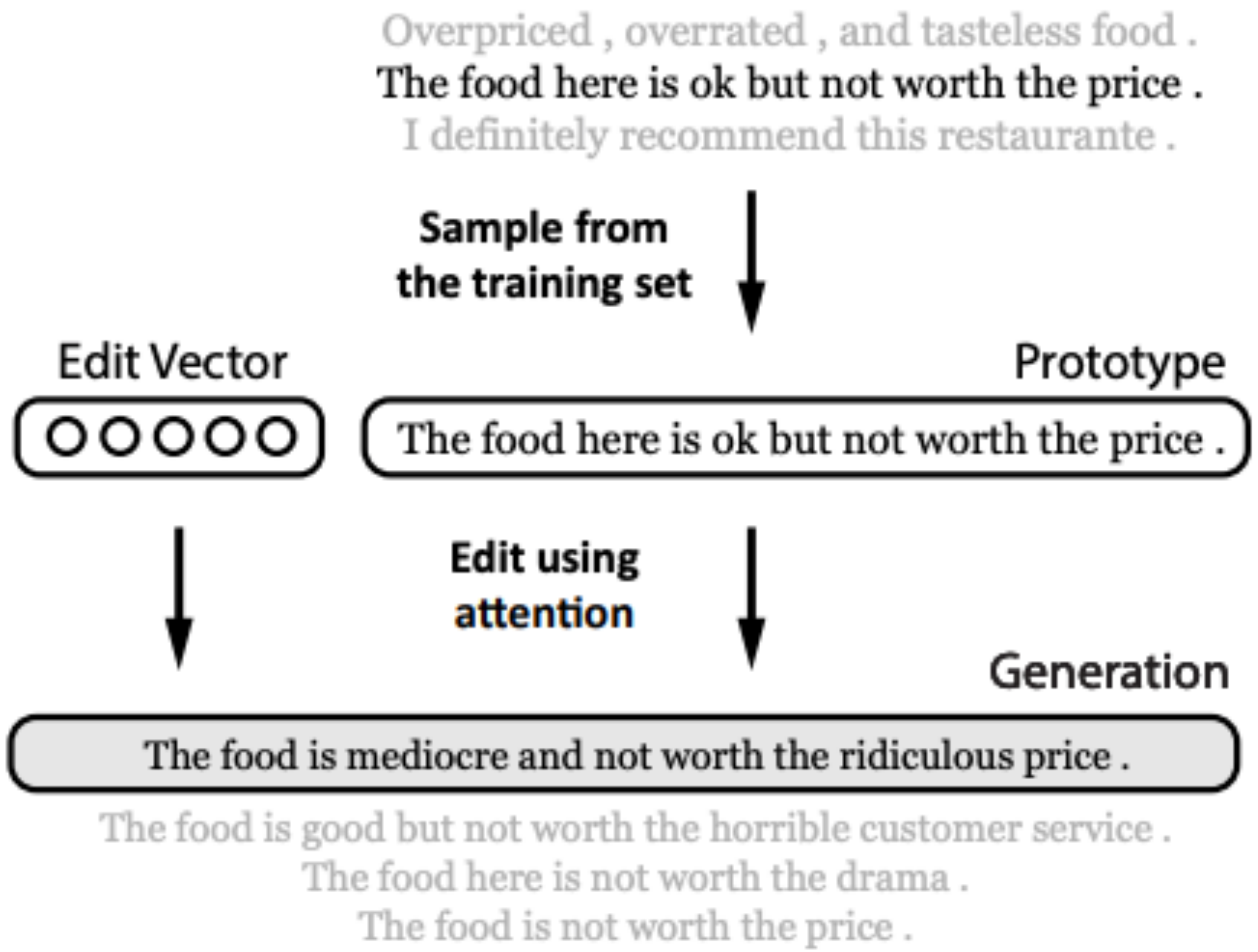
Other problems

- **Unreachable subword problem**: there are some subwords for which under no circumstances is it possible to produce a subword (given any context).
- **Mode collapse**: tuning the LM might cause the model parameters to reach a state where Greedy and Sampling based generation produce the same output.
- **Softmax over very large vocabulary sizes**: Vocabulary sizes have reduced since subword segmentation has become the standard way to set up the vocabulary for LMs; However for very large vocabulary sizes, the compute efficiency for softmax might need careful consideration, e.g. use hierarchical softmax.

Alternatives to autoregressive generation

Retrieve and Edit

- Retrieve prototype sentence x' from a corpus
- Sample edit vector z (encodes type of edit to be perform).
- Use neural editor to combine edit vector z and prototype sentence x' to get new sentence x.



Generating Sentences by Editing Prototypes
https://arxiv.org/pdf/1709.08878.pdf
Guu et al, TACL 2020]

Non-autoregressive generation (with transformers)

- Can generate words in a non-autogressive manner
- Relies on the idea of masked language model
- Predict length of output
- Iterative refinements / masking
  - Predict length of output
  - Predict all words $P(y_i | x)$ $\hat{y}_t^0 = \arg\max_{y_t} \log p(y_t^0 | X)$
  - Iteratively refine sequence of predictions based on input and previous predictions

$$\hat{y}_t^l = \arg\max_{y_t} \log p(y_t^l | \hat{Y}^{l-1}, X)$$

- Efficient decoding since parts of the decoding can run in parallel

Each iteration, can just mask out low-confidence words

Mask-Predict: Parallel Decoding of Conditional Masked Language Models
https://arxiv.org/pdf/1904.09324.pdf
[Ghazvininejad et al, EMNLP 2019]

Evaluation

Content Overlap Metrics

Model-based Metrics

Human Evaluations

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Automatic evaluation metrics

Content overlap metrics:

- Word (n-gram) overlap: BLEU, ROUGE, METEOR, CIDEr
- Structured overlap: PYRAMID, SPICE, SPIDER

Model based metrics:

- Embedding similarity: Embedding average, Word Mover Distance, BERTSCORE, etc.
- Metric predictor: BLEURT

Content overlap metrics

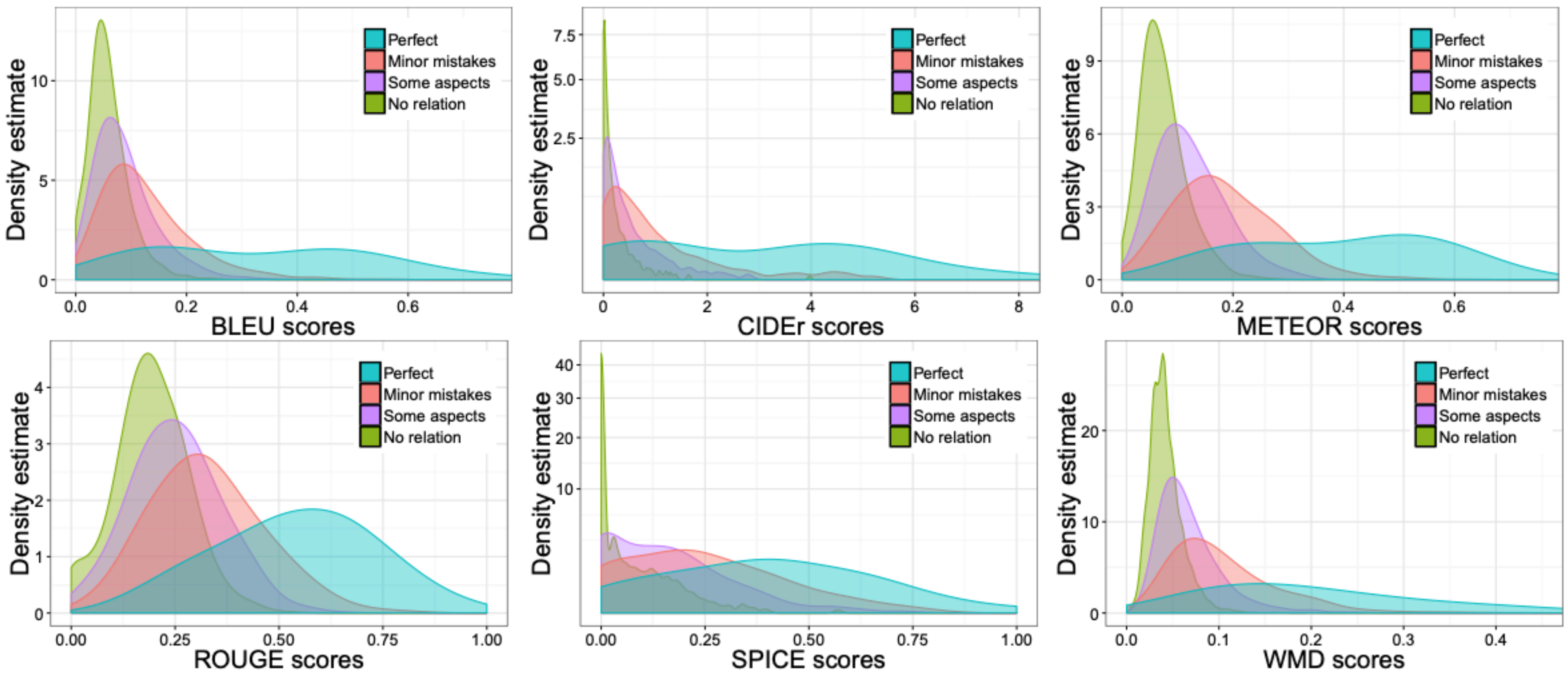
Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .

- Compute a score that indicates the similarity between *generated* and *gold-standard (human-written) text*
- Fast and efficient and widely used
- Two broad categories:
  - *N*-gram overlap metrics (e.g., **BLEU**, ROUGE, METEOR, CIDEr, etc.)
  - Semantic overlap metrics (e.g., PYRAMID, SPICE, SPIDEr, etc.)

Slide credit Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

	FLICKR-8K			COMPOSITE		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
WMD	0.68	0.60	0.48	**0.43**	0.43	0.32
SPICE	**0.69**	**0.64**	**0.56**	0.40	0.42	**0.34**
CIDEr	0.60	0.56	0.45	0.32	0.42	0.32
METEOR	**0.69**	0.58	0.47	0.37	**0.44**	0.33
BLEU	0.59	0.44	0.35	0.34	0.38	0.28
ROUGE	0.57	0.44	0.35	0.40	0.39	0.29



Re-evaluating Automatic Metrics for Image Captioning
[Kilickaya et al, EACL 2017]

N-gram overlaps are not good metrics

Word overlap-based metrics: BLEU, ROUGE, METEOR, CIDEr, etc

- Not ideal for machine translation
- But they get even progressively worse for tasks that are more open-ended than machine translation
  - Worse for summarization, as longer output texts are harder to measure
  - Much worse for dialogue, which is more open-ended that summarization
  - Much, much worse story generation, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

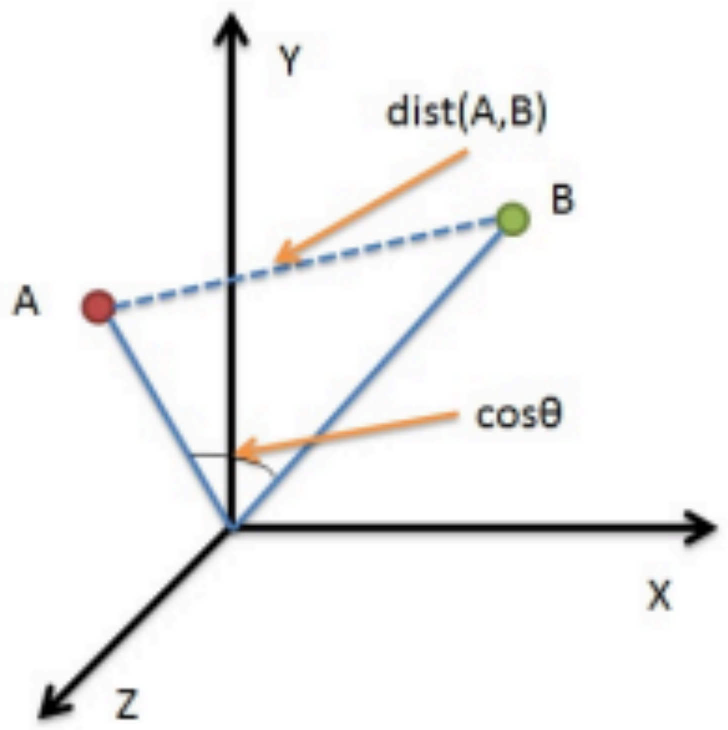
Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Model-based metrics

- Use learned representations of words and sentences to compute semantic similarity between generated and reference texts
- No more n-gram bottleneck because text units are represented as embeddings!
- Even though embeddings are pretrained, distance metrics used to measure the similarity can be fixed

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

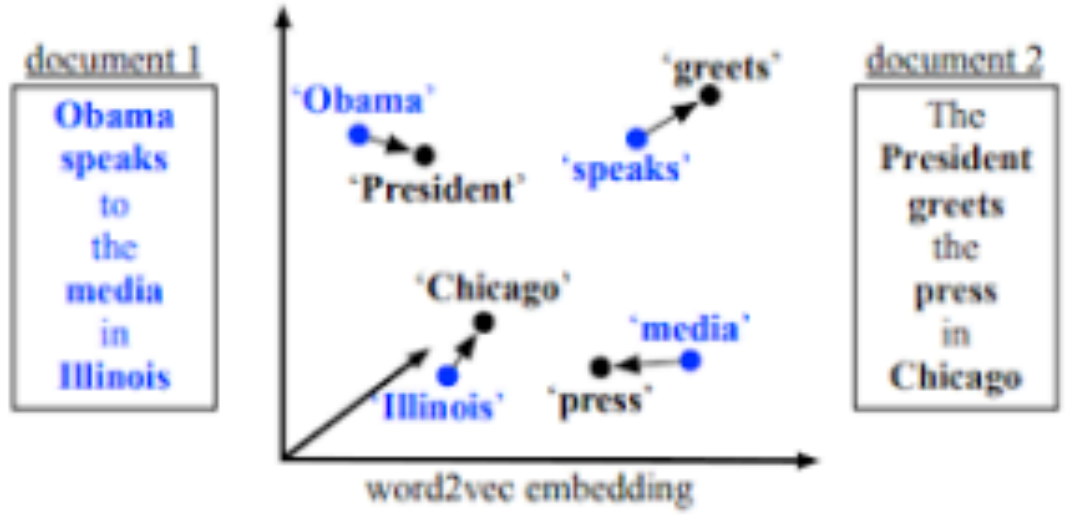
Model-based metrics: Word distance functions



Vector Similarity:

Embedding based similarity for semantic distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Liu et al., 2016)
- **MEANT** (Lo, 2017)
- **YISI** (Lo, 2019)

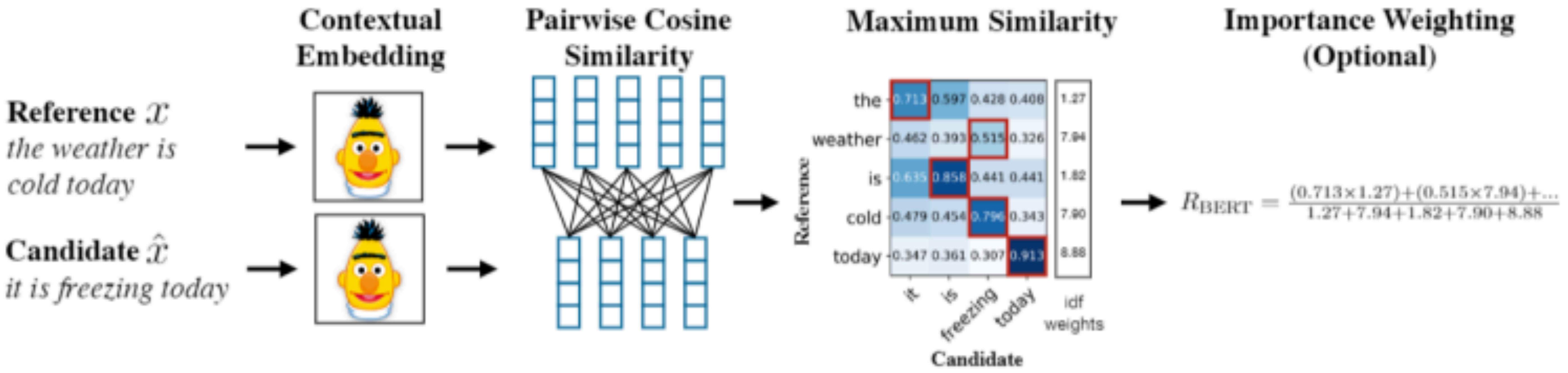


Word Mover's Distance:

Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.
(Kusner et.al., 2015; Zhao et al., 2019)

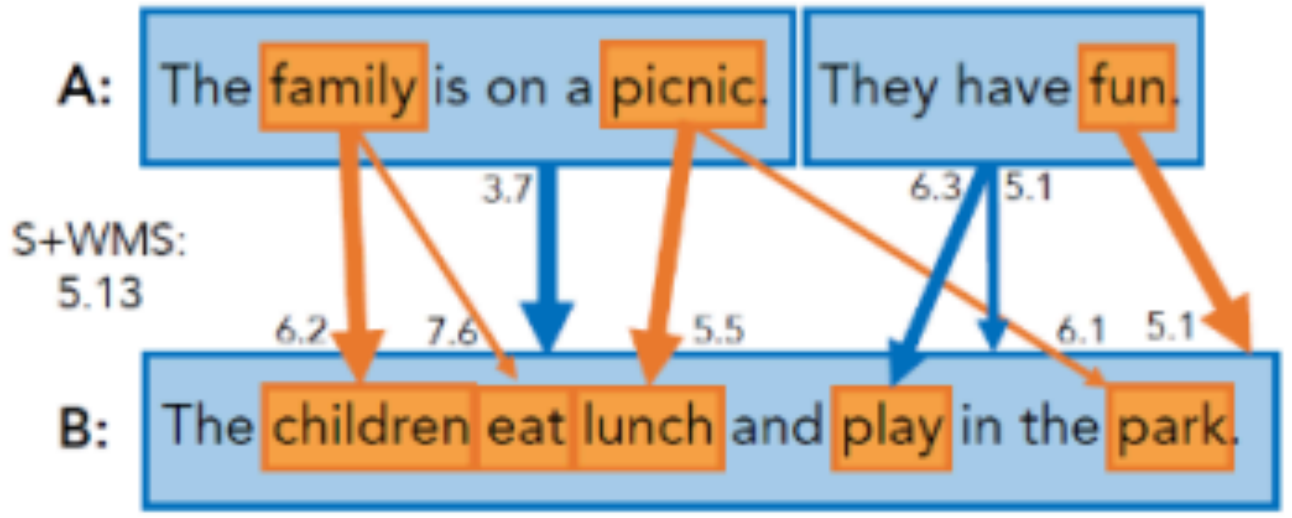
BERTSCORE:

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
(Zhang et.al. 2020)



Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Model-based metrics: Beyond word matching



Sentence Movers Similarity :

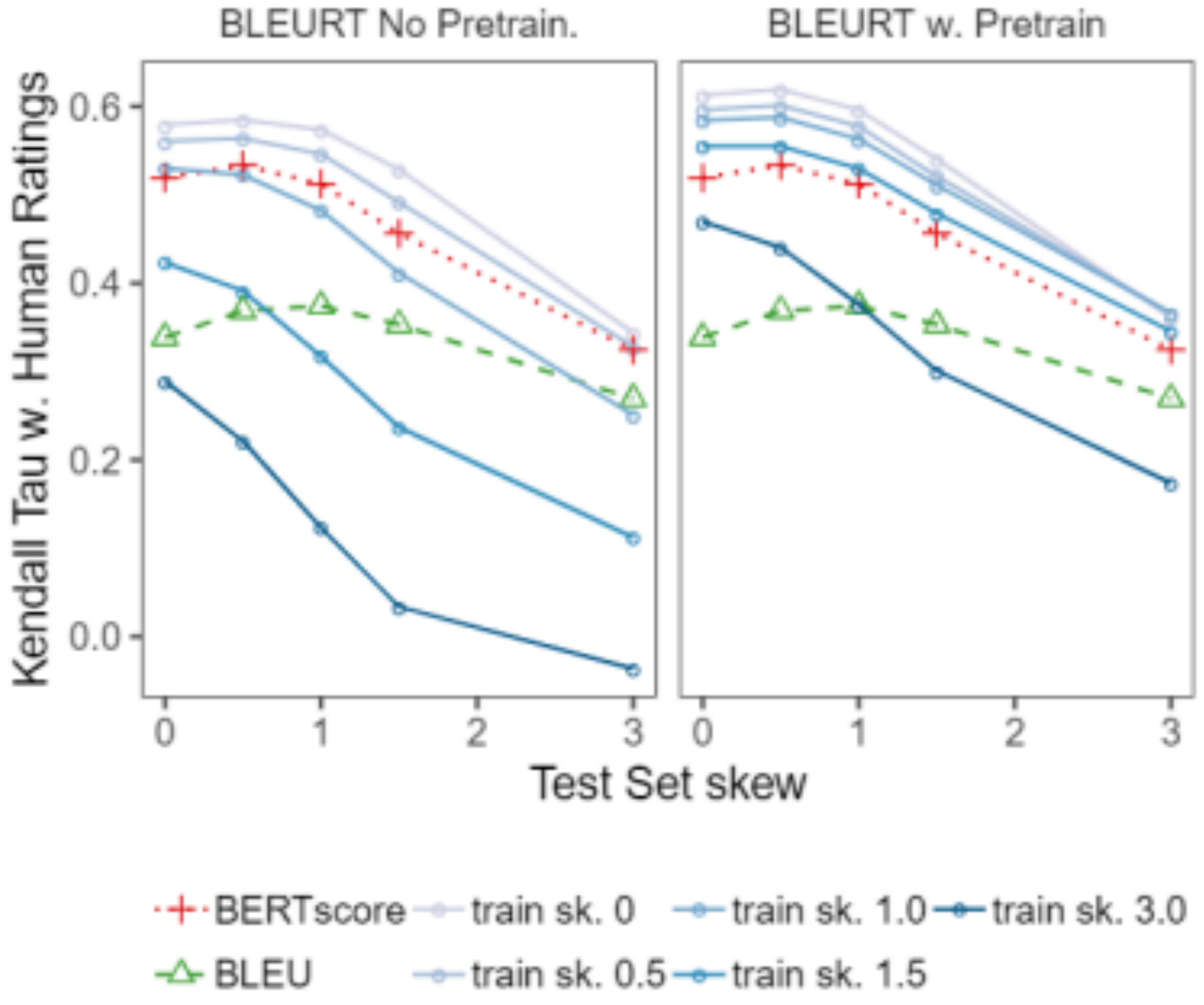
Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations.

(Clark et.al., 2019)

BLEURT:

A regression model based on BERT returns a score that indicates to what extend the candidate text is grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)

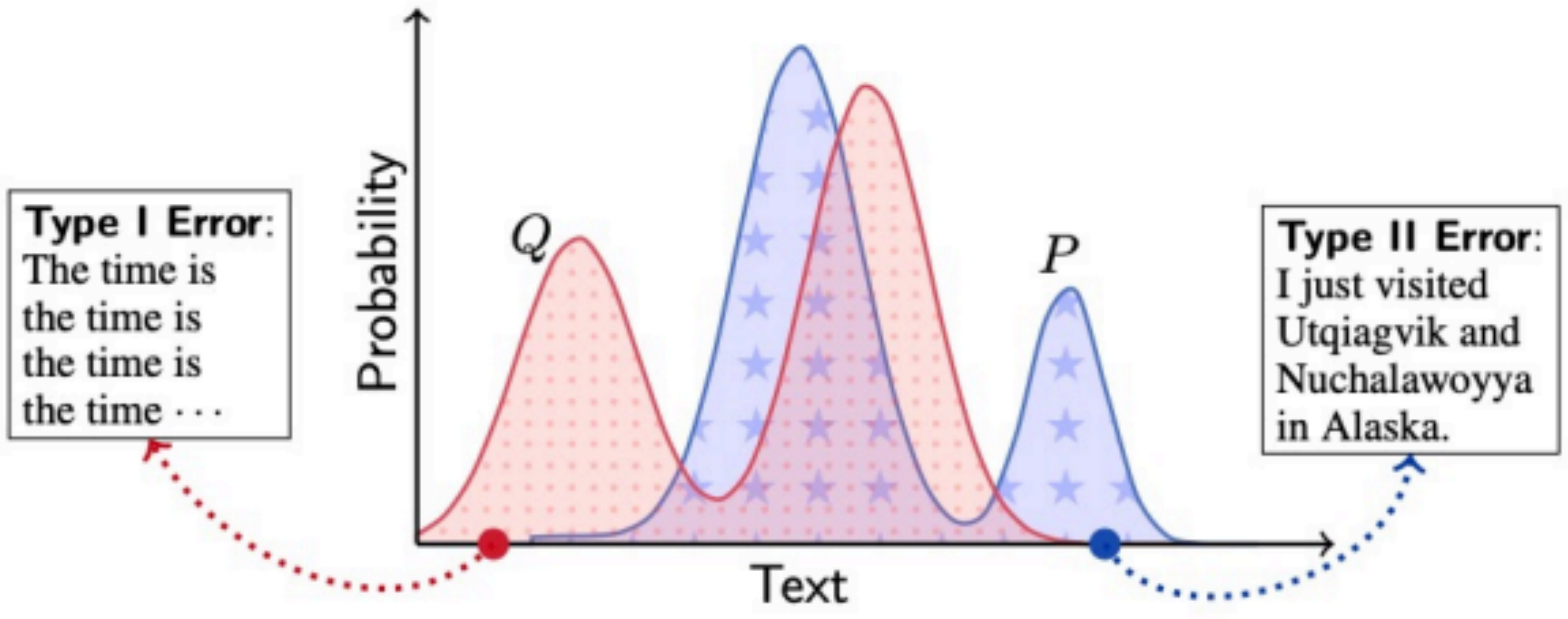


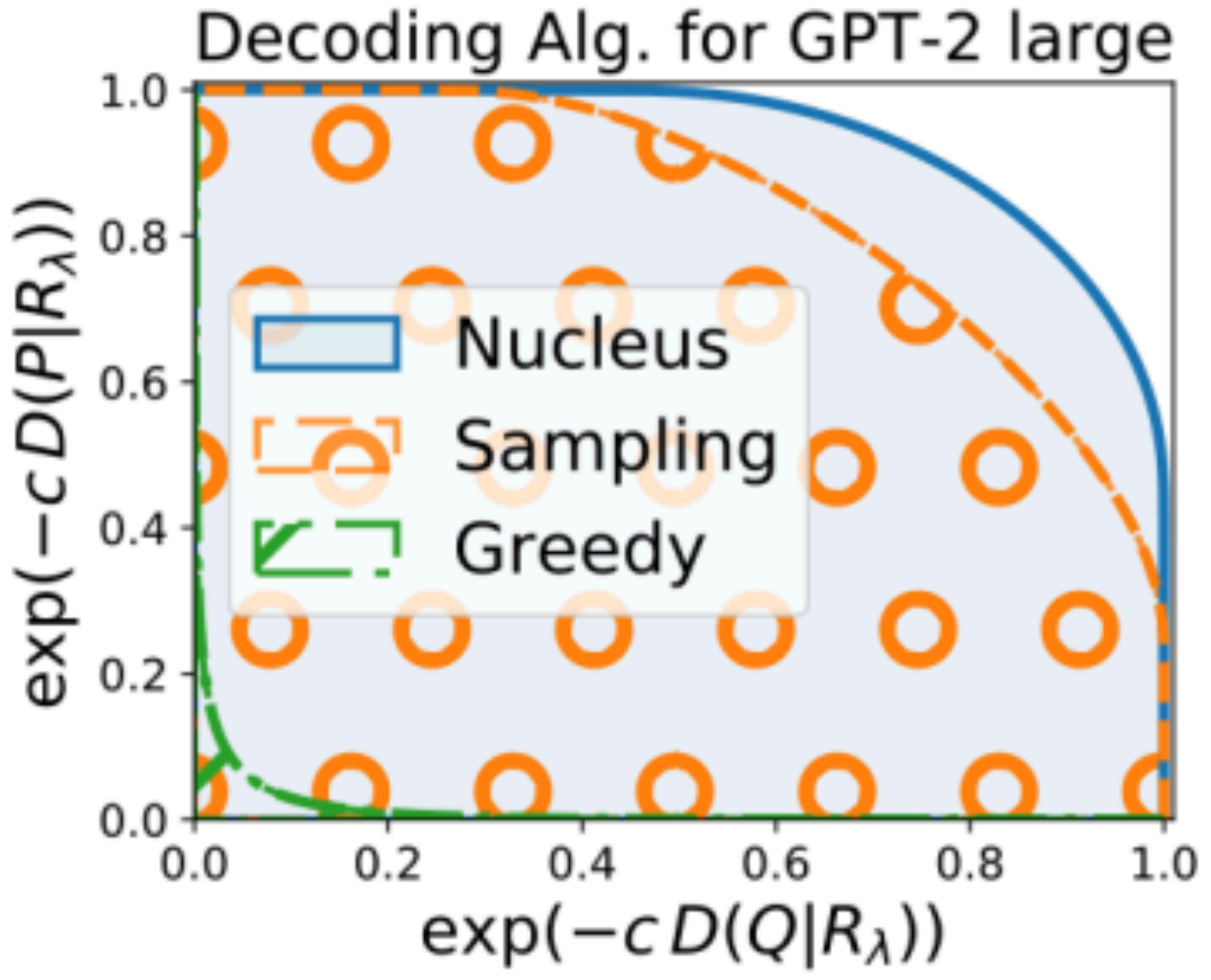
Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Evaluating open-ended text generation

MAUVE

MAUVE computes information divergence in a quantized embedding space, between the generated text and the gold reference text (Pillutla et.al., 2022).



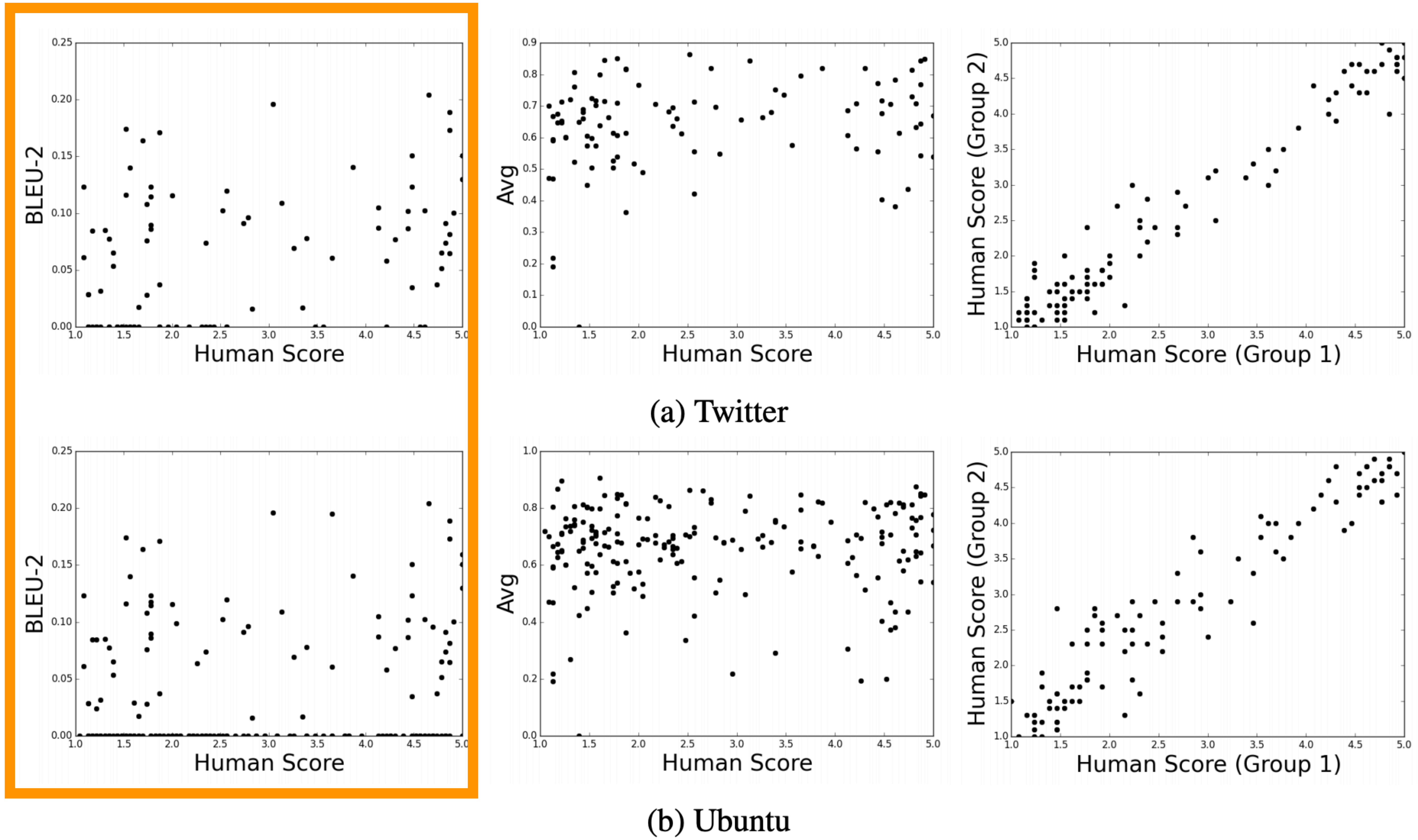


Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Issues with Automatic Evaluation

Automatic Evaluation:
Word overlap metrics are bad for dialogue

No correlation between **human judgement** and **BLEU**



(a) Twitter

(b) Ubuntu

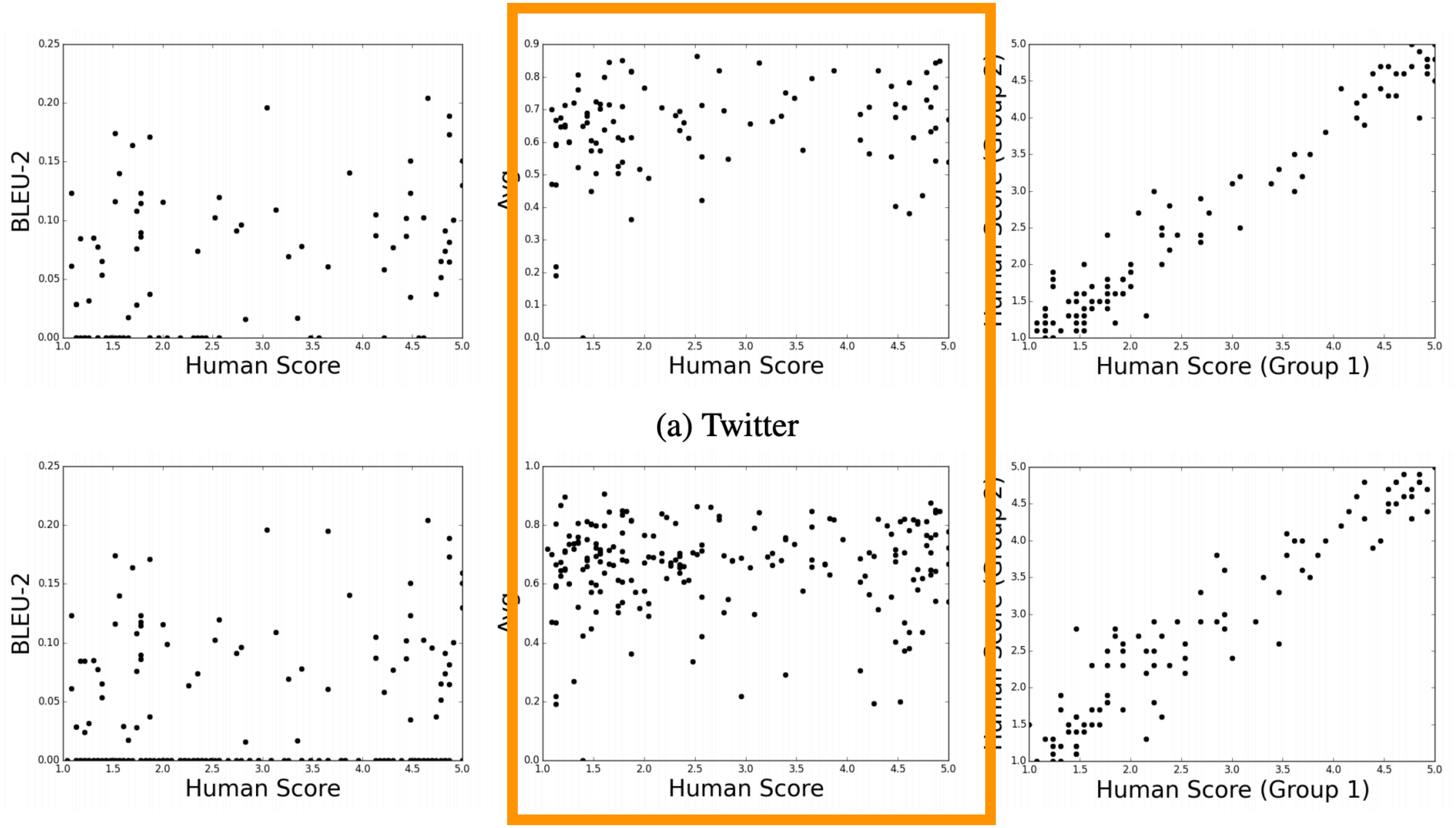
BLEU **Embedding Average** **Human**

[How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, Liu et al 2017, https://arxiv.org/pdf/1603.08023.pdf]

Issues with Automatic Evaluation

Automatic Evaluation:
Embedding metrics are also poor for dialogue

No correlation between **human judgement** and **embedding average**



(a) Twitter

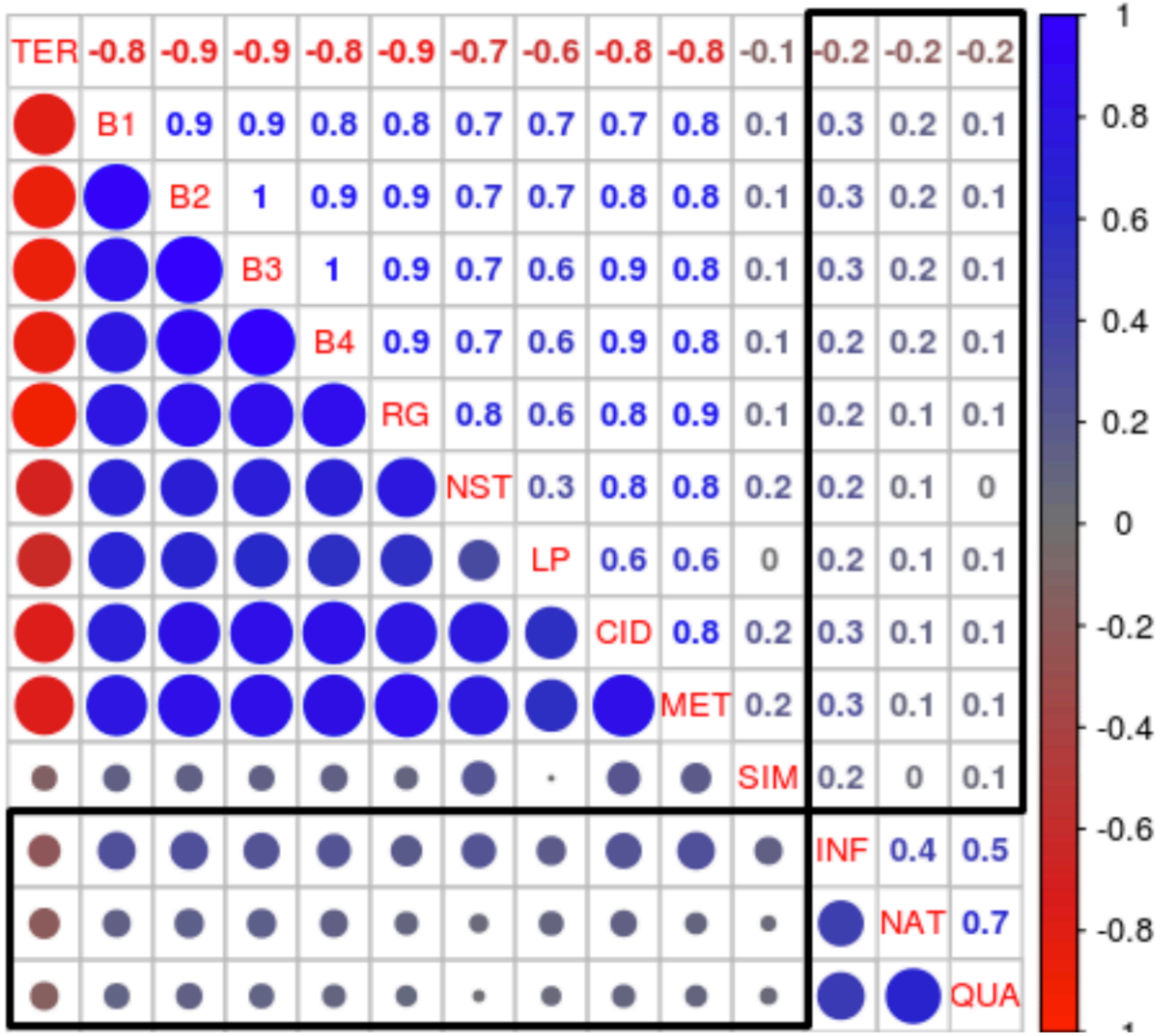
(b) Ubuntu

BLEU **Embedding Average** **Human**

[How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, Liu et al 2017, https://arxiv.org/pdf/1603.08023.pdf]

Issues with Automatic Evaluation

Word Based Metrics



Spearman correlations of word based metrics and human ratings

Word Overlap Metrics

- highly correlated with each other
- Not so correlated with human ratings

Human Ratings

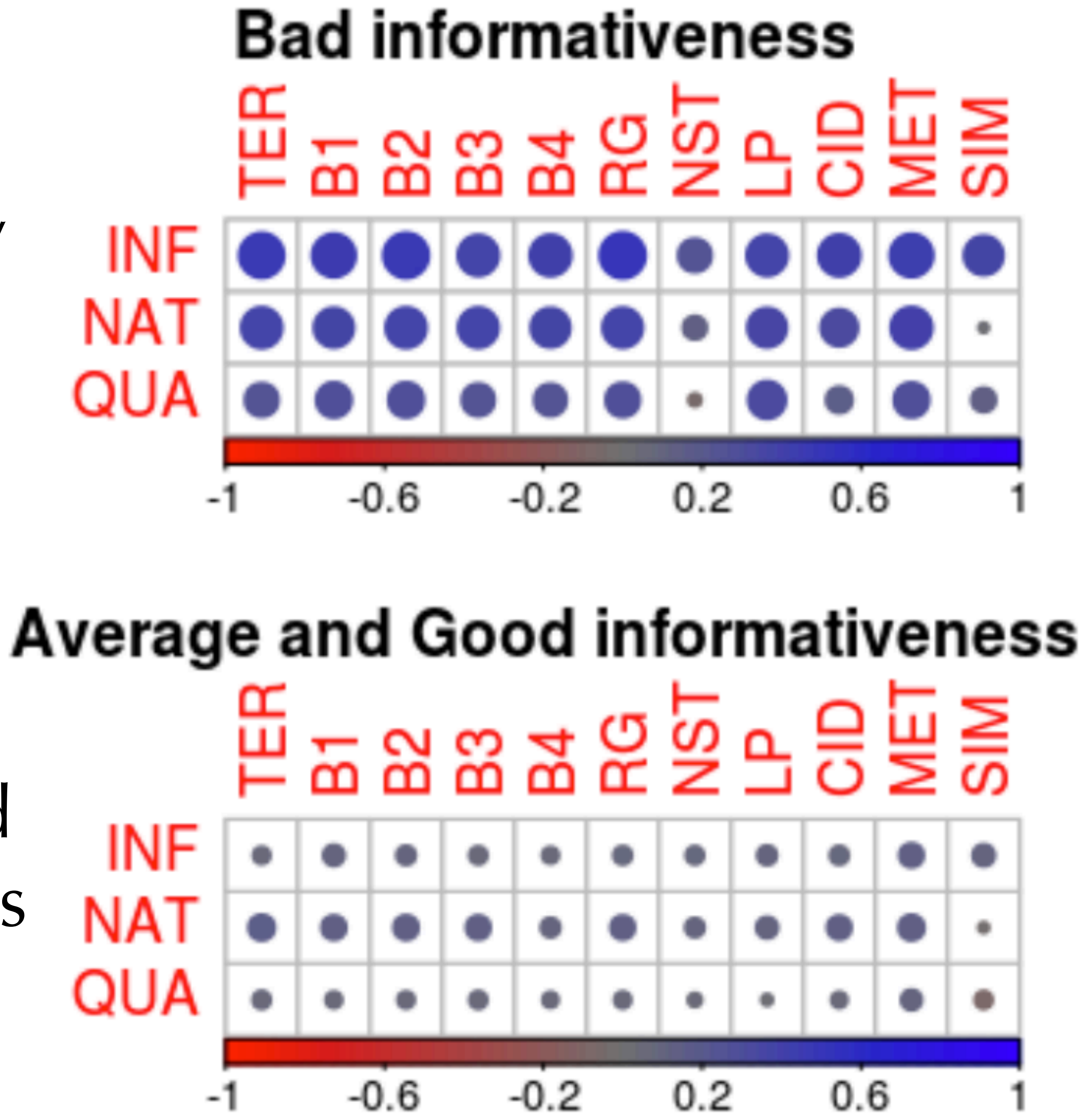
- Informativeness
- Naturalness
- Quality

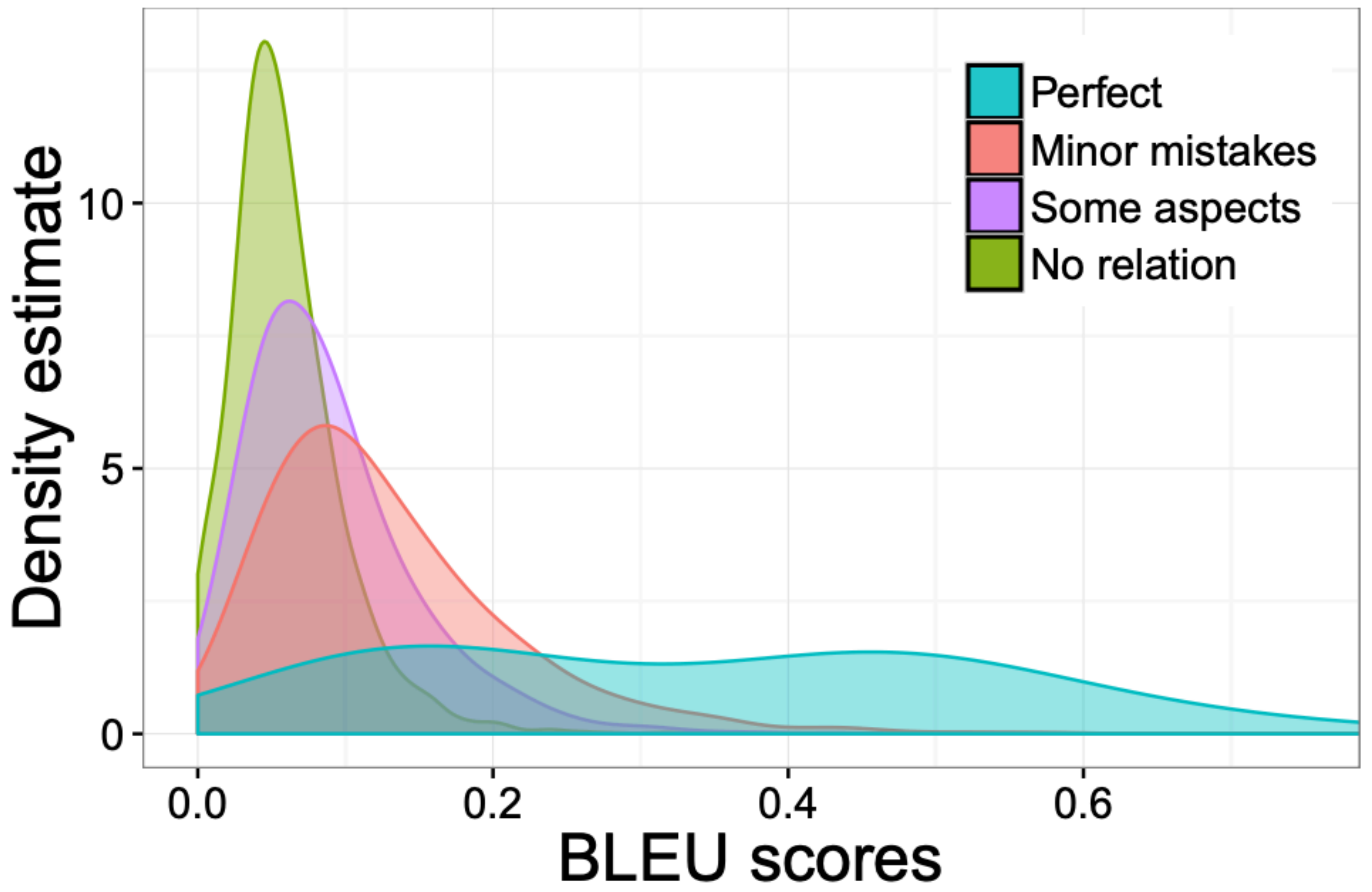
[Why We Need New Evaluation Metrics for NLG, Novikova et al 2017, https://arxiv.org/pdf/1707.06875.pdf]

Issues with automatic Evaluation

High correlation with human judgement for low quality generations

Poor correlation with human judgement for mid to high quality generations





[Why We Need New Evaluation Metrics for NLG, Novikova et al 2017, https://arxiv.org/pdf/1707.06875.pdf]

Re-evaluating Automatic Metrics for Image Captioning
[Kilickaya et al, EACL 2017]

Human evaluation

What kind of human evaluation can be done?

- Can get ratings from chat **participants** or external **observers**.
- Can ask humans to rate various aspects of the chat (**likert scale**) or to compare two chats / responses (**AB testing**)
- Dimensions: fluency, coherence / consistency, factuality and correctness, commonsense, style / formality, grammaticality, typicality, redundancy

Issues with human evaluation

- slow, expensive
- not repeatable (subjective/inconsistent)
- difficult to form well-targeted questions that are not open to misinterpretation

When developing new automatic metrics, human evaluation is used as gold

- New automated metrics must correlate well with human evaluation.

Evaluation takeaways

- **Content overlap metrics** provide a good starting point for evaluating the quality of generated text, but they're not good enough on their own.
- **Model-based metrics** can be more correlated with human judgment, but metric may not be not interpretable
- **Human judgments** are critical
  - But humans are inconsistent and judgments are expensive
- If you are developing a NLG system, you should
  - Look at your model generations. Don't just rely on numbers!
  - Publicly release large samples of the output of systems that you create!

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)