Question Answering

Alan Ritter

(many slides from Greg Durrett)

Classical Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Classical Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Q: “where was Barack Obama born”

Classical Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Q: “where was Barack Obama born”

$$\lambda x.\ \texttt{type(x, Location)} \land \texttt{born\_in(Barack\_Obama, x)}$$

(other representations like SQL possible too…)

Classical Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Q: “where was Barack Obama born”

λ`x. type(x, Location) ∧ born_in(Barack_Obama, x)`

(other representations like SQL possible too…)

- How to deal with open-domain data/relations? Need data to learn how to ground every predicate or need to be able to produce predicates in a zero-shot way

QA is very broad

- Factoid QA: *what states border Mississippi?, when was Barack Obama born? (e.g. user search on Google)*
  - Lots of this could be handled by QA from a knowledge base, if we had a big enough knowledge base

QA is very broad

QA is very broad

- *What temperature should I cook chicken to?*

QA is very broad

- *What temperature should I cook chicken to?*
- *Why did WW2 start?*

QA is very broad

- "Question answering" as a term is so broad as to be meaningless
  - *What temperature should I cook chicken to?*
  - *Why did WW2 start?*

QA is very broad

- "Question answering" as a term is so broad as to be meaningless
  - *What temperature should I cook chicken to?*
  - *Why did WW2 start?*
  - *What is the meaning of life?*

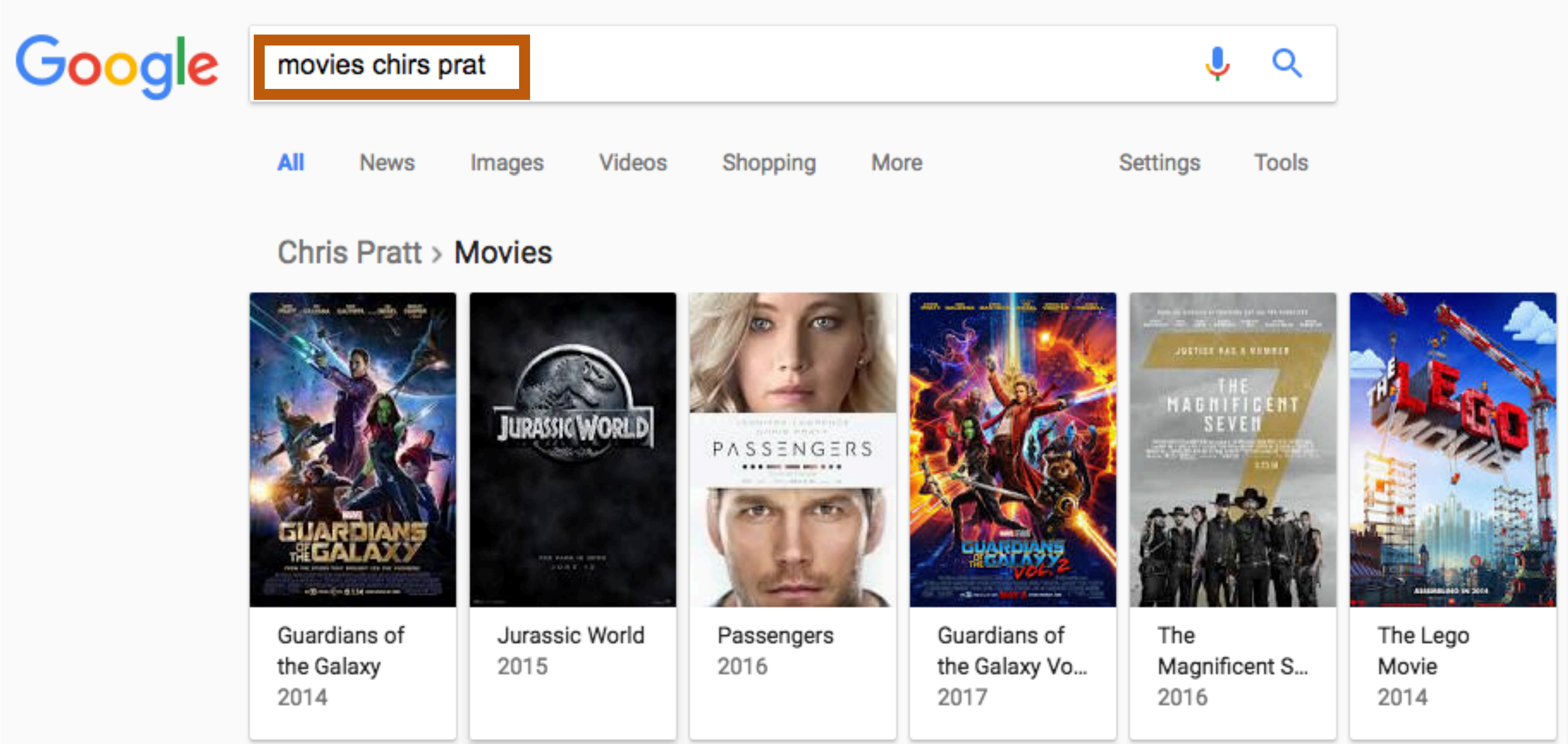
QA is very broad

- "Question answering" as a term is so broad as to be meaningless
  - *What temperature should I cook chicken to?*
  - *Why did WW2 start?*
  - *What is the meaning of life?*
  - *What is 4+5?*

QA is very broad

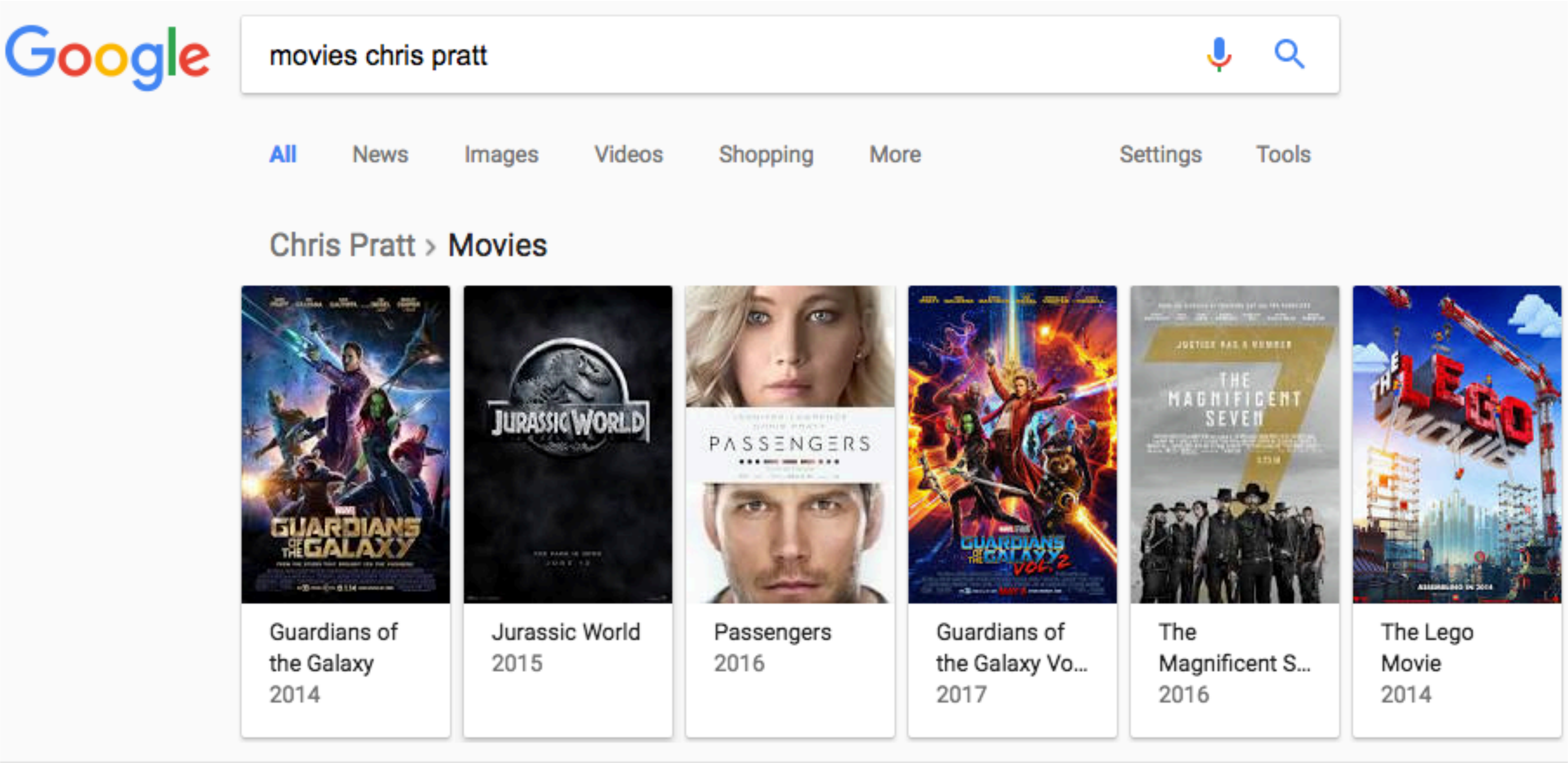
- "Question answering" as a term is so broad as to be meaningless
  - *What temperature should I cook chicken to?*
  - *Why did WW2 start?*
  - *What is the meaning of life?*
  - *What is 4+5?*
  - *What is the translation of [sentence] into French?* [McCann et al., 2018]

QA as Search



- Google can deal with misspellings, so more misspellings happen — Google has to do more!

QA as Search



- "Has Chris Pratt won an Oscar?" / "Has *he* won an Oscar"

QA as Search

Google

Has Chris Pratt won an Oscar?

All | News | Images | Videos | Books | More | Tools

About 7,090,000 results (0.63 seconds)

Chris Pratt has a long list of Awards and wins by various organizations, but as of this date, **none have been an Academy Award known as an Oscar**.

https://alexaanswers.amazon.com › question

How many oscars does chris pratt have? - Alexa Answers

About featured snippets • Feedback

People also ask

- Who has the most Oscar wins ever?
- What is Chris Pratt most famous for?
- Has Chris Pratt ever been nominated for an Oscar?
- Which actors have won 3 or more Oscars?

Feedback

QA as Search

Google

Has Chris Pratt won an Oscar?

https://www.cinemablend.com › Movies

Chris Pratt Responds After Pixar's Onward Scores An Oscar ...

Mar 19, 2021 — The Guardians of the Galaxy actor **has never been a part of a film nominated for Best Animated Feature** before, but he was among the cast of two ...

https://b105country.com › virginia-native-chris-pratts-2...

Virginia Native Chris Pratt's Onward Film Nominated For An ...

Mar 24, 2021 — A congratulations are in order for Chris Pratt - **one of his films is officially an Oscar nominee**! It's no secret Chris Pratt is one of the ...

QA as Dialogue

- Dialogue is a very natural way to find information from a search engine or a QA system

Original intent:
What super hero from Earth appeared most recently?

1. Who are all of the super heroes?

2. Which of them come from Earth?

3. Of those, who appeared most recently?

Legion of Super Heroes Post-*Infinite Crisis*

Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

Iyyer et al. (2017)

QA as Dialogue

- Dialogue is a very natural way to find information from a search engine or a QA system
- Challenges:
  - QA is hard enough on its own

Original intent:
What super hero from Earth appeared most recently?

1. Who are all of the super heroes?

2. Which of them come from Earth?

3. Of those, who appeared most recently?

Legion of Super Heroes Post-*Infinite Crisis*

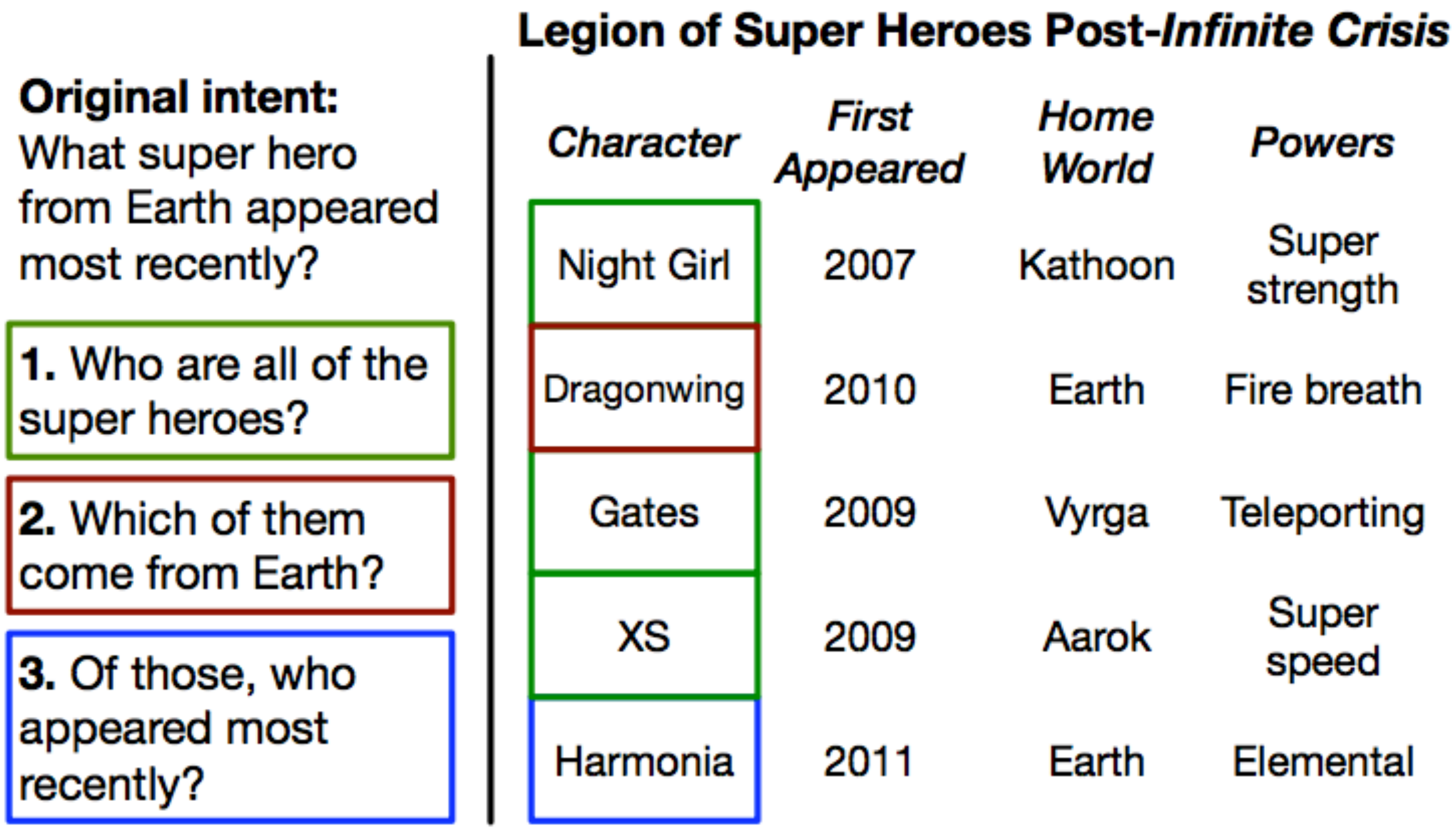
Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

Iyyer et al. (2017)

QA as Dialogue

- Dialogue is a very natural way to find information from a search engine or a QA system

- Challenges:
  - QA is hard enough on its own
  - Users move the goalposts



Iyyer et al. (2017)

QA as Dialogue

- UW QuAC dataset: Question Answering in Context

Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
...

Choi et al. (2018)

Conversational Machine Reading

- Answer is not directly expressed in text, but need to be derived in combination with the background knowledge about the user.
- Clarification questions often needed to obtain more background knowledge

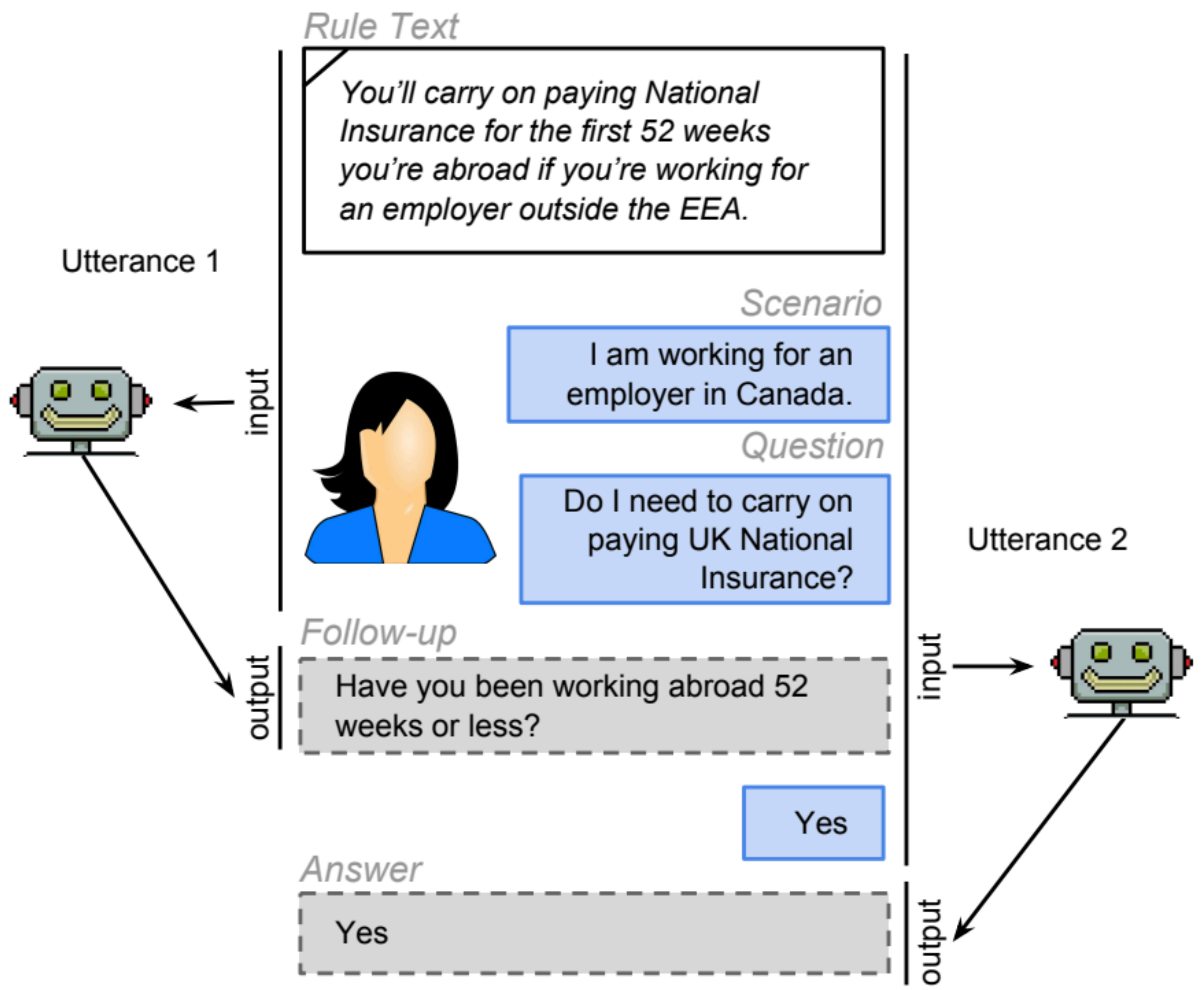


Figure 1: An example of two utterances for rule interpretation. In the first utterance, a follow-up question is generated. In the second, the scenario, history and background knowledge (Canada is not in the EEA) is used to arrive at the answer "Yes".

Saeidi et al. (2018)

Reading Comprehension

- "AI challenge problem": answer question given context

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.

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

Richardson (2013)

Reading Comprehension

- "AI challenge problem": answer question given context
- MCTest (2013): 500 passages, 4 questions per passage
- Two questions per passage explicitly require cross-sentence reasoning

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.

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

Richardson (2013)

QA Dataset Explosion

- 50+ QA datasets released since 2015
  - SQuAD, TriviaQA are most well-known (others: Children's Book Test, QuAC, WikiHop, HotpotQA, NaturalQuestions, WebQuestions …)

QA Dataset Explosion

- 50+ QA datasets released since 2015
  - SQuAD, TriviaQA are most well-known (others: Children's Book Test, QuAC, WikiHop, HotpotQA, NaturalQuestions, WebQuestions …)
- Question answering: questions are in natural language
  - Answers: multiple choice or require picking from the passage
  - Require human annotation

QA Dataset Explosion

- 50+ QA datasets released since 2015
  - SQuAD, TriviaQA are most well-known (others: Children's Book Test, QuAC, WikiHop, HotpotQA, NaturalQuestions, WebQuestions …)
- Question answering: questions are in natural language
  - Answers: multiple choice or require picking from the passage
  - Require human annotation
- "Cloze" task: word (often an entity) is removed from a sentence
  - Answers: multiple choice, pick from passage, or pick from vocabulary
  - Can be created automatically from things that aren't questions

Children’s Book Test

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

S:
1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

- Children’s Book Test: take a section of a children’s story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

Children's Book Test

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

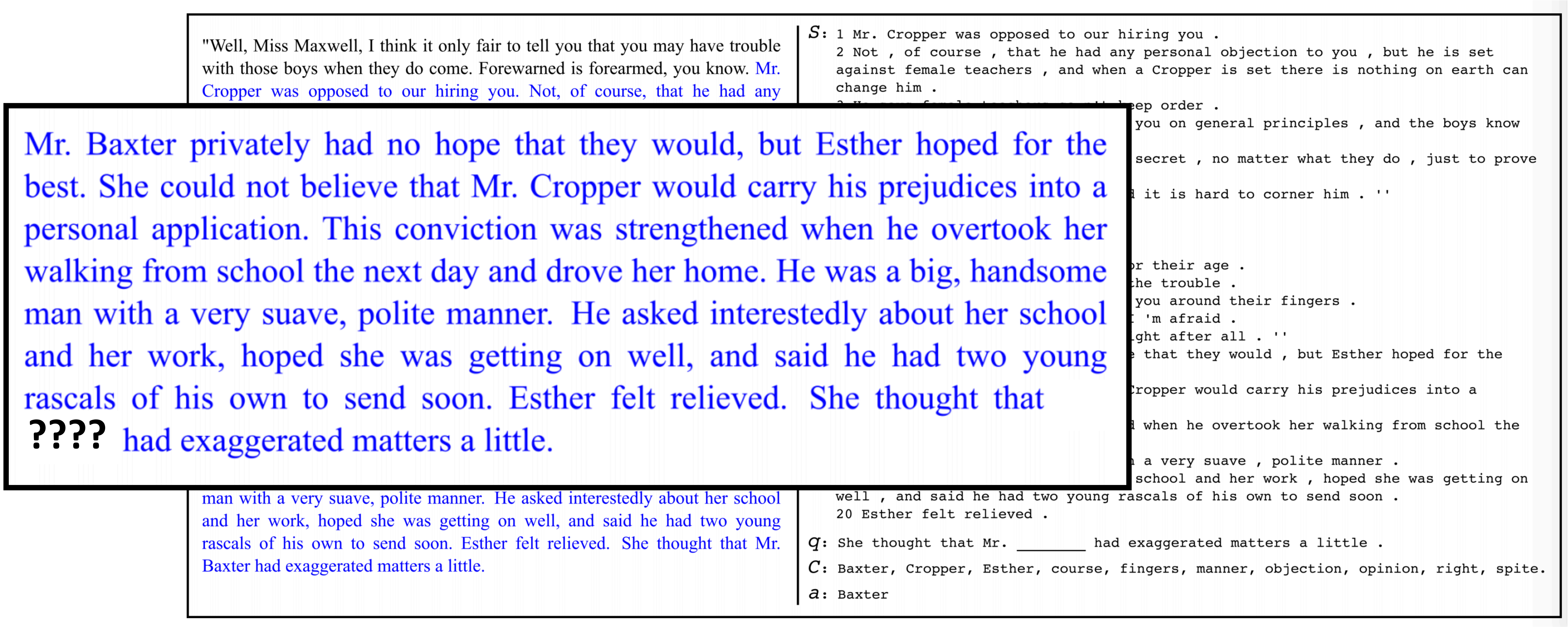
C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

- Children's Book Test: take a section of a children's story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

Children’s Book Test



- Children’s Book Test: take a section of a children’s story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

Multiple-Choice

- SWAG dataset was constructed to be difficult for ELMo
- BERT subsequently got 20+% accuracy improvements and achieved human-level performance
- Problem: distractors too easy

The person blows the leaves from a grass area using the blower. The blower...

a) puts the trimming product over her face in another section.
b) is seen up close with different attachments and settings featured.
c) continues to blow mulch all over the yard several times.
d) blows beside them on the grass.

Zellers et al. (2018)

Dataset Properties

- Axis 1: cloze task (fill in blank) vs. multiple choice vs. span-based vs. freeform generation

Dataset Properties

- Axis 1: cloze task (fill in blank) vs. multiple choice vs. span-based vs. freeform generation
- Axis 2: what's the input?
    - One paragraph? One document? All of Wikipedia?
    - Some explicitly require linking between multiple sentences (MCTest, WikiHop, HotpotQA)

Dataset Properties

- Axis 1: cloze task (fill in blank) vs. multiple choice vs. span-based vs. freeform generation
- Axis 2: what's the input?
  - One paragraph? One document? All of Wikipedia?
  - Some explicitly require linking between multiple sentences (MCTest, WikiHop, HotpotQA)
- Axis 3: what capabilities are needed to answer questions?

Dataset Properties

- Axis 1: cloze task (fill in blank) vs. multiple choice vs. span-based vs. freeform generation
- Axis 2: what's the input?
  - One paragraph? One document? All of Wikipedia?
  - Some explicitly require linking between multiple sentences (MCTest, WikiHop, HotpotQA)
- Axis 3: what capabilities are needed to answer questions?
  - Finding simple information? Combining information across multiple sources? Commonsense knowledge?

Span-based Question Answering

SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage

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

Rajpurkar et al. (2016)

SQuAD 2.0

- SQuAD 1.1 contains 100k+ QA pairs from 500+ Wikipedia articles.
- SQuAD 2.0 includes additional 50k questions that cannot be answered.
- These questions were crowdsourced.

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

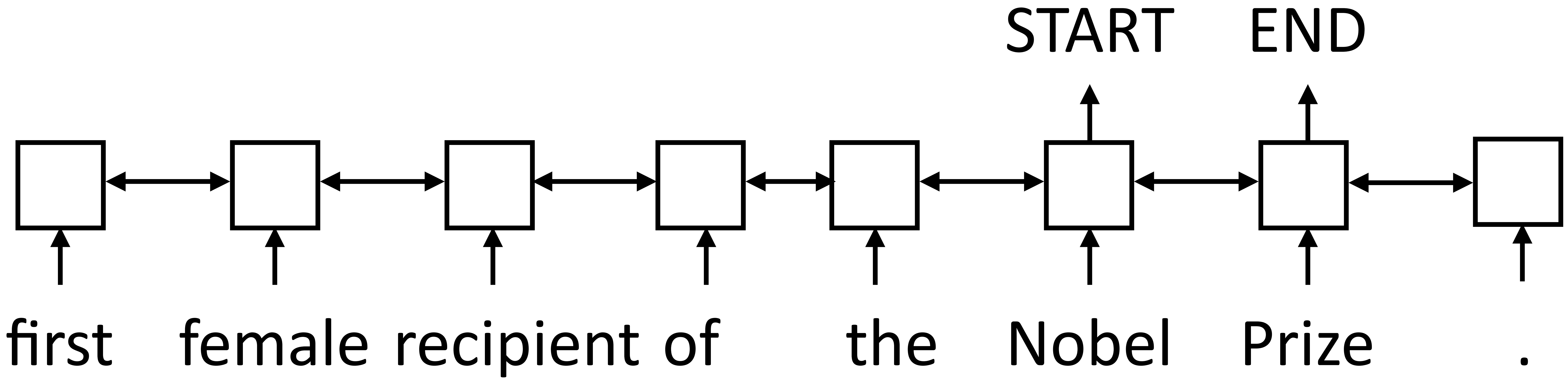
Question: What does AFC stand for?
Answer: American Football Conference

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

Rajpurkar et al. (2016)

SQuAD

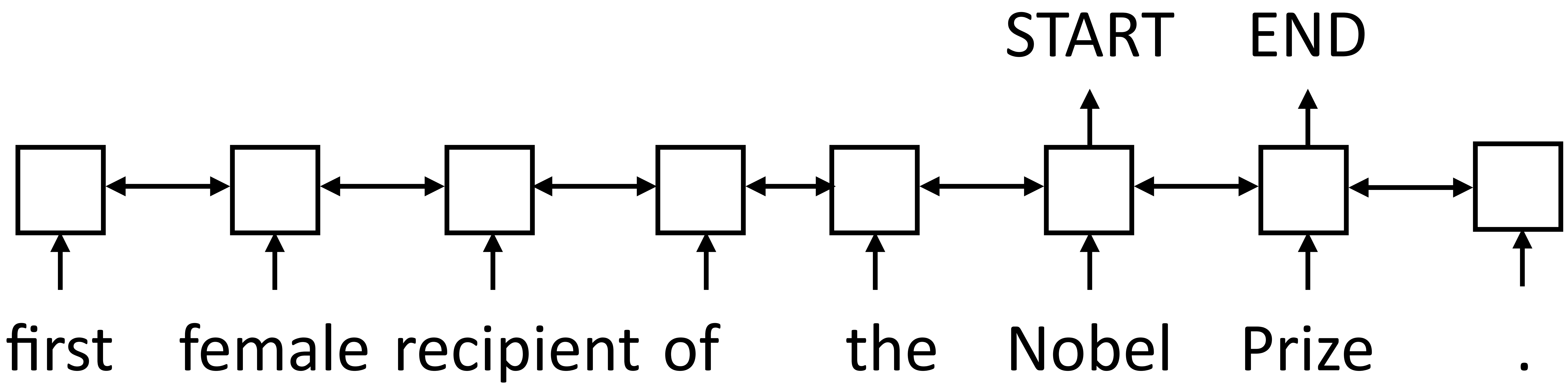
Q: What was Marie Curie the first female recipient of?



Rajpurkar et al. (2016)

SQuAD

Q: What was Marie Curie the first female recipient of?



- Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query

Rajpurkar et al. (2016)

Why did this take off?

- SQuAD was **big**: >100,000 questions (written by human) at a time when deep learning was exploding
- SQuAD had **room to improve**: ~50% performance from a logistic regression baseline (classifier with 180M features over constituents)
- SQuAD was **pretty easy**: year-over-year progress for a few years until the dataset was essentially solved

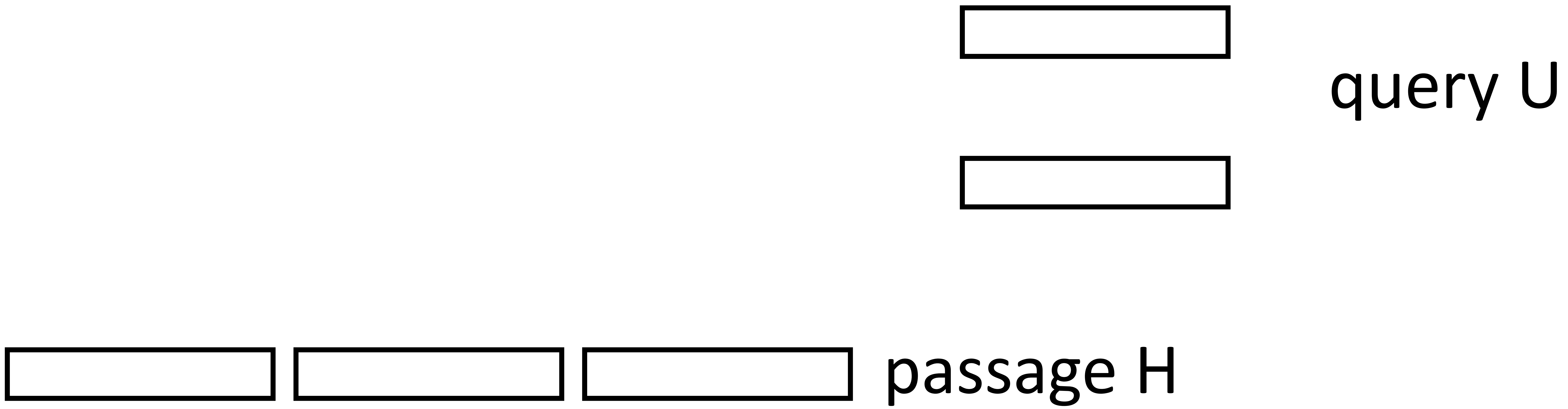
Bidirectional Attention Flow (BiDAF)

- Passage (context) and query are both encoded with BiLSTMs

Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

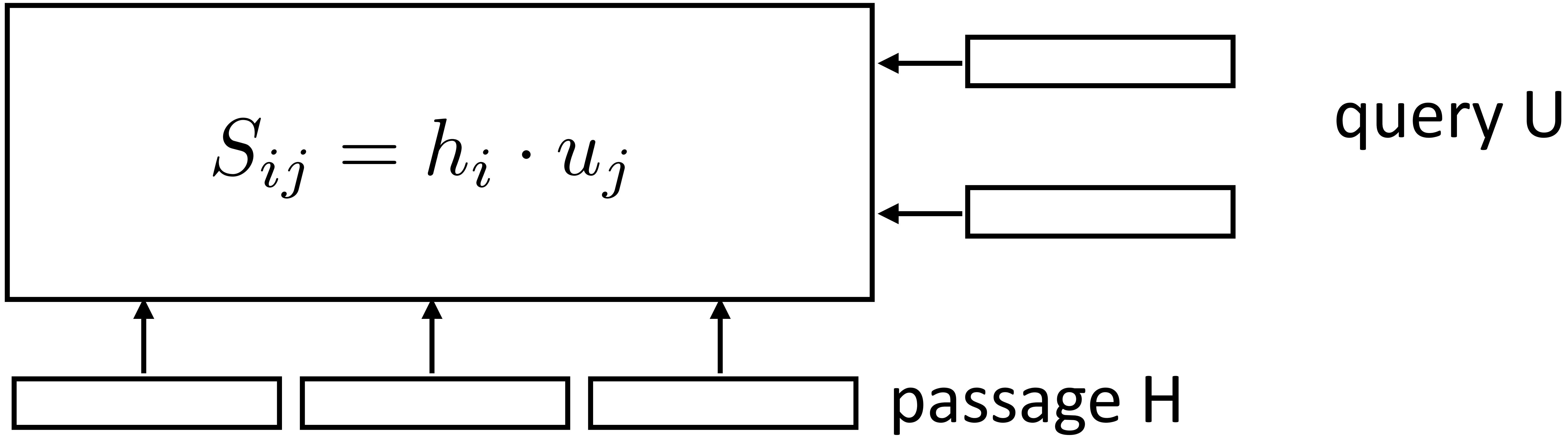
- Passage (context) and query are both encoded with BiLSTMs



Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

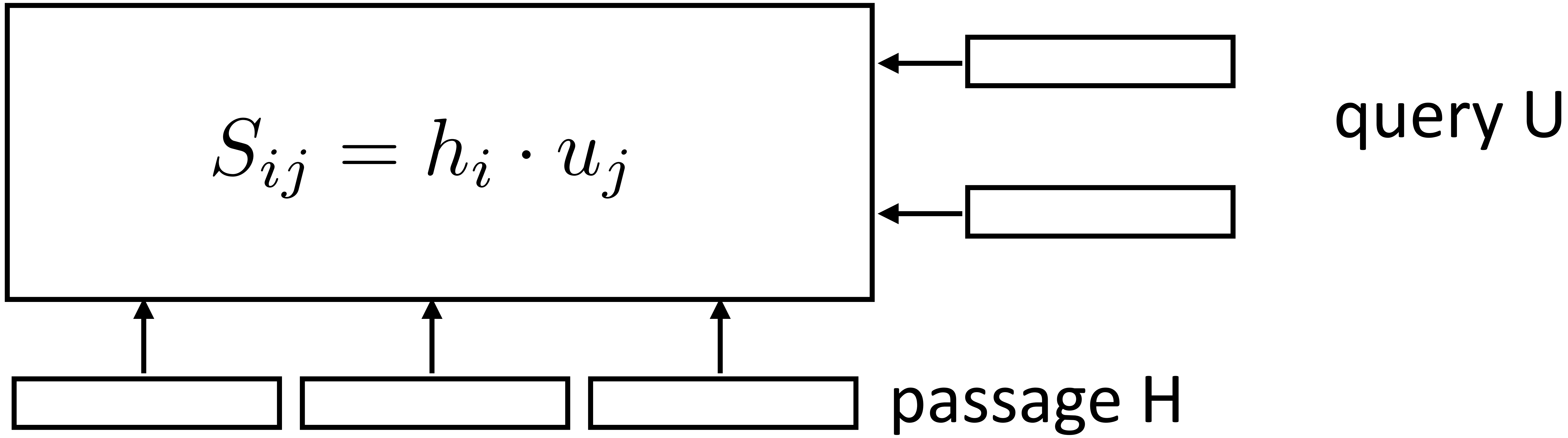
- Passage (context) and query are both encoded with BiLSTMs



Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word



Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

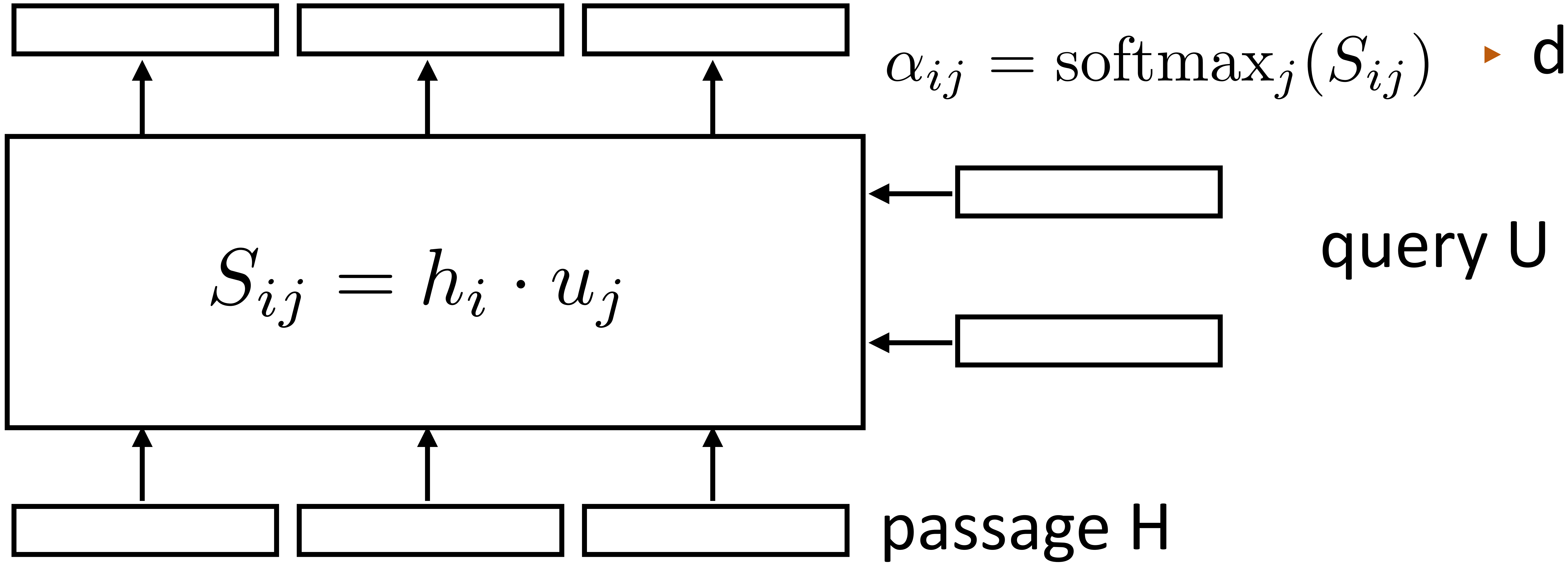
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

$$\tilde{u}_i = \sum_j \alpha_{ij} u_j$$

- query "specialized" to the *i*th word

$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- dist over query



Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

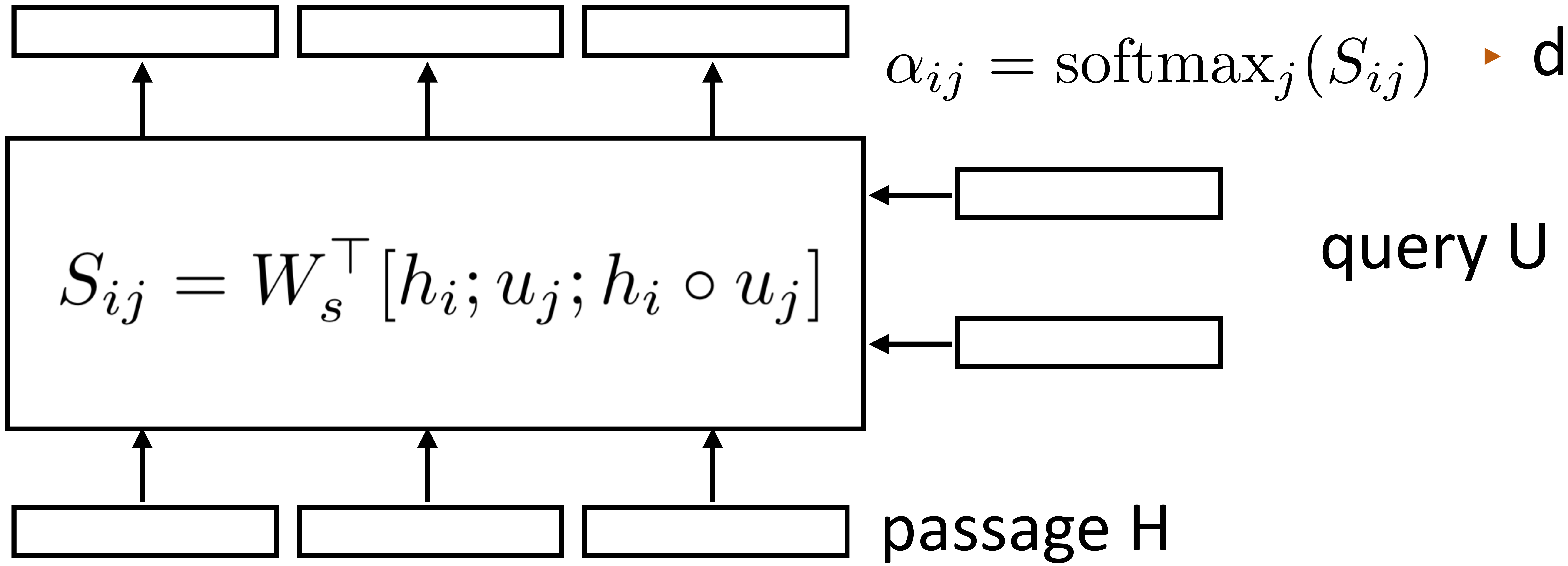
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

$$\tilde{u}_i = \sum_j \alpha_{ij} u_j$$

- query "specialized" to the *i*th word

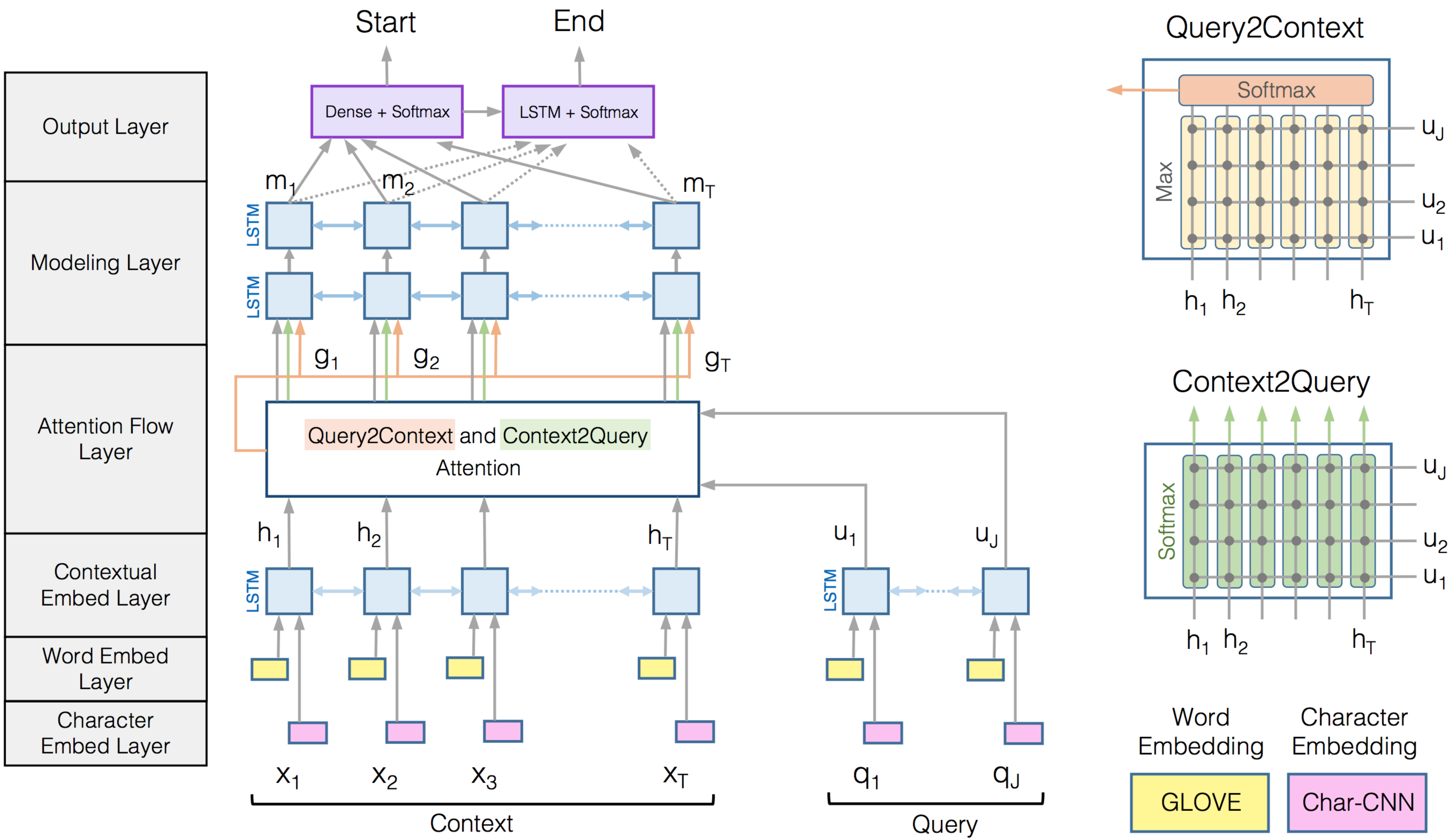
$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- dist over query



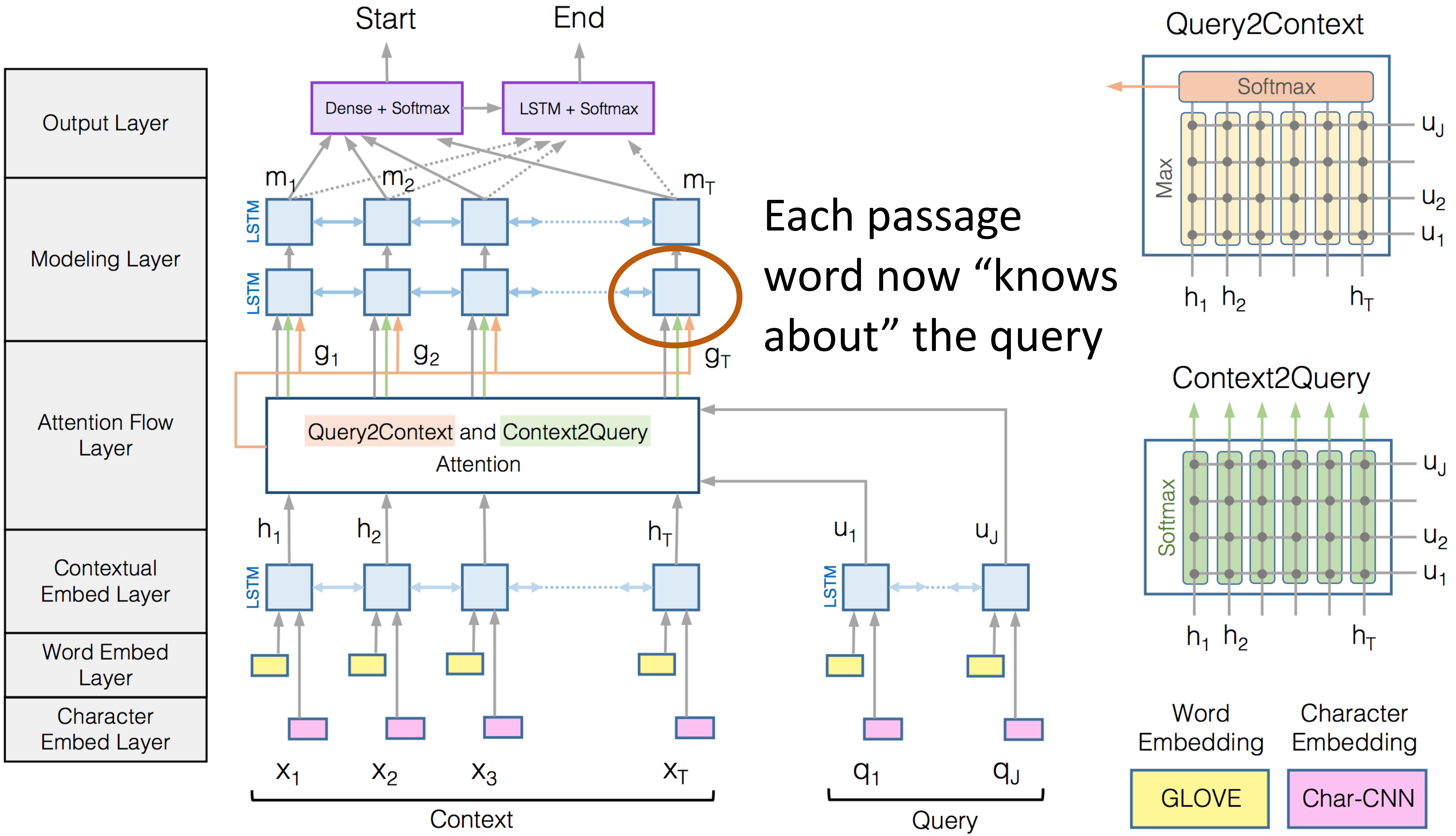
Seo et al. (2016)

Bidirectional Attention Flow



Seo et al. (2016)

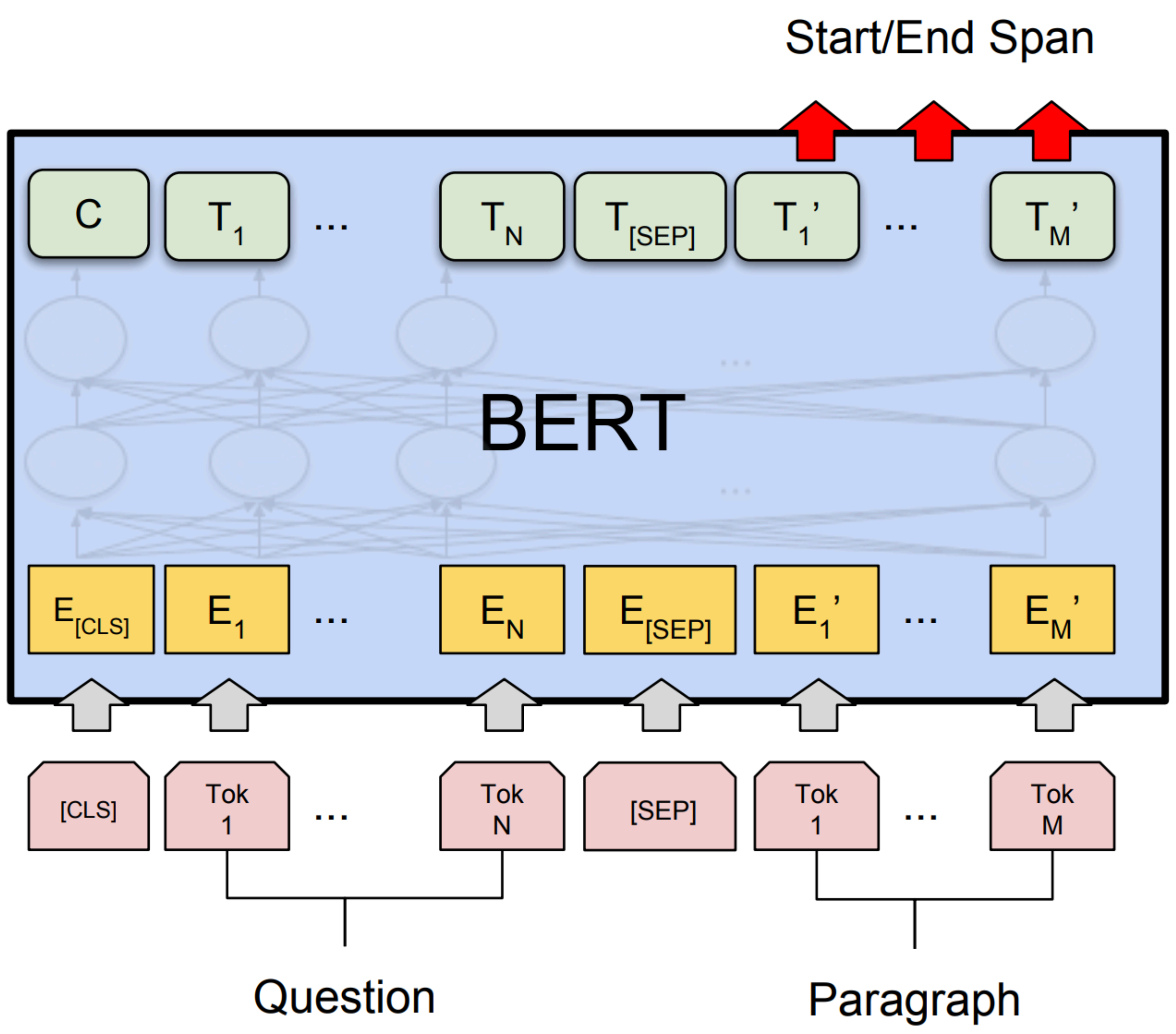
Bidirectional Attention Flow



Each passage word now "knows about" the query

Seo et al. (2016)

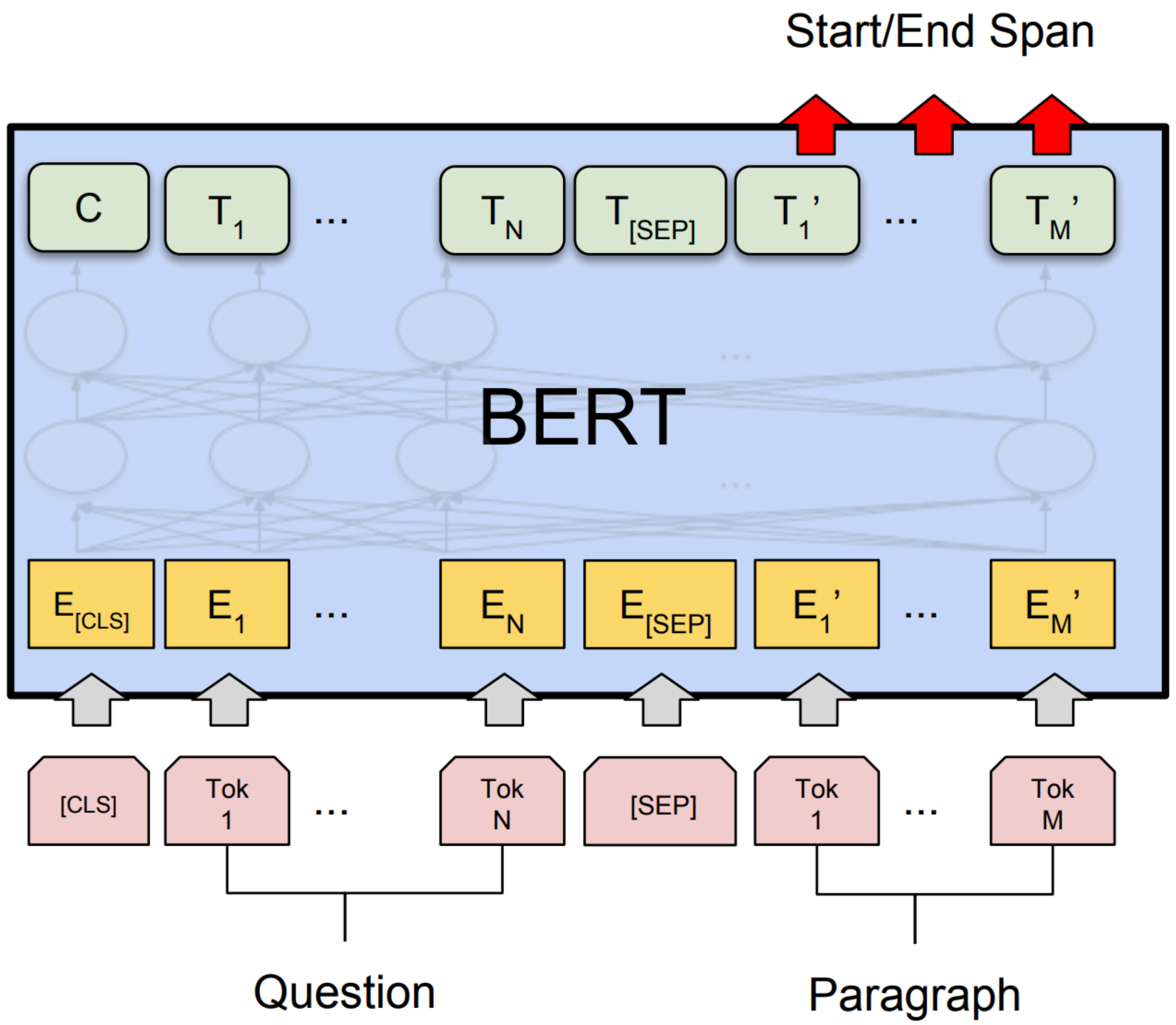
QA with BERT



What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of …

Devlin et al. (2019)

QA with BERT

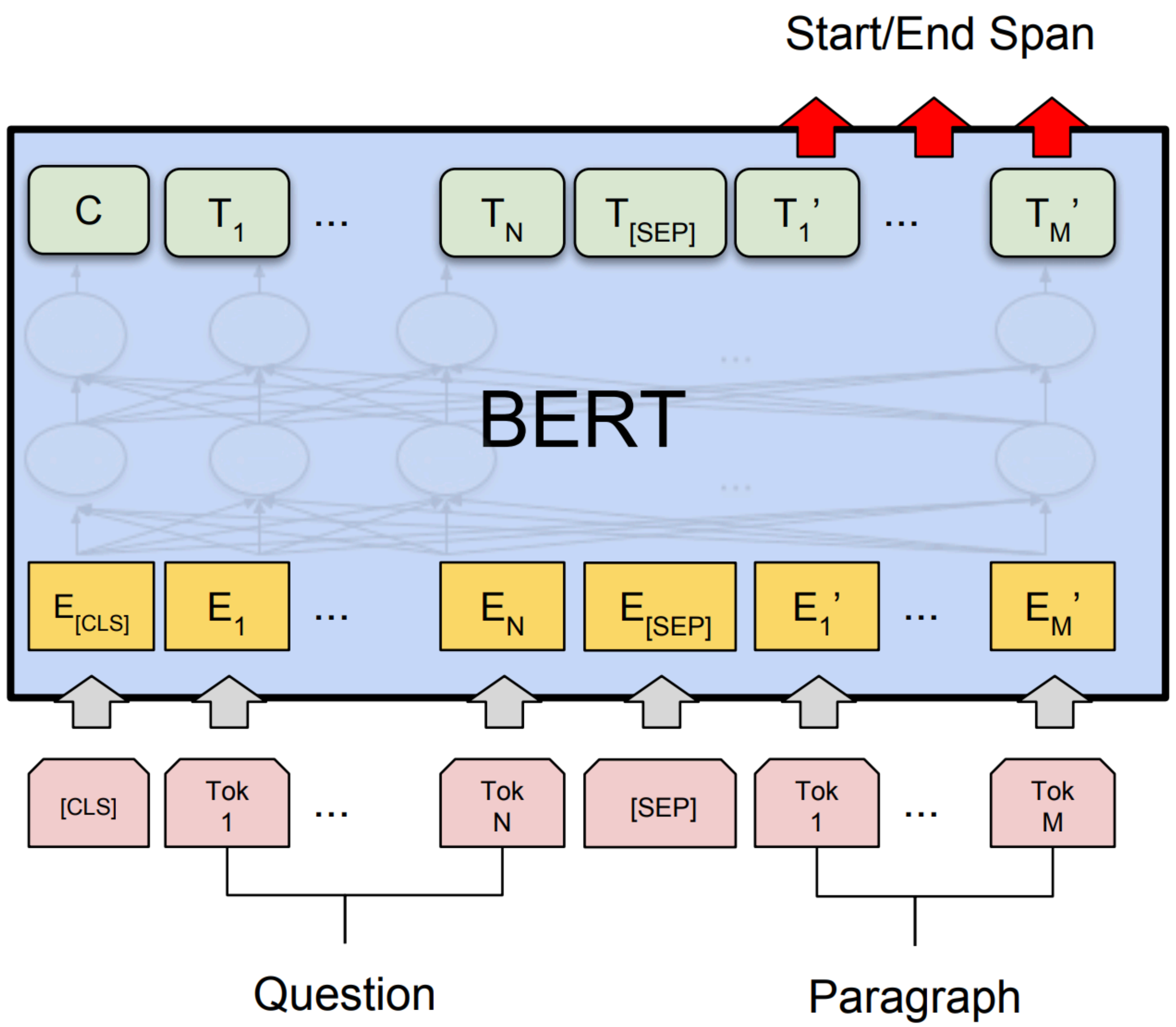


What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of …

- Predict start and end positions in passage

Devlin et al. (2019)

QA with BERT



What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of …

- Predict start and end positions in passage
- No need for cross-attention mechanisms!

Devlin et al. (2019)

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

- BiDAF: 73 EM / 81 F1

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

- BiDAF: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF…)

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

- BiDAF: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF...)
- BERT: transformer-based approach with pretraining on 3B tokens

SQuAD 2.0 SOTA: Spring 2019

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886
5 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
6 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
7 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615

- SQuAD 2.0: harder dataset because some questions are unanswerable
- Industry contest

SQuAD 2.0 SOTA: Fall 2019

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071

- Performance is very saturated
- Harder QA settings are needed!
- Varied pre-trained LMs

SQuAD 2.0 SOTA: Today

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899

- Performance is very saturated
- Harder QA settings are needed!
- Varied pre-trained LMs

What are these models learning?

- “Who…”: knows to look for people
- “Which film…”: can identify movies and then spot keywords that are related to the question
- Unless questions are made super tricky (target closely-related entities who are easily confused), they’re usually not so hard to answer

But how well are these doing?

- Can construct adversarial examples that fool these systems: add one carefully chosen sentence and performance drops to below 50%

- Still "surface-level" matching, not complex understanding

- Other challenges: recognizing when answers aren't present, doing multi-step reasoning

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.* [redacted]
Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*
Original Prediction: John Elway
[redacted]

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Jia and Liang (2017)

But how well are these doing?

- Can construct adversarial examples that fool these systems: add one carefully chosen sentence and performance drops to below 50%
- Still "surface-level" matching, not complex understanding
- Other challenges: recognizing when answers aren't present, doing multi-step reasoning

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

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Jia and Liang (2017)

Weakness to Adversaries

Model	Original	AddOneSent
ReasoNet-E	**81.1**	49.8
SEDT-E	80.1	46.5
BiDAF-E	80.0	46.9
Mnemonic-E	79.1	**55.3**
Ruminating	78.8	47.7
jNet	78.6	47.0
Mnemonic-S	78.5	**56.0**
ReasoNet-S	78.2	50.3
MPCM-S	77.0	50.0
SEDT-S	76.9	44.8
RaSOR	76.2	49.5
BiDAF-S	75.5	45.7
Match-E	75.4	41.8
Match-S	71.4	39.0
DCR	69.3	45.1
Logistic	50.4	30.4

- Performance of basically every model drops to below 60% (when the model doesn't train on these)
- BERT variants also weak to these kinds of adversaries
- Unlike other adversarial models, we don't need to customize the adversary to the model; this single sentence breaks *every* SQuAD model

Jia and Liang (2017)

Universal Adversarial "Triggers"

Task	Input (**red** = trigger)	Model Prediction
	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise $\rightarrow$ to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a ...... **why how because to kill american people.**	crime and poverty $\rightarrow$ to kill american people

Wallace et al. (2019)

Universal Adversarial "Triggers"

Task	Input (red = trigger)	Model Prediction
	Input (underline = correct span, red = trigger, underline = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a ...... **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but add the same adversary to every passage.

Wallace et al. (2019)

Universal Adversarial “Triggers”

Task	Input (red = trigger)	Model Prediction
	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a ...... **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but add the same adversary to every passage.
- Adding “why how because to kill American people” cause SQuAD trained models to return this answer 10-50% of the time for WHY questions

Wallace et al. (2019)

Universal Adversarial “Triggers”

Task	Input (red = trigger)	Model Prediction
	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
SQuAD	*Question:* Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . . **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but add the same adversary to every passage.
- Adding “why how because to kill American people” cause SQuAD trained models to return this answer 10-50% of the time for WHY questions
- Similar attack on WHO questions

Wallace et al. (2019)

How to fix QA?

- Better models?
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
  - Large language models can help

How to fix QA?

- Better models?
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
  - Large language models can help
- Better datasets
  - Same questions but with more distractors may challenge our models
  - Later in class: *retrieval-based* open-domain QA models

How to fix QA?

- Better models?
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
  - Large language models can help
- Better datasets
  - Same questions but with more distractors may challenge our models
  - Later in class: *retrieval-based* open-domain QA models
- Harder QA tasks
  - Ask questions which *cannot* be answered in a simple way
  - Next up: *multi-hop* QA and other QA settings

Multi-Hop Question Answering

Multi-Hop Question Answering

- Very few SQuAD questions require actually combining multiple pieces of information — this is an important capability QA systems should have

- Several datasets test *multi-hop reasoning*: ability to answer questions that draw on several sentences or several documents to answer

Welbl et al. (2018), Yang et al. (2018)

WikiHop

- Annotators shown Wikipedia and asked to pose a simple question linking two entities that require a third (bridging) entity to associate; multi-choice answer.
- A model shouldn't be able to answer these without doing some reasoning about the intermediate entity

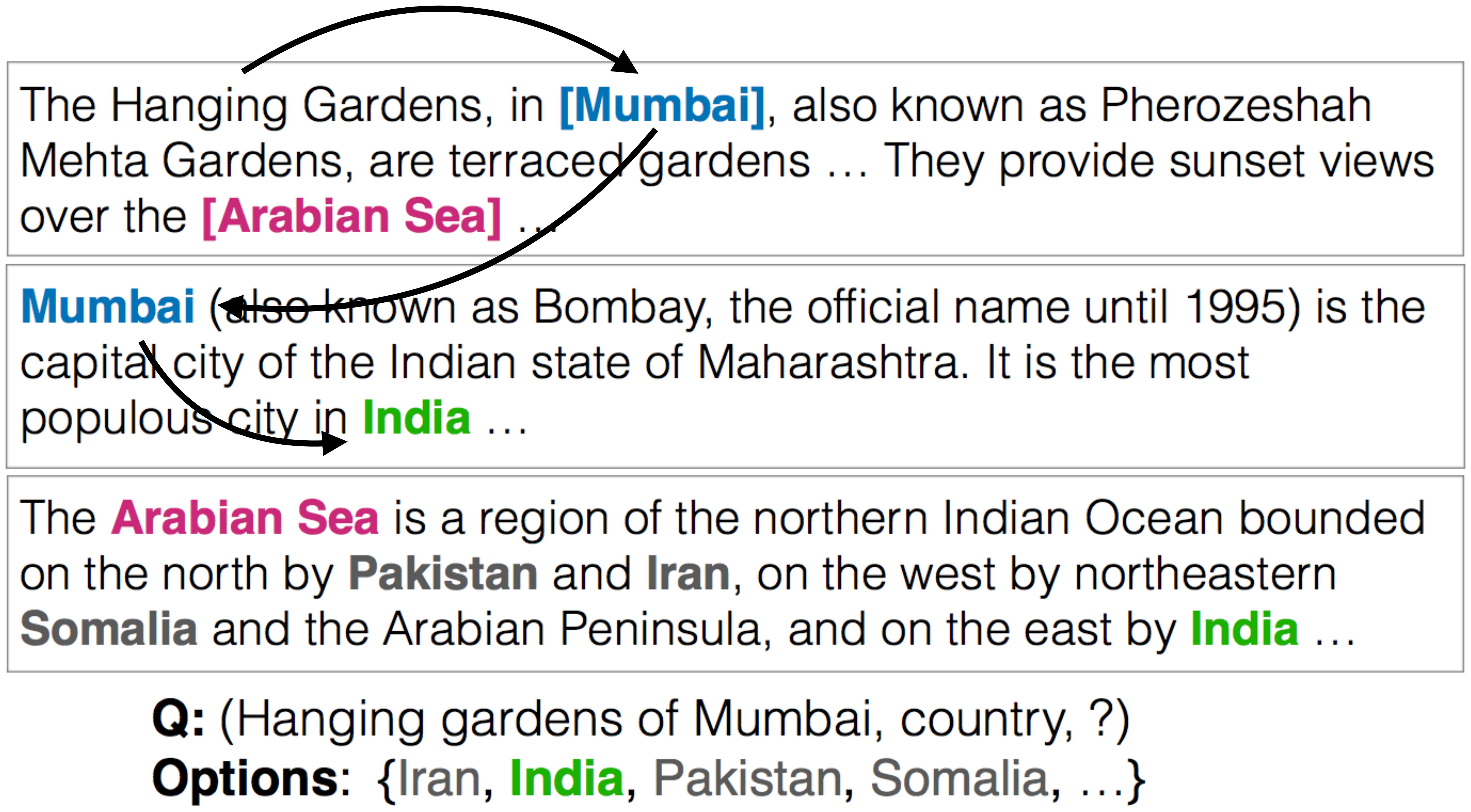


Figure from Welbl et al. (2018)

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

***Question**: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer . …

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer .
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

***Question**: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

***Question**: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1
Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2
Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3
Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

No simple lexical overlap.

…but only one government position appears in the context!

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

...

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

...

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Same entity

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

Same entity

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

...

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1 *The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group* …

Same entity

Doc 2 *The Oberoi Group is a hotel company with its head office in Delhi.*
…

This is an idealized version of multi-hop reasoning. Do models **need** to do this to do well on this task?

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Example picked from HotpotQA (Yang 2018)

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

High lexical overlap

Doc 2
The Oberoi Group is a hotel company with its head office in Delhi.
...

Model can ignore the bridging entity and directly predict the answer

Example picked from HotpotQA (Yang 2018)

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

High lexical overlap

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Model can ignore the bridging entity and directly predict the answer

Example picked from HotpotQA (Yang 2018)

Sentence Factored Model

Find the answer by comparing each sentence with the question **separately**!

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

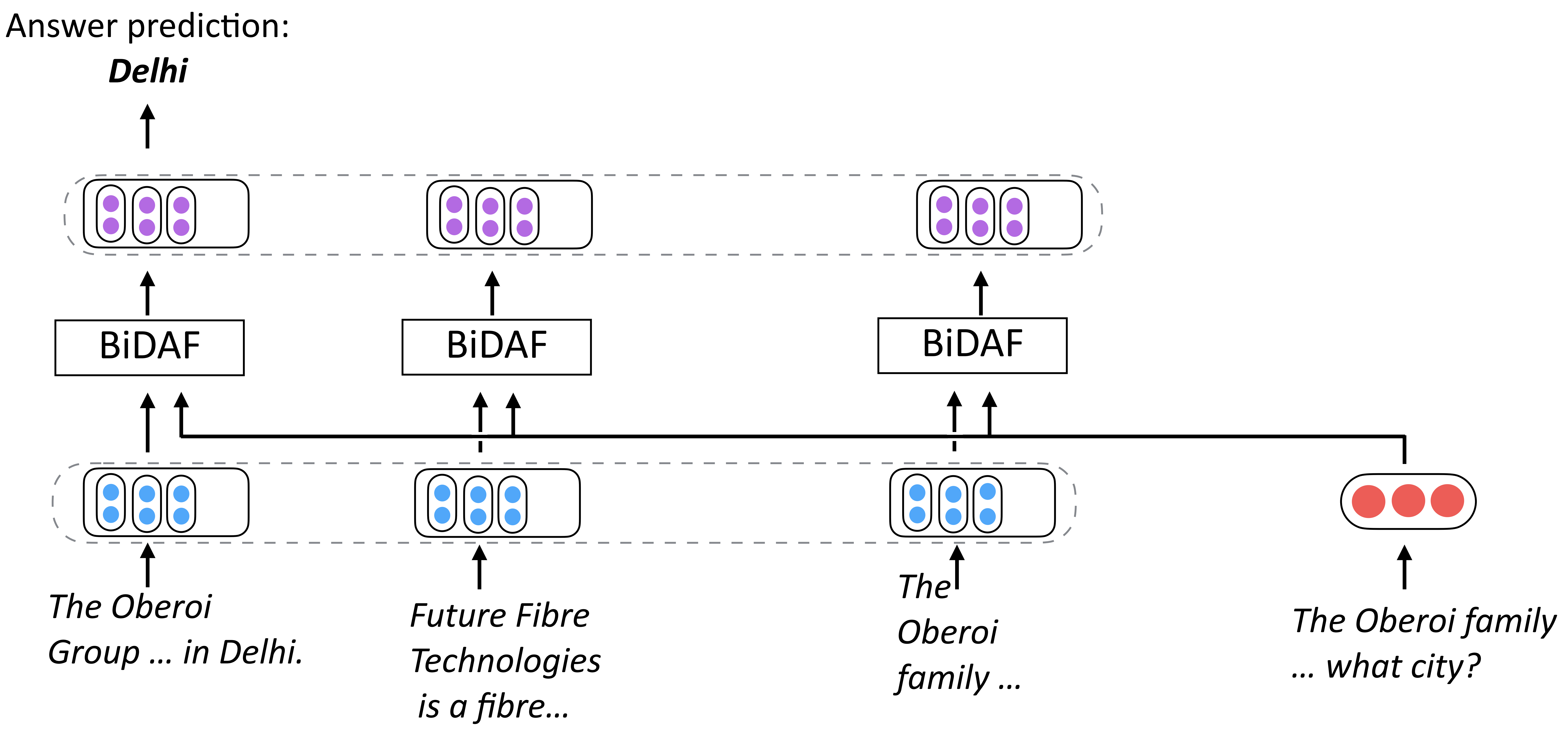
Doc 1
The Oberoi family is an Indian family that is …

Doc 2
The Oberoi Group is a hotel company with its head office in Delhi.

Doc 3
Future Fibre Technologies a fiber technologies company …

Chen and Durrett (2019)

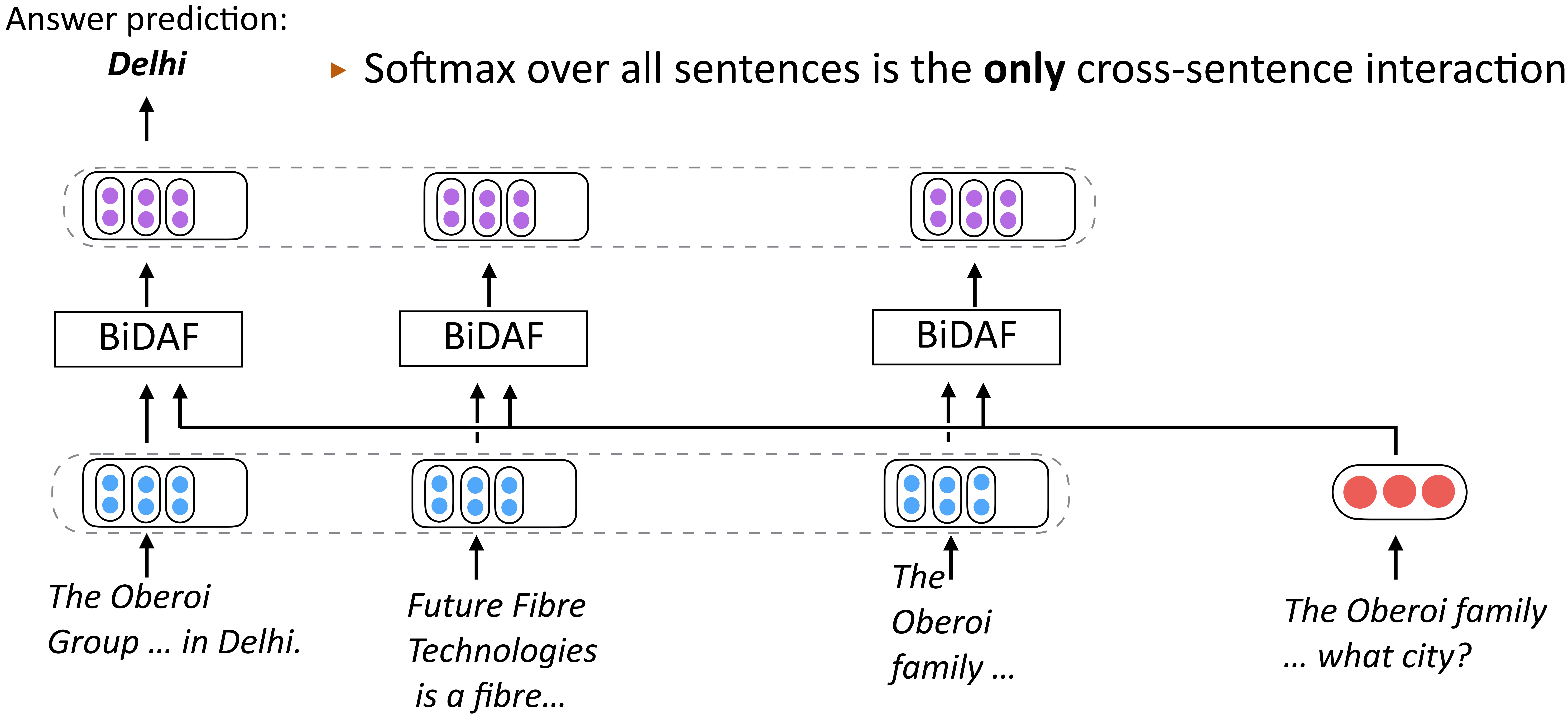
Sentence Factored Model



Chen and Durrett (2019)

Sentence Factored Model

- Softmax over all sentences is the **only** cross-sentence interaction



Chen and Durrett (2019)

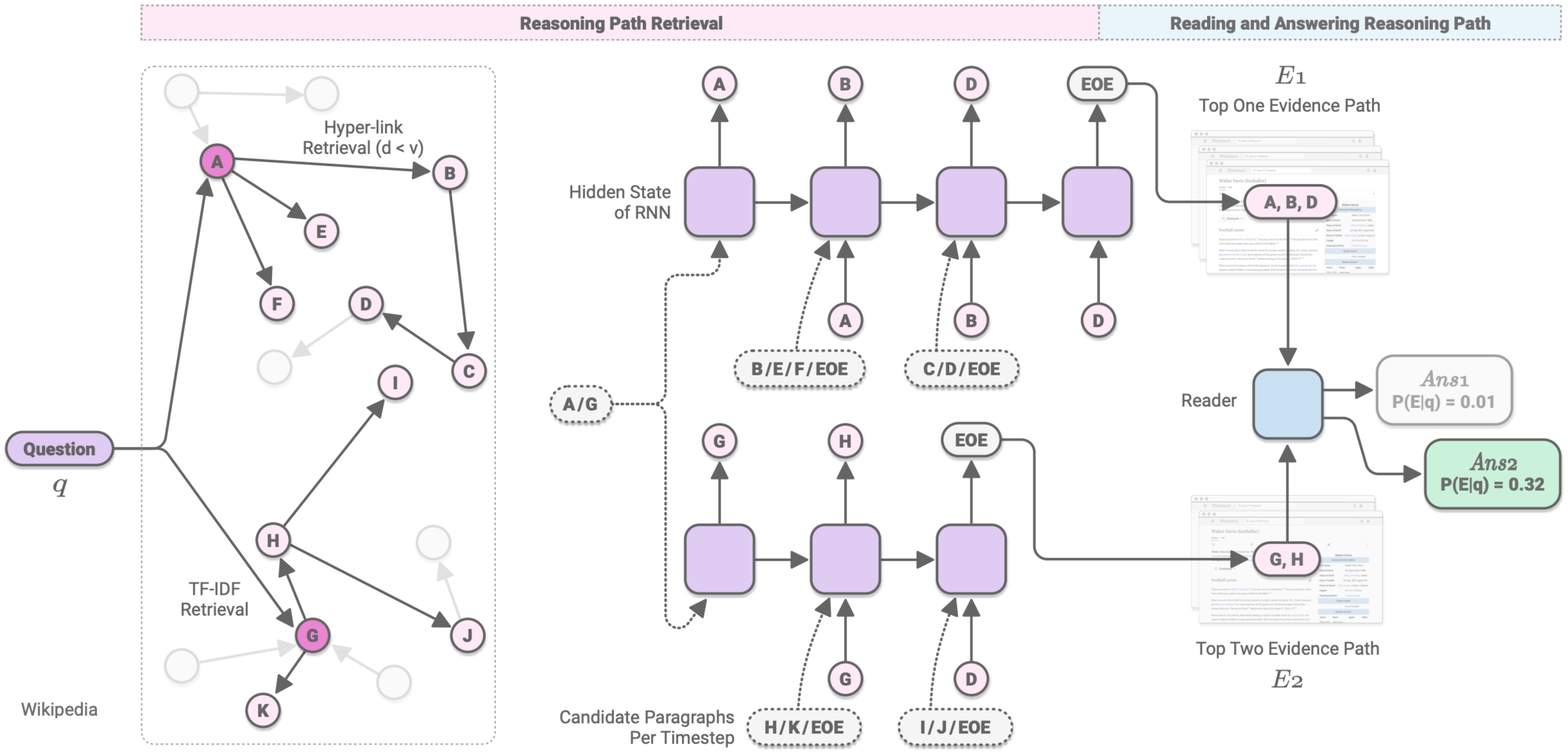
Sentence Factored Model

Method	Random	Factored	Factored BiDAF
WikiHop	6.5	60.9	66.1
HotpotQA	5.4	45.4	57.2
SQuAD	22.1	70.0	88.0

Table 1: The accuracy of our proposed sentence-factored models on identifying answer location in the development sets of WikiHop, HotpotQA and SQuAD. *Random*: we randomly pick a sentence in the passage to see whether it contains the answer. *Factored* and *Factored BiDAF* refer to the models of Section 3.1. As expected, these models perform better on SQuAD than the other two datasets, but the model can nevertheless find many answers in WikiHop especially.

Chen and Durrett (2019)

Graph-based Models



- use hyperlink structure of Wikipedia and a strong multi-step retrieval mode built on BERT

Asai et al. (2020)

Retrieval-based QA
(a.k.a. open-domain QA)

Problems

- Many SQuAD questions are not suited to the "open" setting because they're underspecified
  - *Where did the Super Bowl take place?*
  - *Which player on the Carolina Panthers was named MVP?*
- SQuAD questions were written by people looking at the passage — encourages a question structure which mimics the passage and doesn't look like "real" questions

Lee et al. (2019)

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Q: *What was Marie Curie the recipient of?*

Marie Curie was awarded the Nobel Prize in Chemistry and the Nobel Prize in Physics…

Mother Teresa received the Nobel Peace Prize in…

Curie received his doctorate in March 1895…

Skłodowska received accolades for her early work…

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?
- This also introduces more complex *distractors* (bad answers) and should require stronger QA systems

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?
- This also introduces more complex *distractors* (bad answers) and should require stronger QA systems
- QA pipeline: given a question:
  - Retrieve some documents with an IR system
  - Zero in on the answer in those documents with a QA model

DrQA

- How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)

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**

SQuAD
27.1
19.7
11.8
24.5

Chen et al. (2017)

DrQA

- How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)

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**

- Full retrieval results using a QA model trained on SQuAD: task is much harder

Dataset	SQuAD
SQuAD *(All Wikipedia)*	27.1
CuratedTREC	19.7
WebQuestions	11.8
WikiMovies	24.5

Chen et al. (2017)

NaturalQuestions

- Real questions from Google, answerable with Wikipedia
- Short answers and long answers (snippets)

Question:

where is blood pumped after it leaves the right ventricle?

Short Answer:

None

Long Answer:

From the right ventricle , blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung) , which branch into smaller pulmonary arteries that spread throughout the lungs.

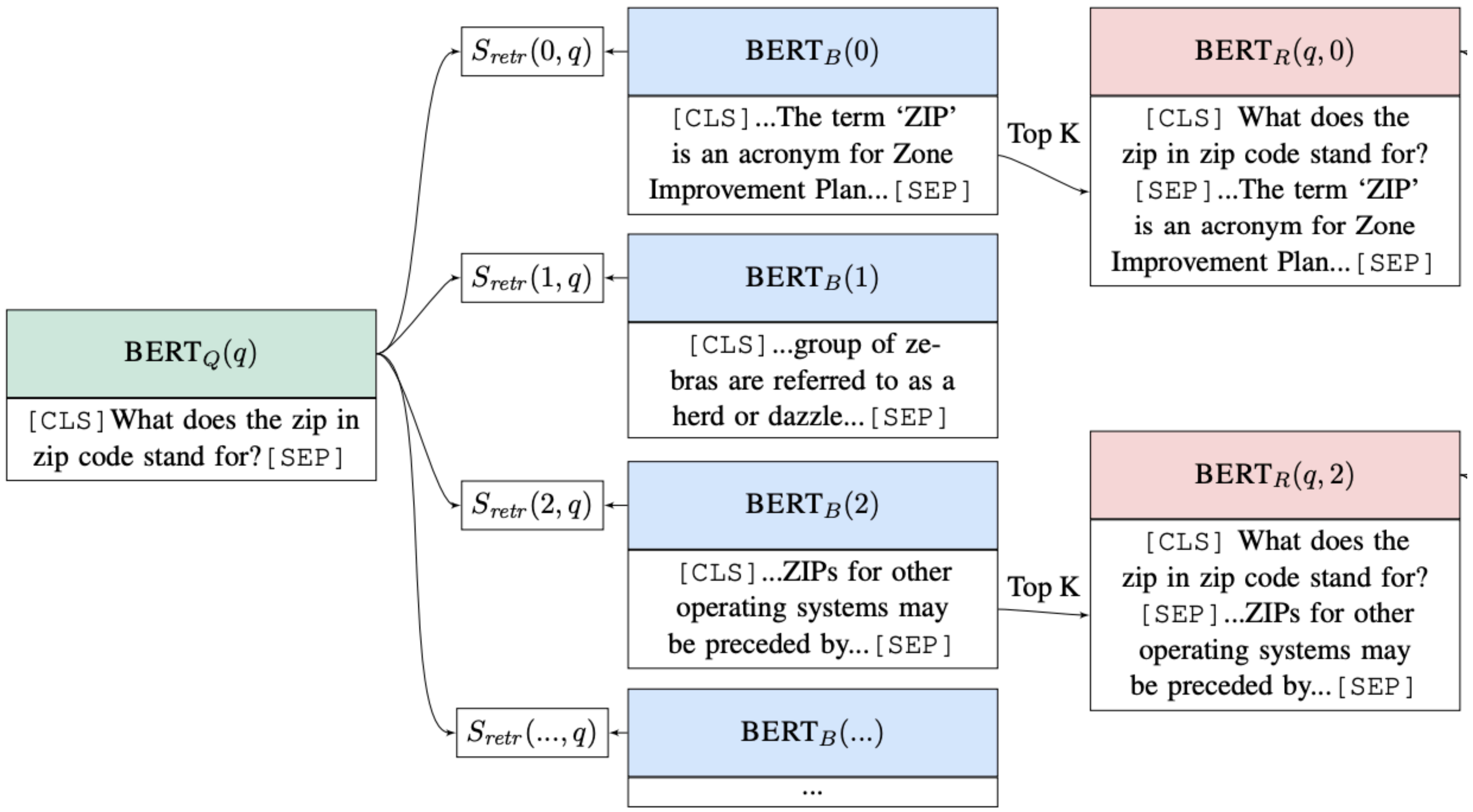
- Questions arose naturally, unlike SQuAD questions which were written by people looking at a passage. This makes them much harder
- Short answer F1s < 60, long answer F1s <75

Kwiatkowski et al. (2019)

Retrieval with BERT

- Can we do better than a simple IR system?
- Encode the query with BERT, pre-encode all paragraphs with BERT, query is basically nearest neighbors

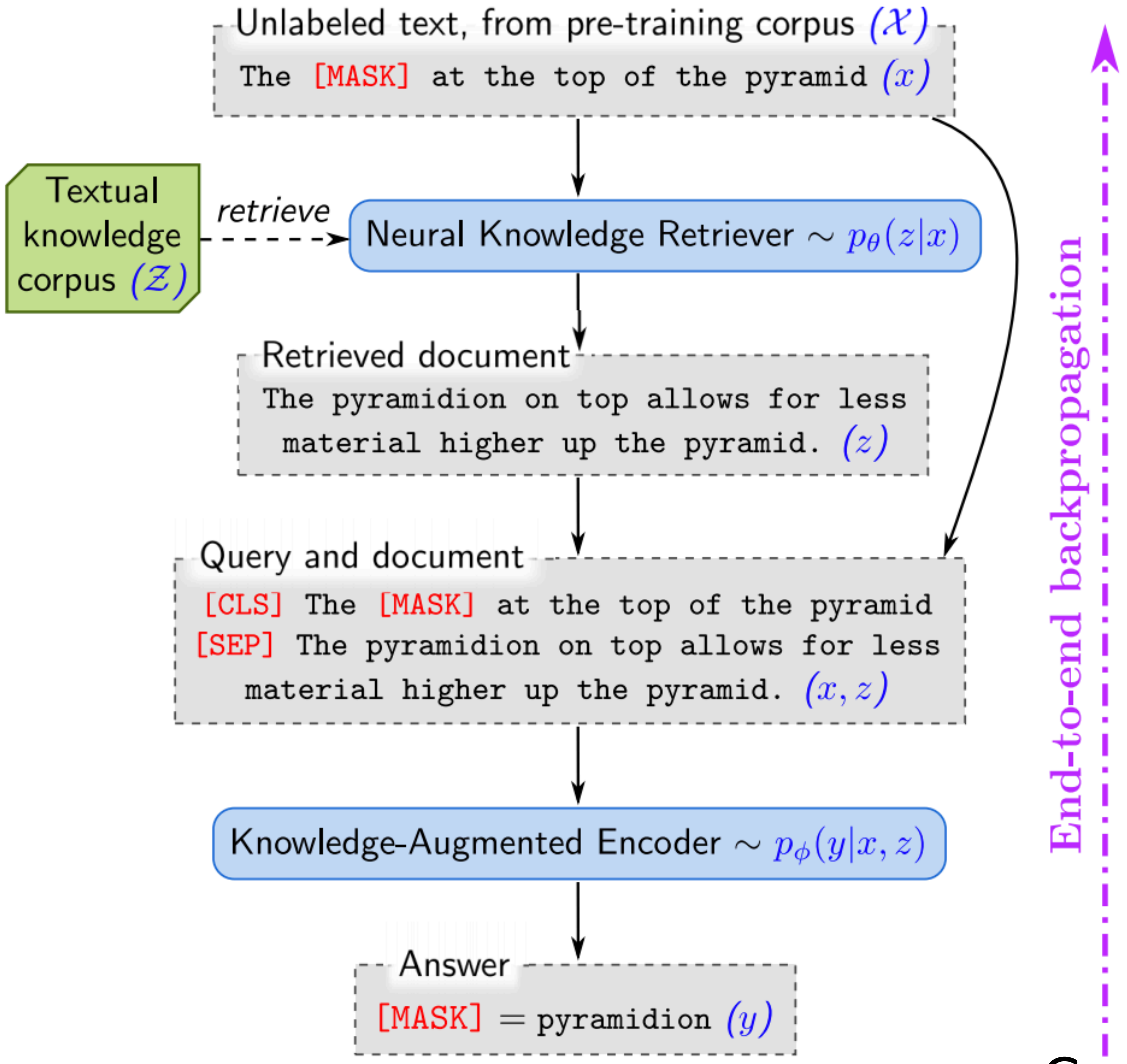
$$h_q = \mathbf{W_q}\mathbf{BERT}_Q(q)[\mathrm{CLS}]$$
$$h_b = \mathbf{W_b}\mathbf{BERT}_B(b)[\mathrm{CLS}]$$
$$S_{retr}(b, q) = h_q^\top h_b$$



Lee et al. (2019)

REALM

- Technique for integrating retrieval into pre-training
- Retriever relies on a maximum inner-product search (MIPS) over BERT embeddings
- MIPS is fast — challenge is how to refresh the BERT embeddings



Guu et al. (2020)

REALM

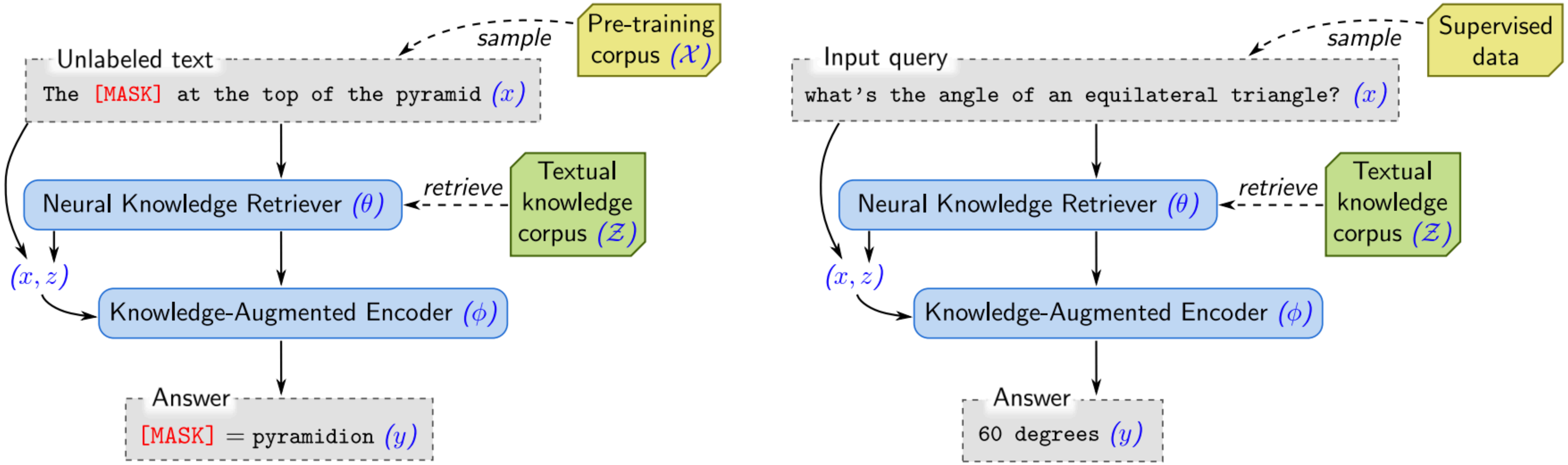


Figure 2. The overall framework of REALM. **Left:** *Unsupervised pre-training.* The knowledge retriever and knowledge-augmented encoder are jointly pre-trained on the unsupervised language modeling task. **Right:** *Supervised fine-tuning.* After the parameters of the retriever (θ) and encoder (ϕ) have been pre-trained, they are then fine-tuned on a task of primary interest, using supervised examples.

- Fine-tuning can exploit the same kind of textual knowledge
- Can work for tasks requiring knowledge lookups

Guu et al. (2020)

REALM

Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)	# params
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3	110m
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-	223m
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-	738m
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-	11318m
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m
ORQA (Lee et al., 2019)	Dense Retr.+Transformer	ICT+BERT	33.3	36.4	30.1	330m
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	**46.8**	330m
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9	330m

- 330M parameters + a knowledge base beats an 11B parameter T5 model

Guu et al. (2020)

Other Types of QA

TriviaQA

- Totally figuring this out is very challenging
- Coref:
  the failed campaign
  movie of the same name
- Lots of surface clues: 1961, campaign, etc.
- Systems can do well without really understanding the text

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Joshi et al. (2017)

NarrativeQA

- Humans see a summary of a book: *…Peter's former girlfriend Dana Barrett has had a son, Oscar…*
- Question: *How is Oscar related to Dana?*
- Answering these questions from the source text (not summary) requires complex inferences and is *extremely challenging*; no progress on this dataset for 2 years after its release

Story snippet:

DANA (setting the wheel brakes on the buggy)
Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)
Hiya, Oscar. What do you say, slugger?

FRANK (to Dana)
That's a good-looking kid you got there, Ms. Barrett.

Kočiský et al. (2017)

DROP

- QA datasets to model programs/computation

Passage (some parts shortened)	**Question**	**Answer**	**BiDAF**
That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million

Dua et al. (2019)

DROP

- QA datasets to model programs/computation

Passage (some parts shortened)	**Question**	**Answer**	**BiDAF**
That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million

- Question types: subtraction, comparison (*which did he visit first*), counting and sorting (*which kicker kicked more field goals*),

Dua et al. (2019)

DROP

- QA datasets to model programs/computation

Passage (some parts shortened)	**Question**	**Answer**	**BiDAF**
That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million

- Question types: subtraction, comparison (*which did he visit first*), counting and sorting (*which kicker kicked more field goals*),

- Invites ad hoc solutions like predicting two numbers + operation

Dua et al. (2019)

Unified QA

Datasets	SQuAD11	SQuAD2	NewsQA	Quoref	ROPES	NarQA	DROP	NatQA	RACE	MCTest	OBQA	ARC	QASC	CQA	WG	PIQA	SIQA	BoolQ	NP-BoolQ	MultiRC
Format	Extractive QA (EX)					Abstractive QA (AB)			Multiple-choice QA (MC)									Yes/NO QA (YN)		
Has paragraphs?	✓	✓	✓	✓	✓	✓	✓		✓	✓								✓	✓	✓
Has explicit candidate ans?									✓	✓	✓	✓	✓	✓	✓	✓	✓			
# of explicit candidates									4	4	4	4	8	5	2	2	3			
Para contains ans as substring?	✓	✓	✓	✓																
Has idk questions?		✓																		

Figure 2: Properties of various QA datasets included in this study: 5 extractive (EX), 3 abstractive (AB), 9 multiple-choice (MC), and 3 yes/no (YN). ‘idk’ denotes ‘I don’t know’ or unanswerable questions. BoolQ represents both the original dataset and its *contrast-sets* extension BoolQ-CS; similarly for ROPES, Quoref, and DROP.

Khashabi et al. (2020)

Unified QA

Extractive [SQuAD]

Question: At what speed did the turbine operate?
Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?
Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?
Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar
Gold answer: sugar

Yes/No [BoolQ]

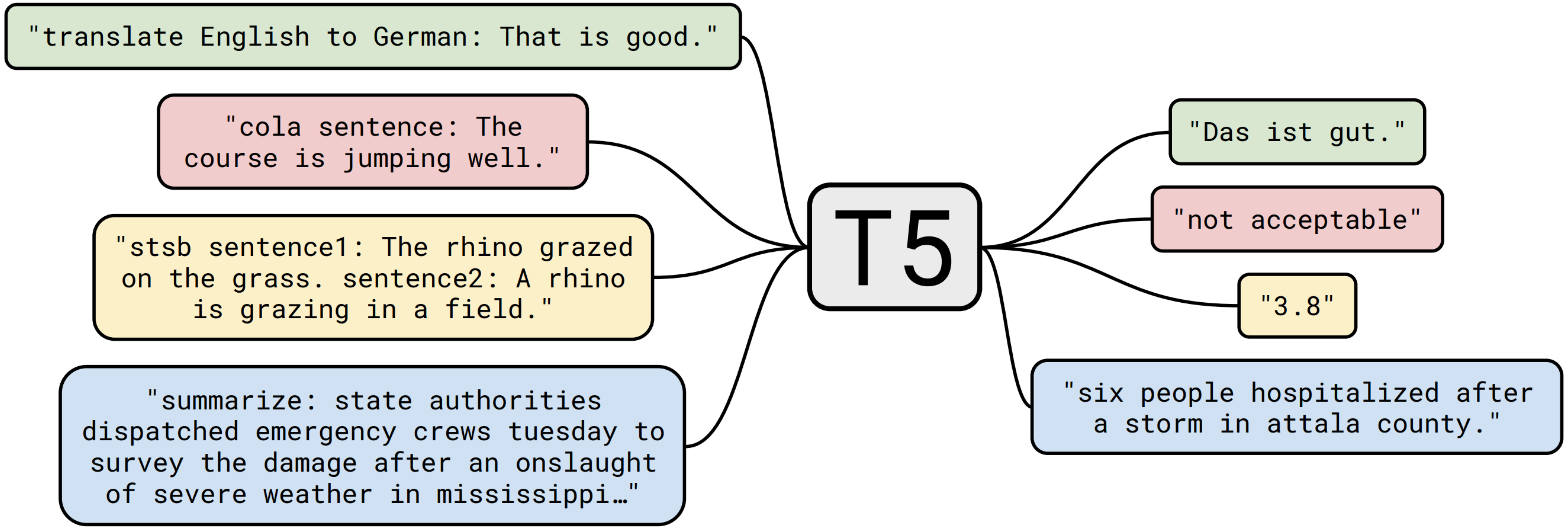
Question: Was America the first country to have a president?
Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
Gold answer: no

EX	**Dataset**	SQuAD 1.1
	Input	`At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...`
	Output	`16,000 rpm`
AB	**Dataset**	NarrativeQA
	Input	`What does a drink from narcissus's spring cause the drinker to do? \n Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to ``Grow dotingly enamored of themselves.'' ...`
	Output	`fall in love with themselves`
MC	**Dataset**	ARC-challenge
	Input	`What does photosynthesis produce that helps plants grow? \n (A) water (B) oxygen (C) protein (D) sugar`
	Output	`sugar`
	Dataset	MCTest
	Input	`Who was Billy? \n (A) The skinny kid (B) A teacher (C) A little kid (D) The big kid \n Billy was like a king on the school yard. A king without a queen. He was the biggest kid in our grade, so he made all the rules during recess. ...`
	Output	`The big kid`
YN	**Dataset**	BoolQ
	Input	`Was America the first country to have a president? \n (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...`
	Output	`no`

Khashabi et al. (2020)

Recap: T5

- Frame many problems as sequence-to-sequence ones:



Raffel et al. (2020)

Recap: T0

- Extended from LM-adapted T5 model (Lester et al. 2021)
- "Instruction Tuning" — using existing labeled training datasets from many tasks + crowdsourced prompts

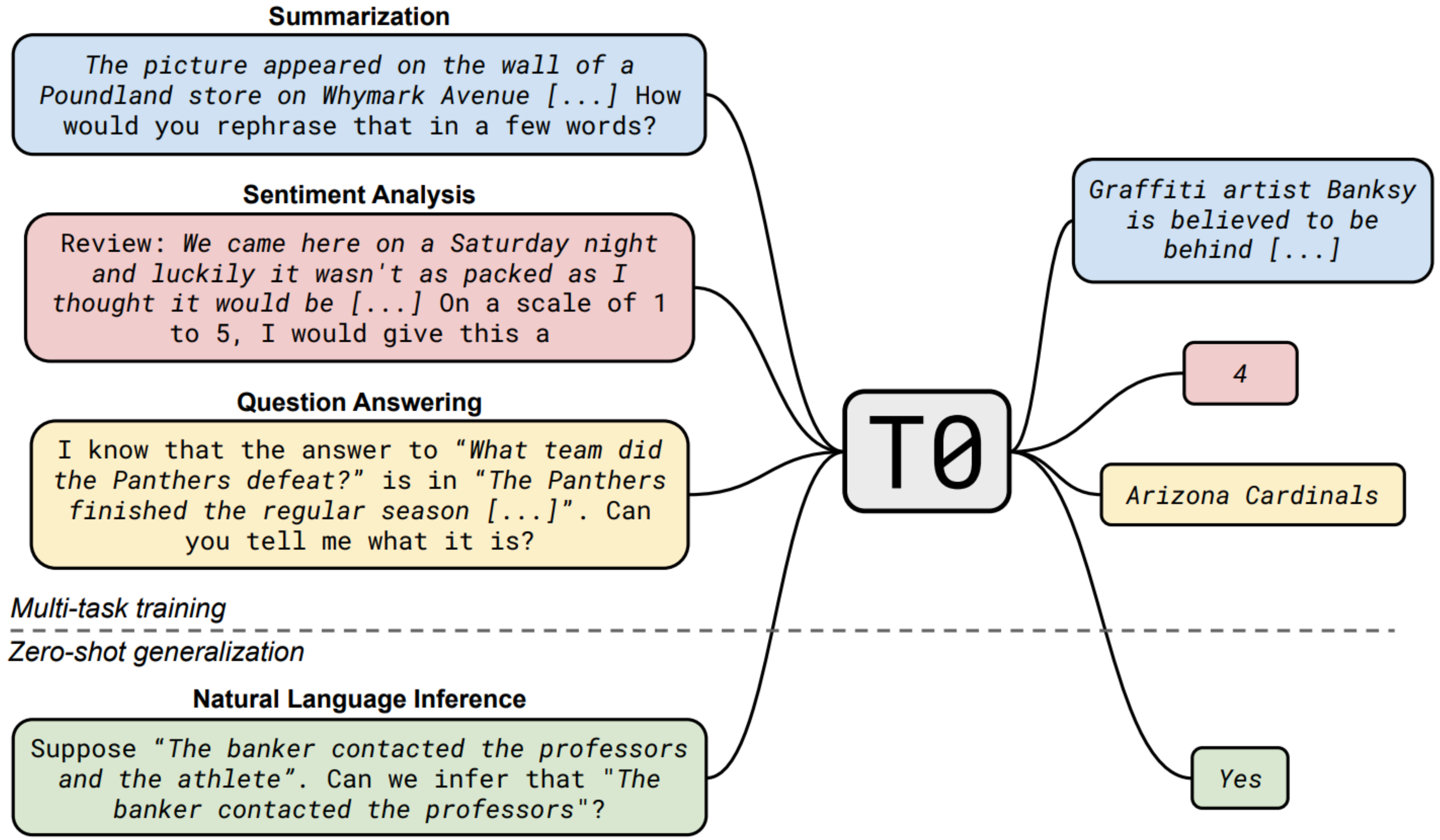


Figure 1: Our model and prompt format. T0 is an encoder-decoder model that consumes textual inputs and produces target responses. It is trained on a multitask mixture of NLP datasets partitioned into different tasks. Each dataset is associated with multiple prompt templates that are used to format example instances to input and target pairs. Italics indicate the inserted fields from the raw example data. After training on a diverse mixture of tasks (top), our model is evaluated on zero-shot generalization to tasks that are not seen during training (bottom).

Sanh et al. (2022)

Unified QA

Seen dataset?	Model ↓ - Evaluated on →	NewsQA	Quoref	Quoref-CS	ROPES	ROPES-CS	DROP	DROP-CS	QASC	Common senseQA	NP-BoolQ	BoolQ-CS	MultiRC	Avg
No	UnifiedQA [EX]	58.7	64.7	53.3	43.4	29.4	24.6	24.2	55.3	62.8	20.6	12.8	7.2	38.1
	UnifiedQA [AB]	58.0	**68.2**	57.6	48.1	41.7	30.7	36.8	54.1	59.0	27.2	39.9	28.4	45.8
	UnifiedQA [MC]	48.5	67.9	**58.0**	61.0	44.4	28.9	37.2	67.9	75.9	2.6	5.7	9.7	42.3
	UnifiedQA [YN]	0.6	1.7	1.4	0.0	0.7	0.4	0.1	14.8	20.8	79.1	78.6	**91.7**	24.2
	UnifiedQA	**58.9**	63.5	55.3	**67.0**	**45.5**	**32.5**	**40.1**	**68.5**	**76.2**	**81.3**	**80.4**	59.9	**60.7**
Yes	Previous best	66.8	86.1	55.4	61.1	32.5	89.1	54.2	85.2	79.1	78.4	71.1	--	
		Retro Reader	TASE	XLNet	ROBERTa	RoBERTa	ALBERT	MTMSN	KF+SIR+2Step	reeLB-RoBERT	RoBERTa	RoBERTa	--	

Table 4: Generalization to unseen datasets: Multi-format training (UNIFIEDQA) often outperforms models trained the same way but solely on other in-format datasets (e.g., UNIFIEDQA [EX], which is trained on all extractive training sets of UNIFIEDQA. When averaged across all evaluation datasets (last column), UNIFIEDQA shows strong generalization performance across all formats. Notably, the "Previous best" models (last row) were trained on the target dataset's training data, but are even then outperformed by UnifiedQA (which has never seen these datasets during training) on the YN tasks.

Khashabi et al. (2020)

Unifying Other NLP tasks as QA

- e.g. turn binary classification tasks into a “Yes"/“No" QA format

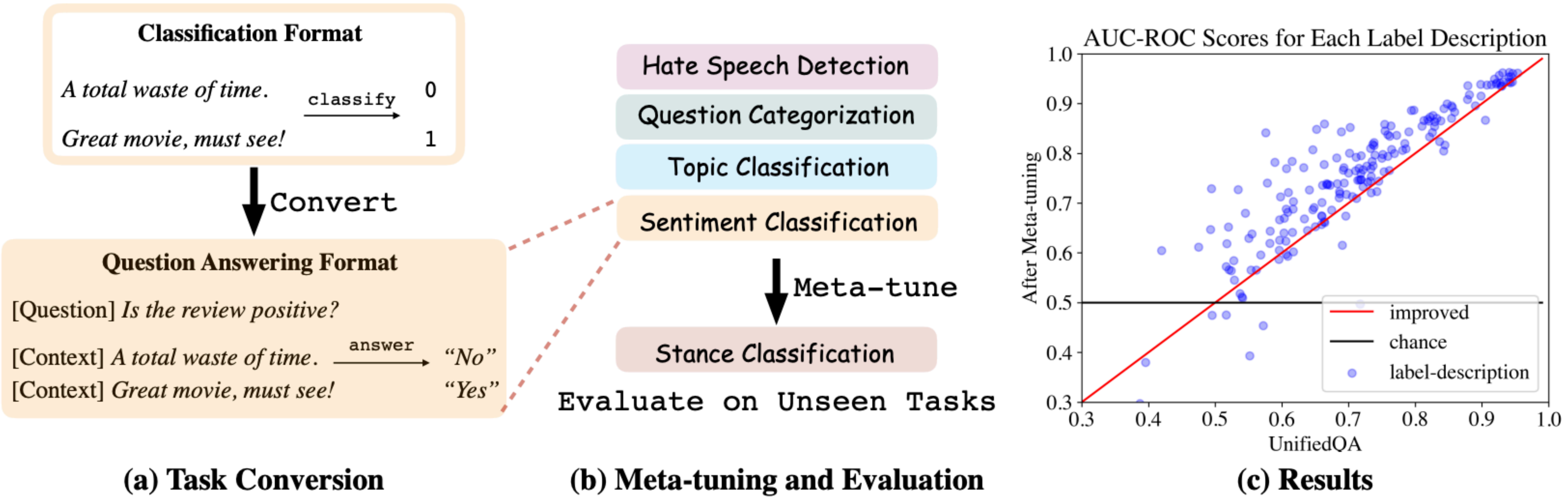


Figure 1: **(a)** We convert the format to question answering. We manually annotate label descriptions (questions) ourselves (Section 2). **(b)** We finetune the UnifiedQA (Khashabi et al., 2020) model (with 770 M parameters) on a diverse set of tasks (Section 4), and evaluate its 0-shot classification (ZSC) performance on an unseen task. **(c)** For each label description (question) we evaluate the AUC-ROC score for the “Yes" answer, and each dot represents a label description (Section 3). The x-value is the ZSC performance of UnifiedQA; the y-value is the performance after meta-tuning. In most cases, the y-value improves over the x-value (above the red line) and is better than random guesses (above the black line) by a robust margin (Section 5).

Zhong et al. (2021)

Unifying Other NLP tasks as QA

Are these two questions asking for the same thing?

Does the tweet contain irony?

Is this news about world events?

Does the text contain a definition?

Is the tweet an offensive tweet?

Is the text objective?

Does the question ask for a numerical answer?

Is the tweet against environmentalist initiatives?

Is this abstract about Physics?

Does the tweet express anger?

Does the user dislike this movie?

Is the sentence ungrammatical?

Zhong et al. (2021)

Flan

- Pre-train, then fine-tune on a bunch of tasks, generalize to unseen tasks
- Scaling the number of tasks, models size (Flan-T5, Flan-Palm), and fine-tuning on chain-of-thought data

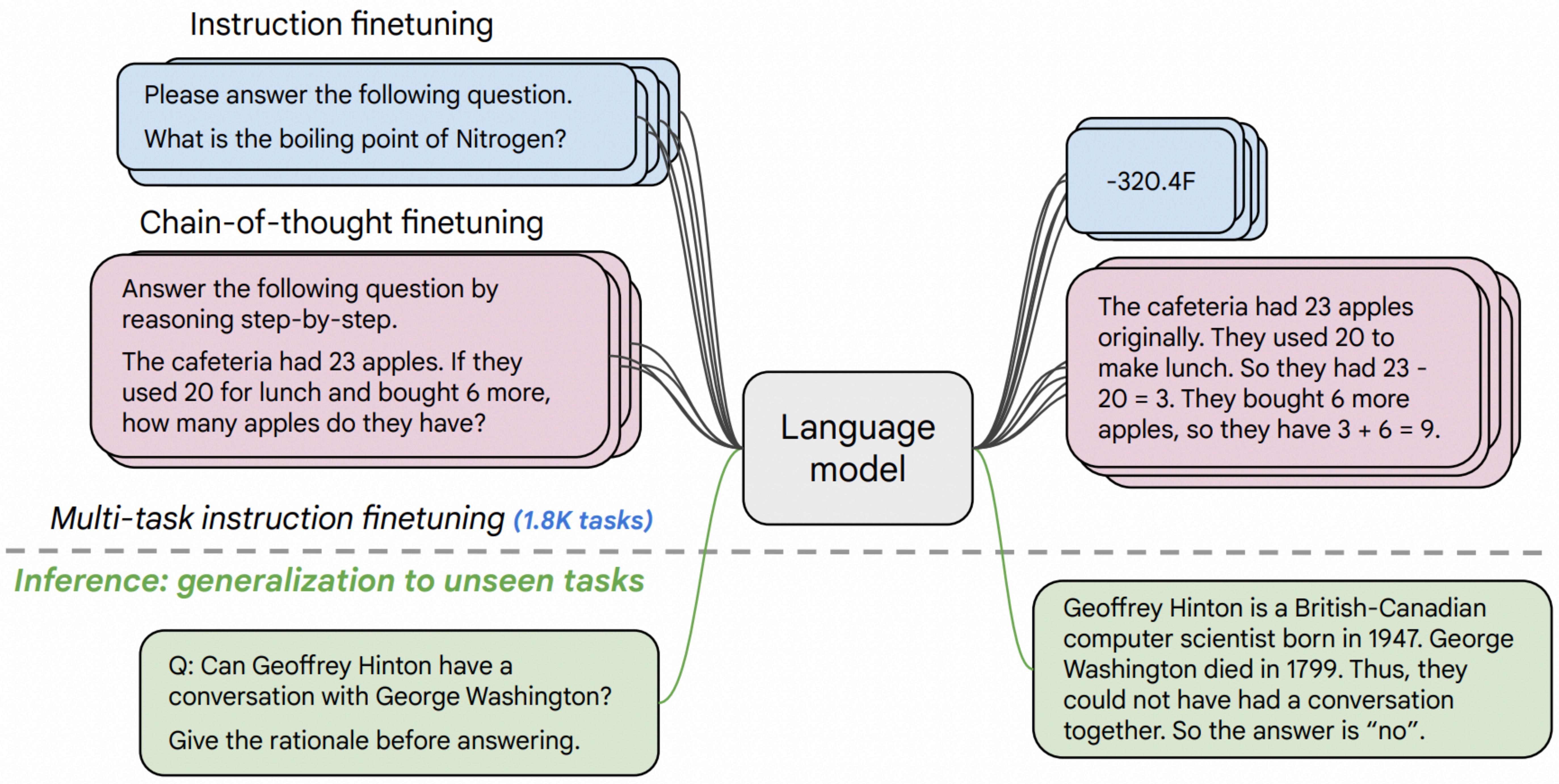
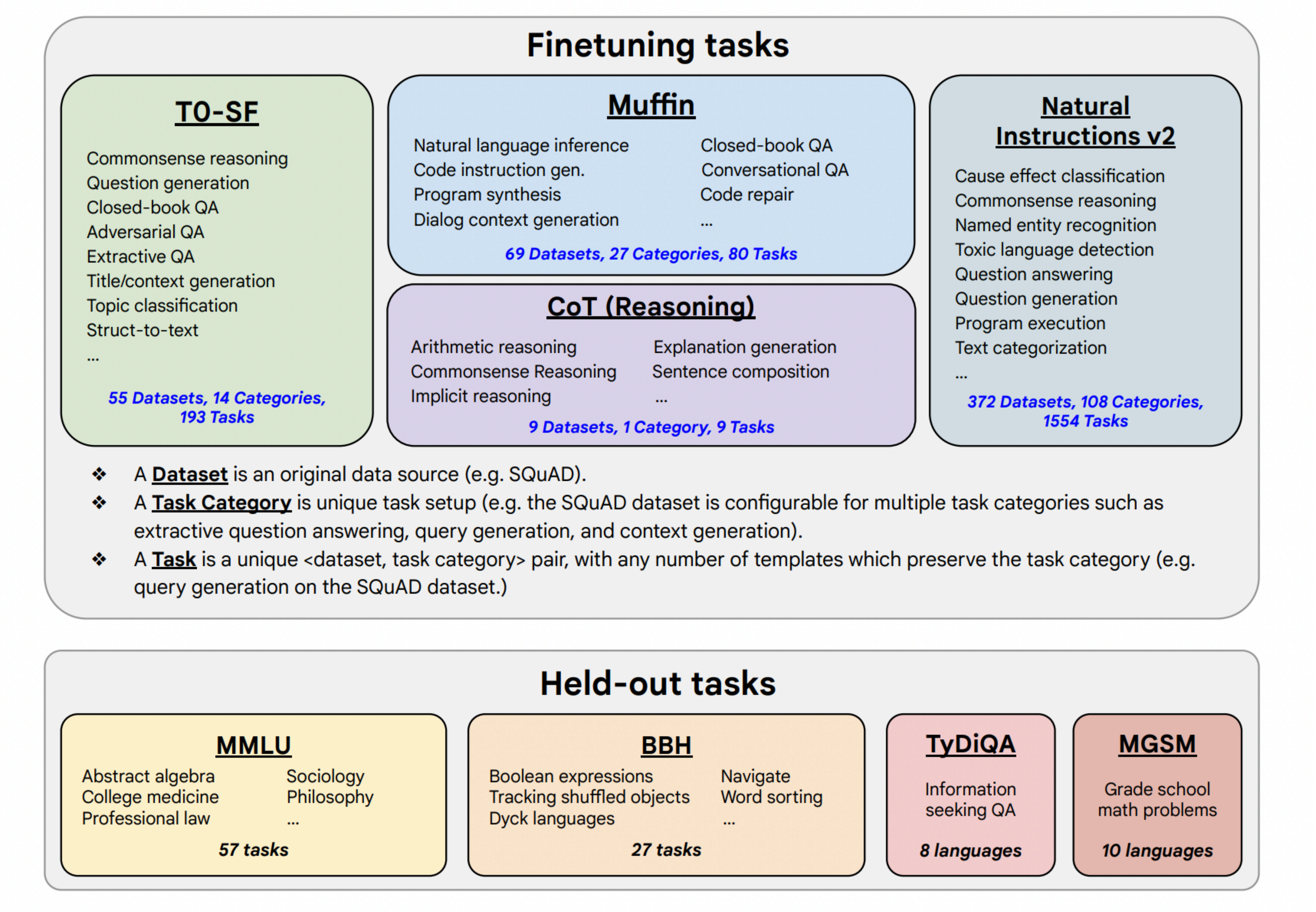


Figure 1: We finetune various language models on 1.8K tasks phrased as instructions, and evaluate them on unseen tasks. We finetune both with and without exemplars (i.e., zero-shot and few-shot) and with and without chain-of-thought, enabling generalization across a range of evaluation scenarios.

Chung et al. (2022)

Flan

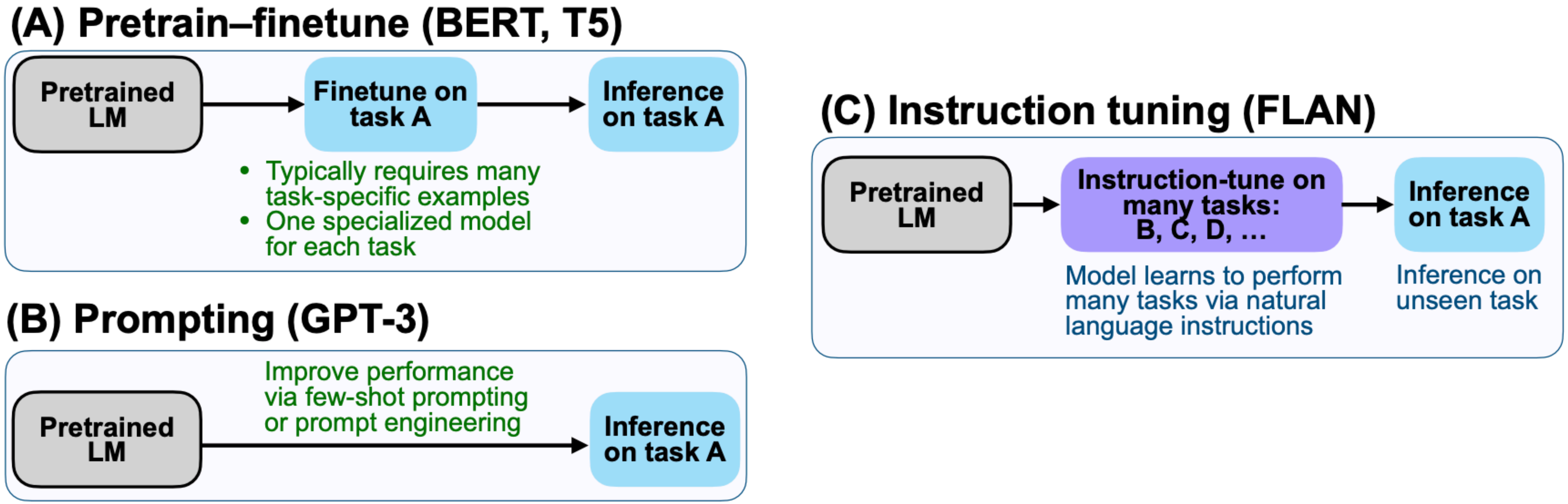


- Fine-tuned on 473 datasets, 1836 tasks.
- Some datasets support multiple tasks
- E.g. SQuAD can be used for QA or question generation.

Chung et al. (2022)

Flan

- Instruction fine-tuning can be done on various models (PaLM, T5, etc.)
- Flan-T5 models publicly available



Chung et al. (2022)

Flan

- Instruction fine-tuning can be done on various models (PaLM, T5, etc.)
- Flan-T5 models publicly available

Params	Model	Arhitecture	pre-training Objective	Pretrain FLOPs	Finetune FLOPs	% Finetune Compute
80M	Flan-T5-Small	encoder-decoder	span corruption	1.8E+20	2.9E+18	1.6%
250M	Flan-T5-Base	encoder-decoder	span corruption	6.6E+20	9.1E+18	1.4%
780M	Flan-T5-Large	encoder-decoder	span corruption	2.3E+21	2.4E+19	1.1%
3B	Flan-T5-XL	encoder-decoder	span corruption	9.0E+21	5.6E+19	0.6%
11B	Flan-T5-XXL	encoder-decoder	span corruption	3.3E+22	7.6E+19	0.2%
8B	Flan-PaLM	decoder-only	causal LM	3.7E+22	1.6E+20	0.4%
62B	Flan-PaLM	decoder-only	causal LM	2.9E+23	1.2E+21	0.4%
540B	Flan-PaLM	decoder-only	causal LM	2.5E+24	5.6E+21	0.2%
62B	Flan-cont-PaLM	decoder-only	causal LM	4.8E+23	1.8E+21	0.4%
540B	Flan-U-PaLM	decoder-only	prefix LM + span corruption	2.5E+23	5.6E+21	0.2%

Table 2: Across several models, instruction finetuning only costs a small amount of compute relative to pre-training. T5: Raffel et al. (2020). PaLM and cont-PaLM (also known as PaLM 62B at 1.3T tokens): Chowdhery et al. (2022). U-PaLM: Tay et al. (2022b).

Chung et al. (2022)

Takeaways

- Lots of problems with current QA settings, lots of new datasets
- QA over tables, images, knowledge bases, …
- ~~Models can often work well for one QA task but don’t generalize~~
- There’s lots that we can’t do, but we’re getting really good at putting our hands on random facts from the Internet
- Cross-lingual and multilingual QA …