Lecture 15: Question Answering (Adversarial, Retrieval, Multi-Hop)

Alan Ritter

(many slides from Greg Durrett)

Recall: 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

One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

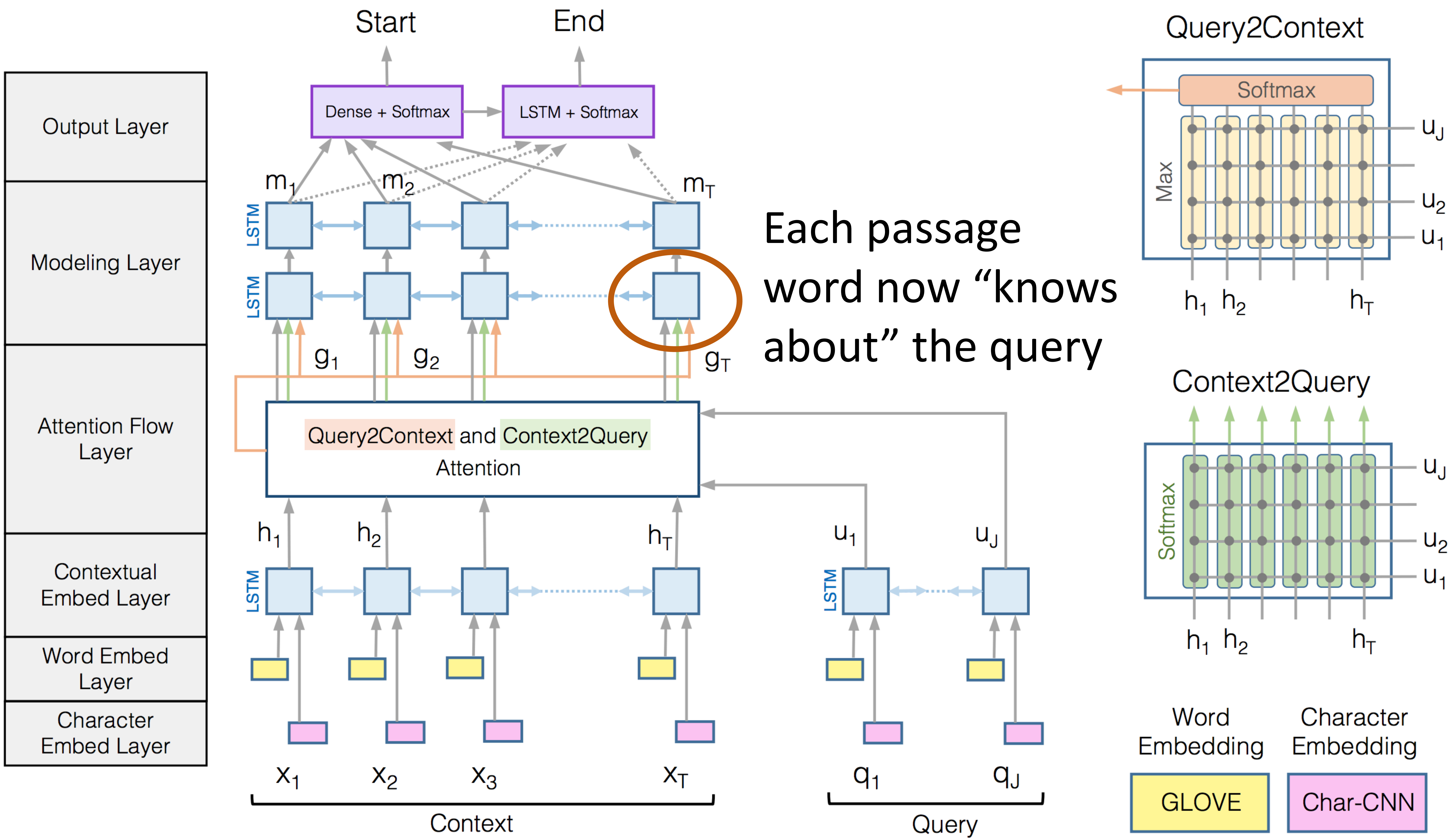
What was Maria Curie the first female recipient of?
Ground Truth Answers: Nobel Prize | Nobel Prize | Nobel Prize

What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 | 1745 | 1745

Who was one of the most famous people born in Warsaw?
Ground Truth Answers: Maria Skłodowska-Curie | Maria Skłodowska-Curie | Maria Skłodowska-Curie

Rajpurkar et al. (2016)

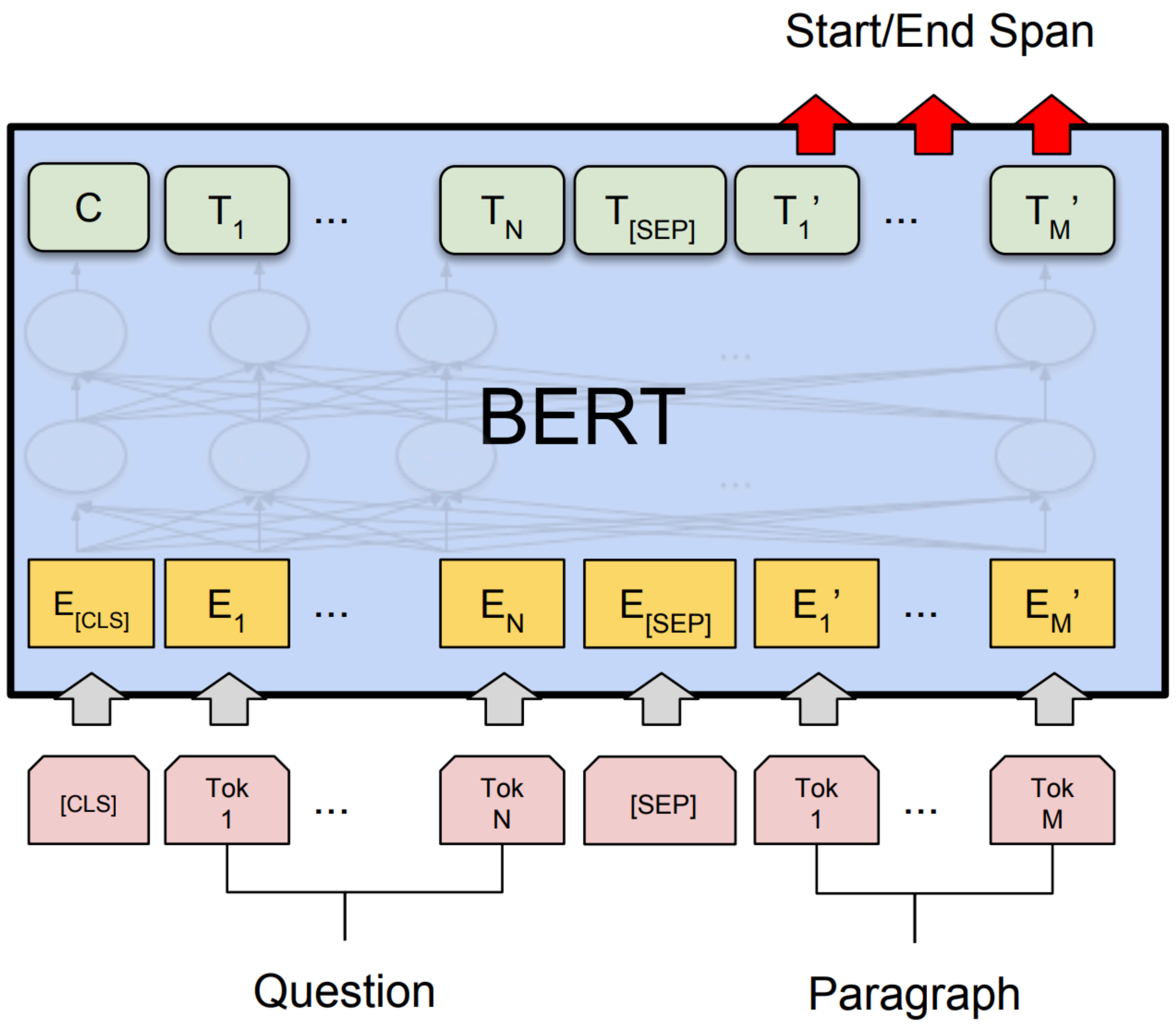
Recall: Bidirectional Attention Flow



Each passage word now "knows about" the query

Seo et al. (2016)

Recall: QA with BERT

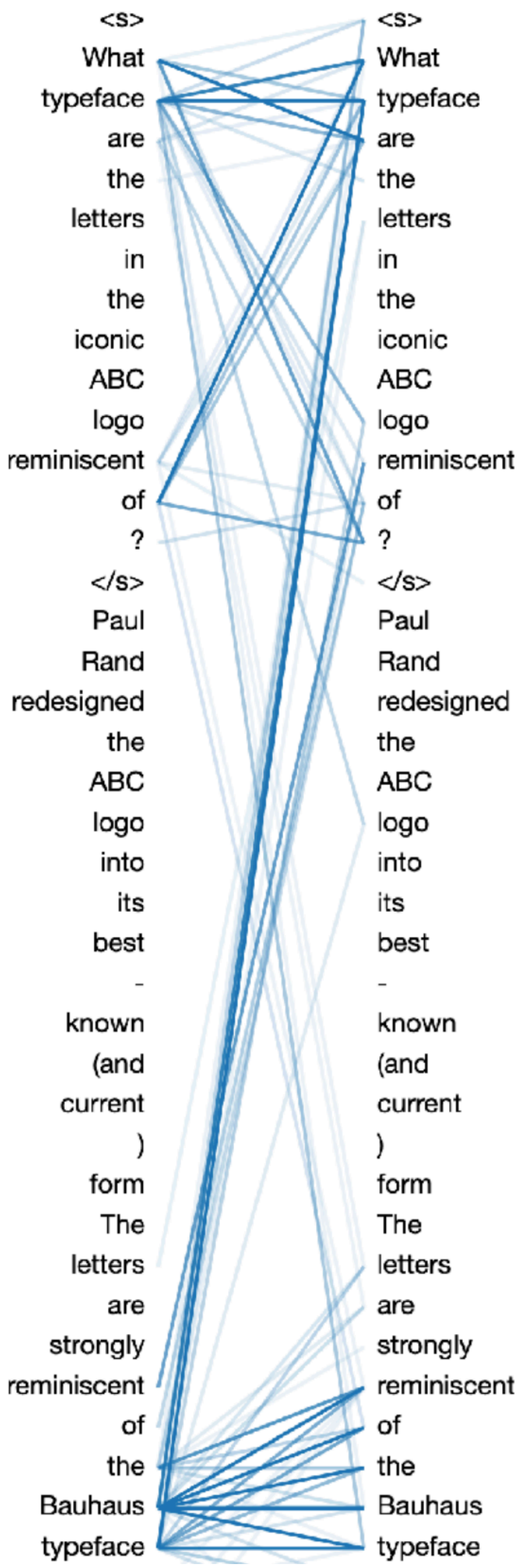


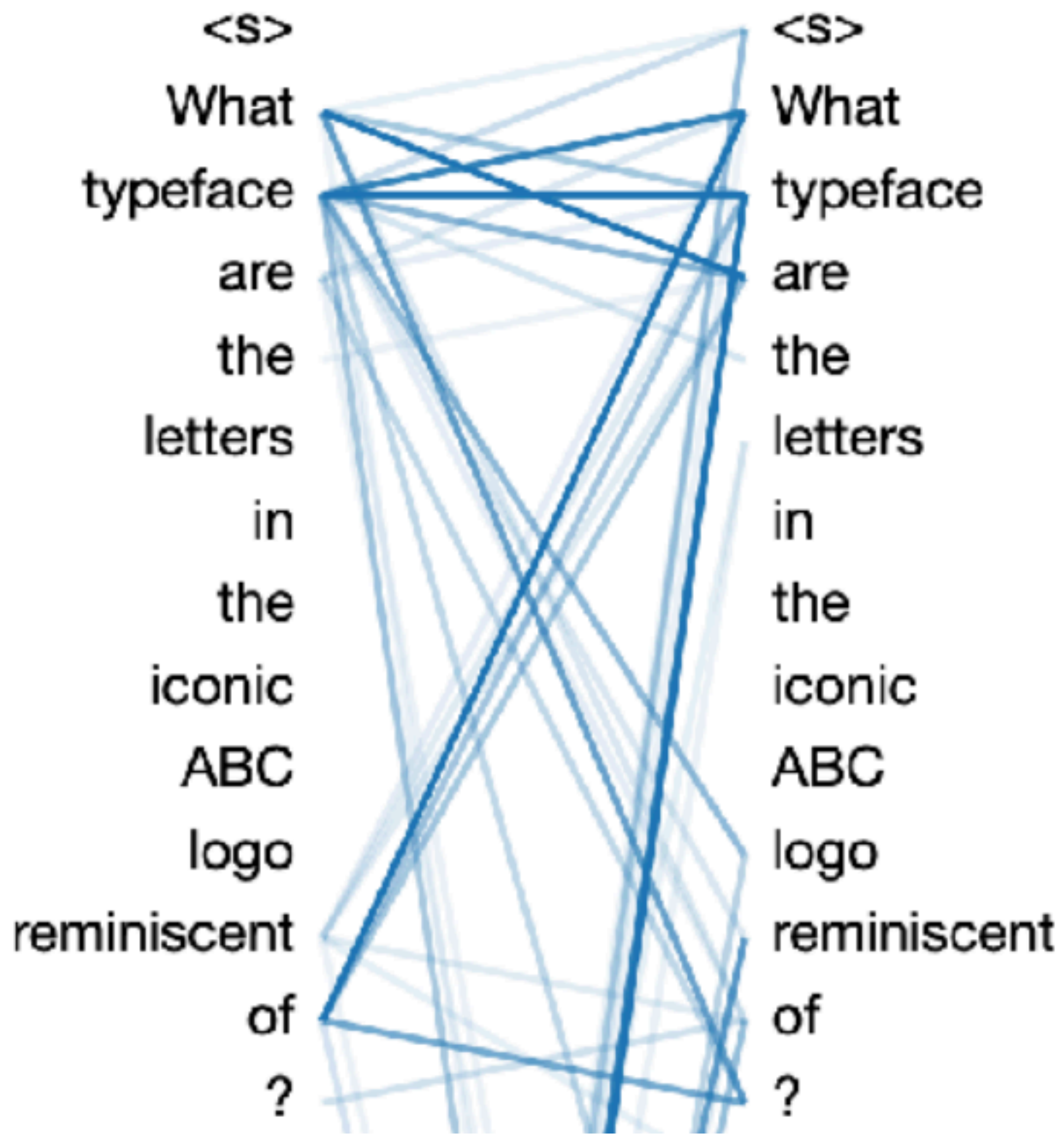
What was Marie Curie the first female recipient of ? [SEP] One of the most famous people born in Warsaw was Marie …

- Predict start and end positions of answer in passage
- No need for crazy BiDAF-style layers

Devlin et al. (2019)

Recall: What are these models learning?





ABC isn't used at all! The model is mostly using the fact that only one typeface is in the context

Ye, Nair, Durrett (2021)

This Lecture

- Problems in QA, especially related to answer type overfitting
- Retrieval-based QA / multi-hop QA
- New QA frontiers

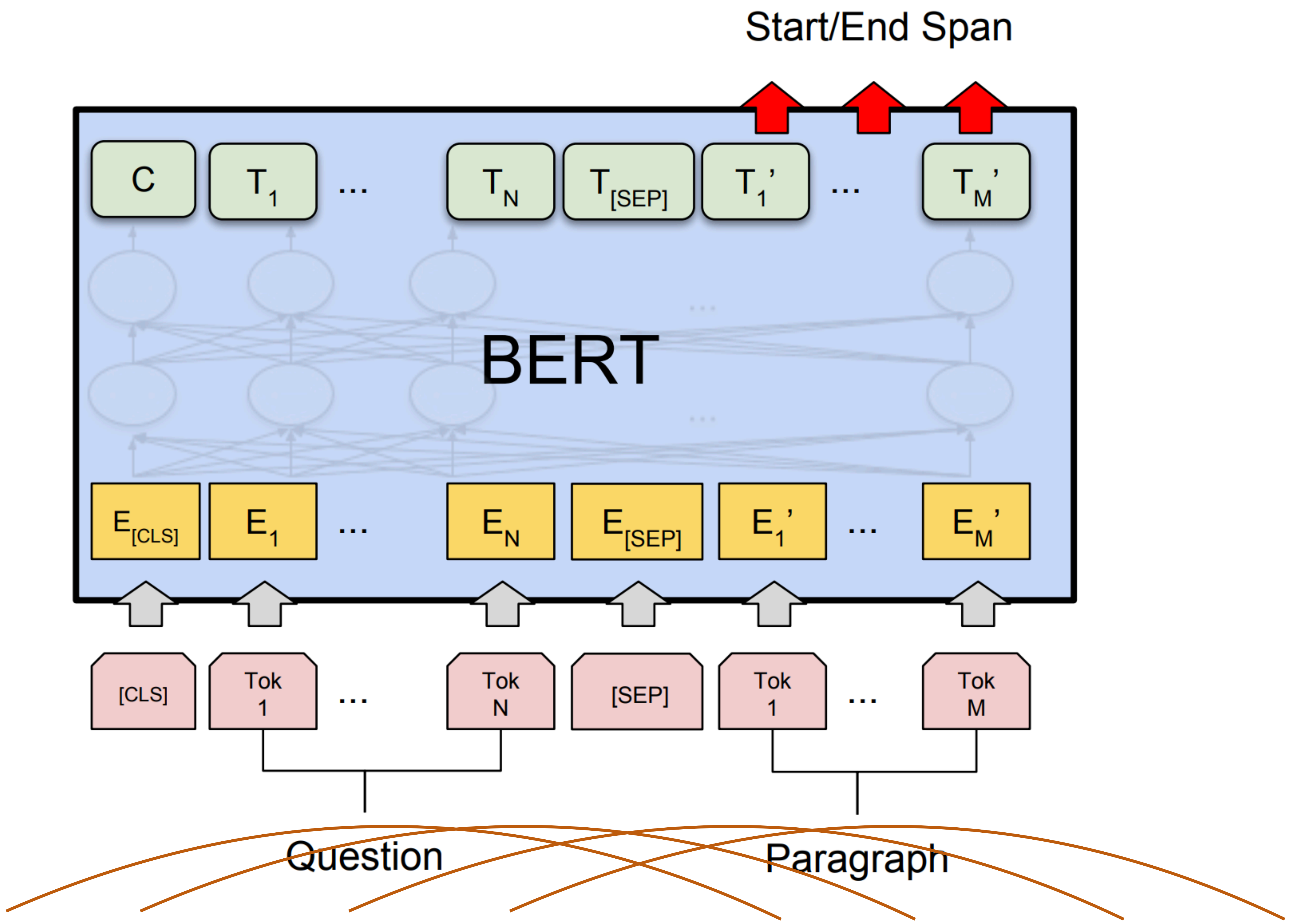
Problems in QA

Adversarial SQuAD

- SQuAD questions are often easy: “*what was she the recipient of?*” passage: “*… recipient of Nobel Prize…*”

Jia and Liang (2017)

Adversarial SQuAD

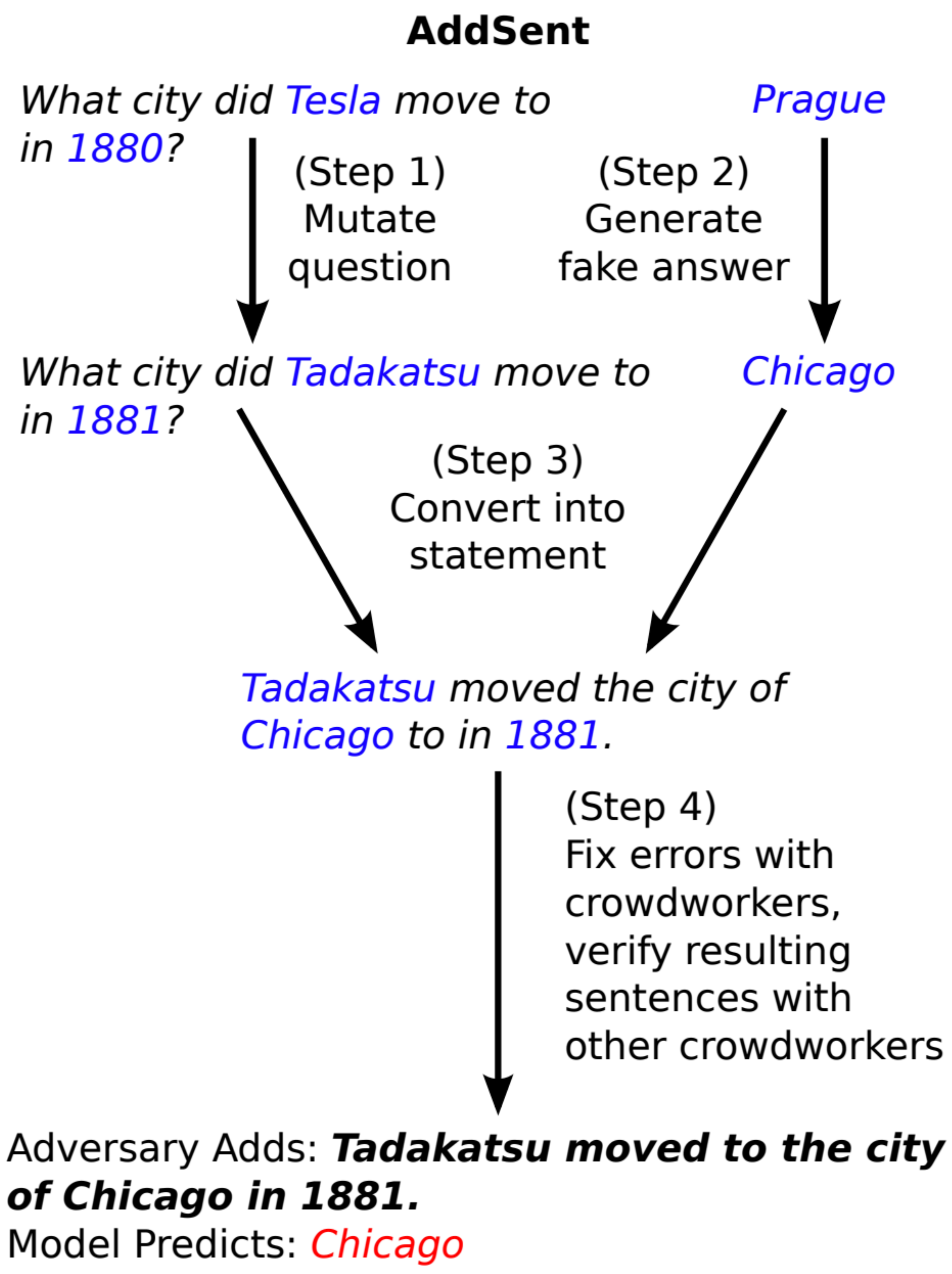


What was Marie Curie the first female recipient of ? [SEP] … first female recipient of **the Nobel Prize** …

- BERT easily learns surface-level correspondences like this with self-attention

Adversarial SQuAD

- SQuAD questions are often easy: “*what was she the recipient of?*” passage: “*… recipient of Nobel Prize…*”
- Can we make them harder by adding a *distractor* answer in a very similar context?
- Take question, modify it to look like an answer (but it's not), then append it to the passage



Jia and Liang (2017)

Adversarial SQuAD

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

- Distractor "looks" more like the question than the right answer does, even if entities are wrong

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"

Input (underline = correct span, **red** = trigger, **underline** = target span)	
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 instead add the same adversary to *every* passage
- Adding "*why how because to kill american people*" causes SQuAD models to return this answer 10-50% of the time when given a "why" question
- Similar attacks on other question types like "who"

Wallace et al. (2019)

How to fix QA?

- Better models?
  - But a model trained on weak data will often still be weak to adversaries
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
- Better datasets
  - Same questions but with more distractors may challenge our models
  - Next up: *retrieval-based* QA models
- Harder QA tasks
  - Ask questions which *cannot* be answered in a simple way
  - Afterwards: *multi-hop* QA and other QA settings

Retrieval Models

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
- 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)

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)

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)

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:

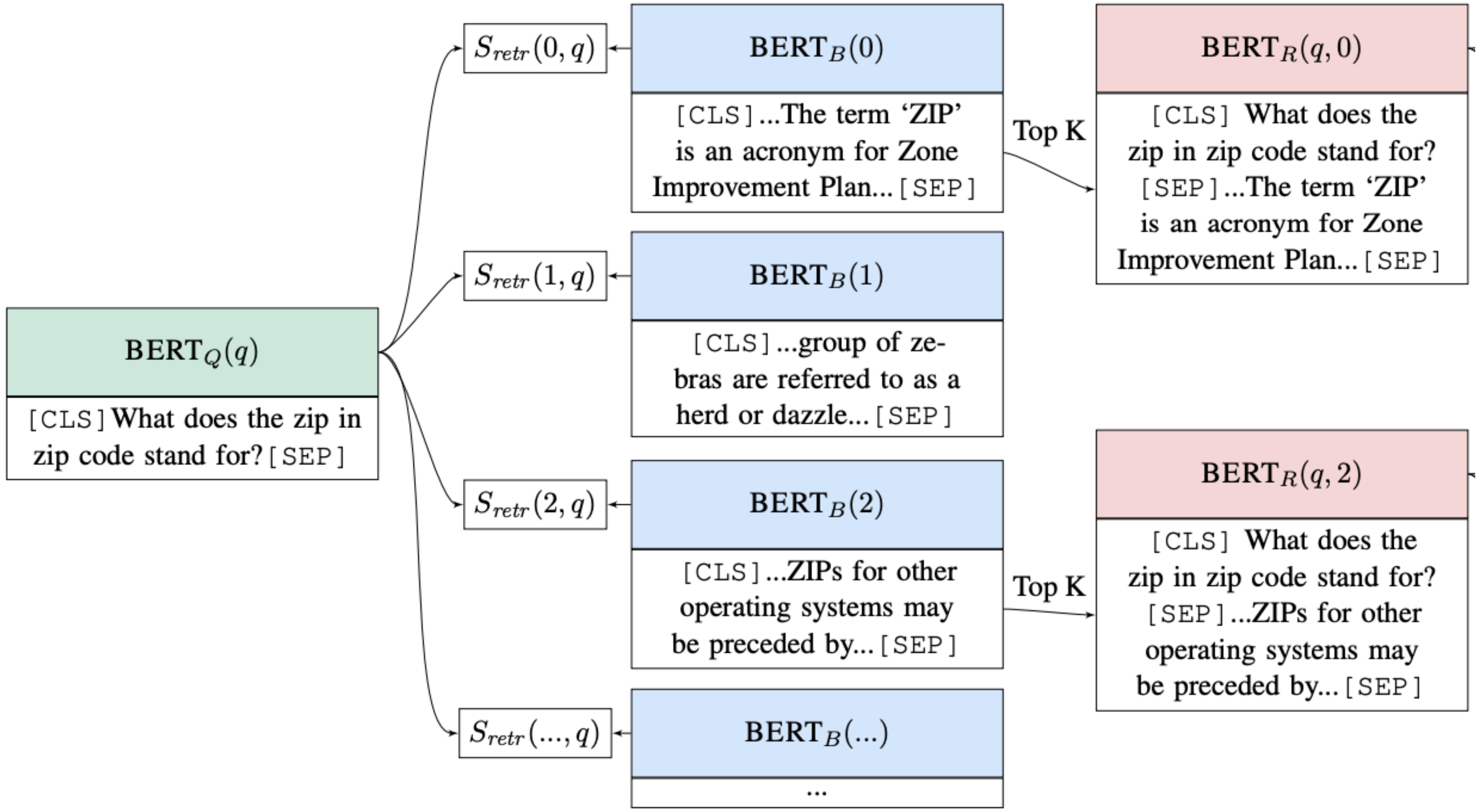
From the right ventricle , blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung) , which branch into smaller pulmonary arteries that spread throughout the lungs.

- 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

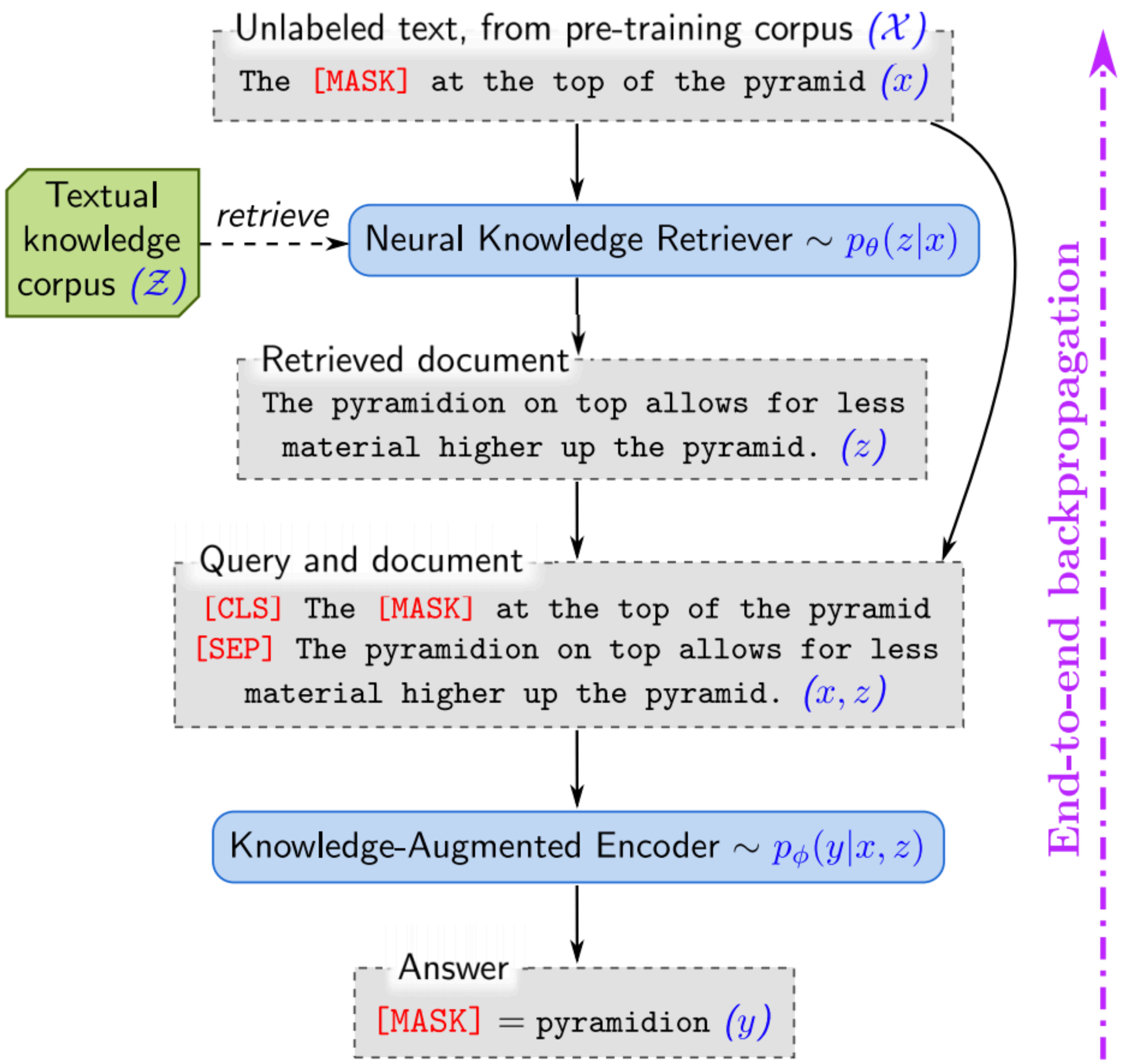


$$h_q = \mathbf{W_q}\mathrm{BERT}_Q(q)[\mathtt{CLS}]$$
$$h_b = \mathbf{W_b}\mathrm{BERT}_B(b)[\mathtt{CLS}]$$
$$S_{retr}(b, q) = h_q^\top h_b$$

Lee et al. (2019)

REALM

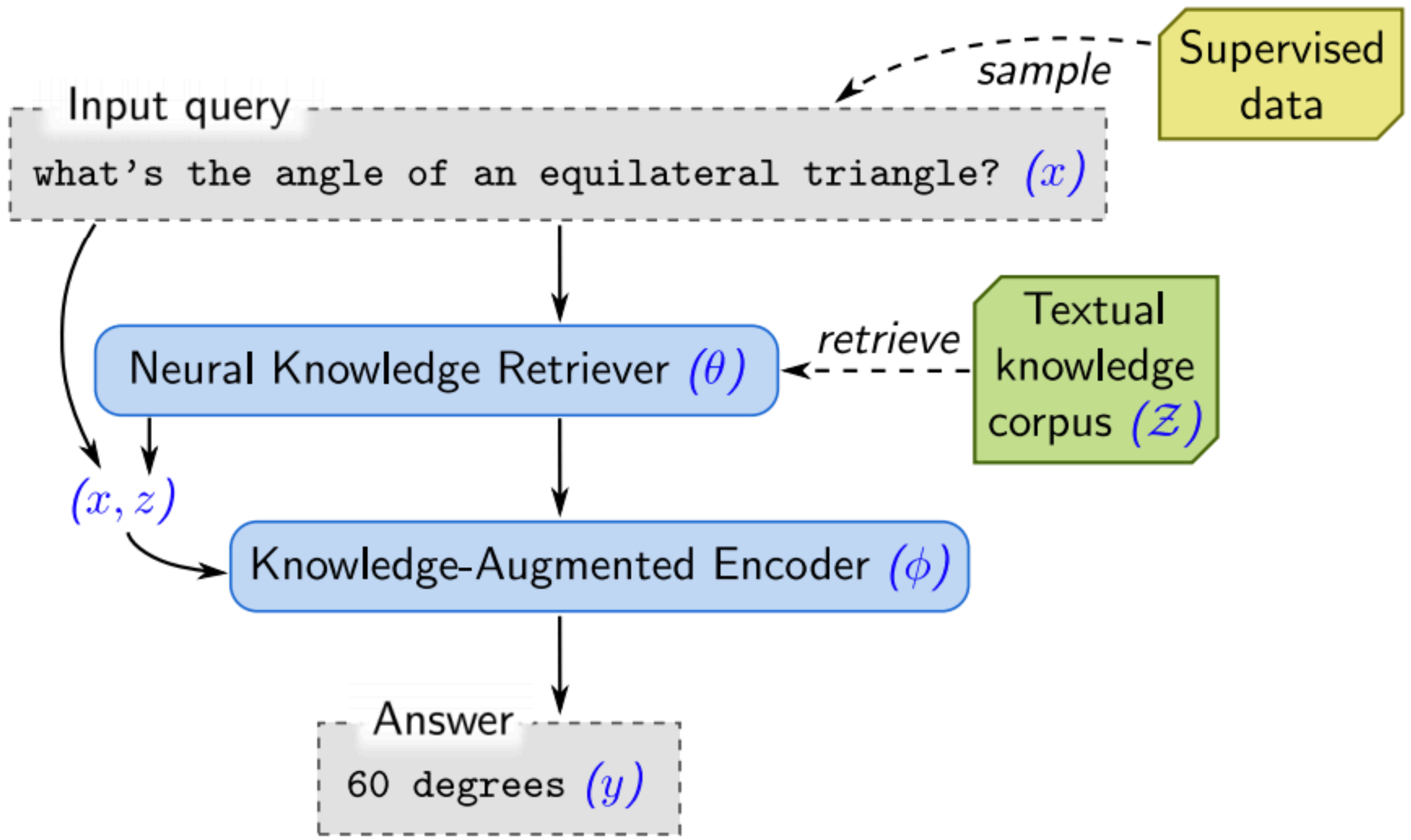
- Technique for integrating retrieval into pre-training
- Retriever relies on a maximum inner-product search (MIPS) over BERT embeddings
- MIPS is fast — challenge is how to refresh the BERT embeddings



Guu et al. (2020)

REALM

- 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)

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
- A model shouldn't be able to answer these without doing some reasoning about the intermediate entity

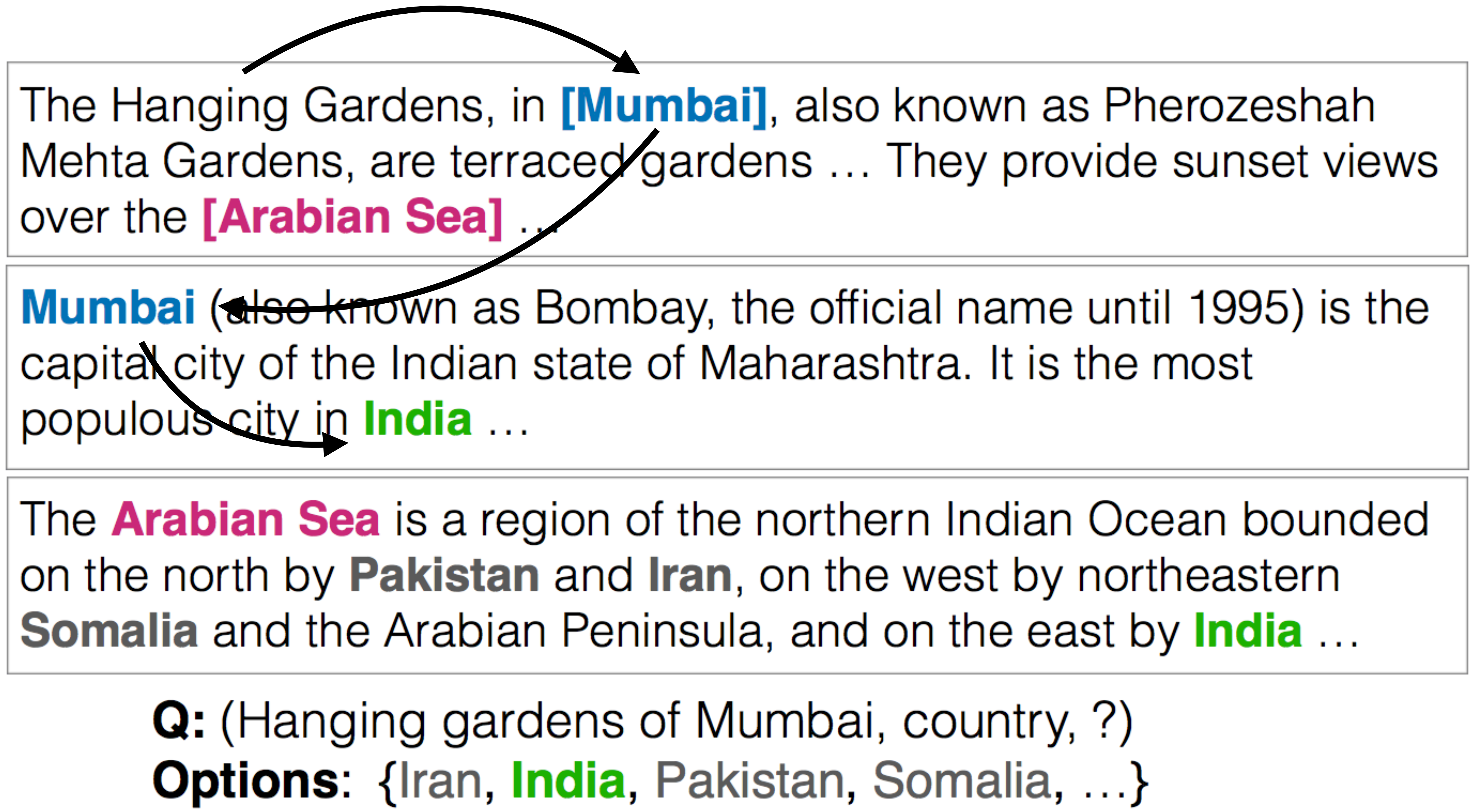


Figure from Welbl et al. (2018)

HotpotQA

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

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

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 fam… …mil… …at is famous for its involvement in hotels, name… …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**: 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]

Results on WikiHop

Dataset	WikiHop-MC
Metric	Accuracy
NoContext	59.70
MC-BiDAF++	61.32
MC-MemNet	61.80
Span2MC-BiDAF++	59.85

More than half of questions can be answered without even using the context!

- SOTA models trained on this **may** be learning question-answer correspondences, not multi-hop reasoning as advertised

Chen and Durrett (2019)

Results on HotpotQA

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

A simple single sentence reasoning model can solve more than half questions on HotpotQA.

Chen and Durrett (2019)

Other Work

- Min et al. ACL 2019 “Compositional Questions do not Necessitate Multi-hop Reasoning”
  - Focuses just on HotpotQA
  - Additionally tries to adversarially harden Hotpot against these attacks. Some limited success, but doesn't solve the problem

New Types of QA

DROP

- One thread of research: let's build QA datasets to help the community focus on modeling particular things

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 (structure the model around predicting differences between numbers)

Dua et al. (2019)

MultiQA

- Maybe we should just look at lots of QA datasets instead?

	CQ	CWQ	ComQA	WikiHop	DROP	SQuAD	NewsQA	SearchQA	TQA-G	TQA-W	HotpotQA
SQuAD	23.6	12.0	20.0	4.6	5.5	-	**31.8**	8.4	37.8	33.4	**11.8**
NewsQA	24.1	12.4	18.9	7.1	4.4	**60.4**	-	10.1	37.6	28.4	8.0
SearchQA	30.3	18.5	25.8	12.4	2.8	23.3	12.7	-	**53.2**	**35.4**	5.2

…

- BERT trained on SQuAD gets <40% performance on any other QA dataset
- Our QA models are pretty good at fitting single datasets with 50k-100k examples, but still aren't learning general question answering

Talmor and Berant (2019)

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 in 2 years

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)

Takeaways

- Lots of problems with current QA settings, lots of new datasets
- Models can often work well for one QA task but don’t generalize
- We still don’t have (solvable) QA settings which seem to require really complex reasoning as opposed to surface-level pattern recognition