THE OHIO STATE UNIVERSITY

CSE 5525: Foundations of Speech and Language Processing

Question Answering I

Huan Sun (CSE@OSU)

Many thanks to Prof. Greg Durrett @ UT Austin for sharing his slides.
Some images/examples were from the two textbooks by (1) Jurafsky and Martin and (2) Eisenstein.

Previously, QA as Semantic Parsing

- Many ways to build these parsers
- One approach: run a "supertagger" (tagging the sentence with complex labels), then run the parser

What	states	border	Texas
$(S/(S\backslash NP))/N$	N	$(S\backslash NP)/NP$	NP
$\lambda f.\lambda g.\lambda x.f(x) \wedge g(x)$	$\lambda x.state(x)$	$\lambda x.\lambda y.borders(y,x)$	$texas$

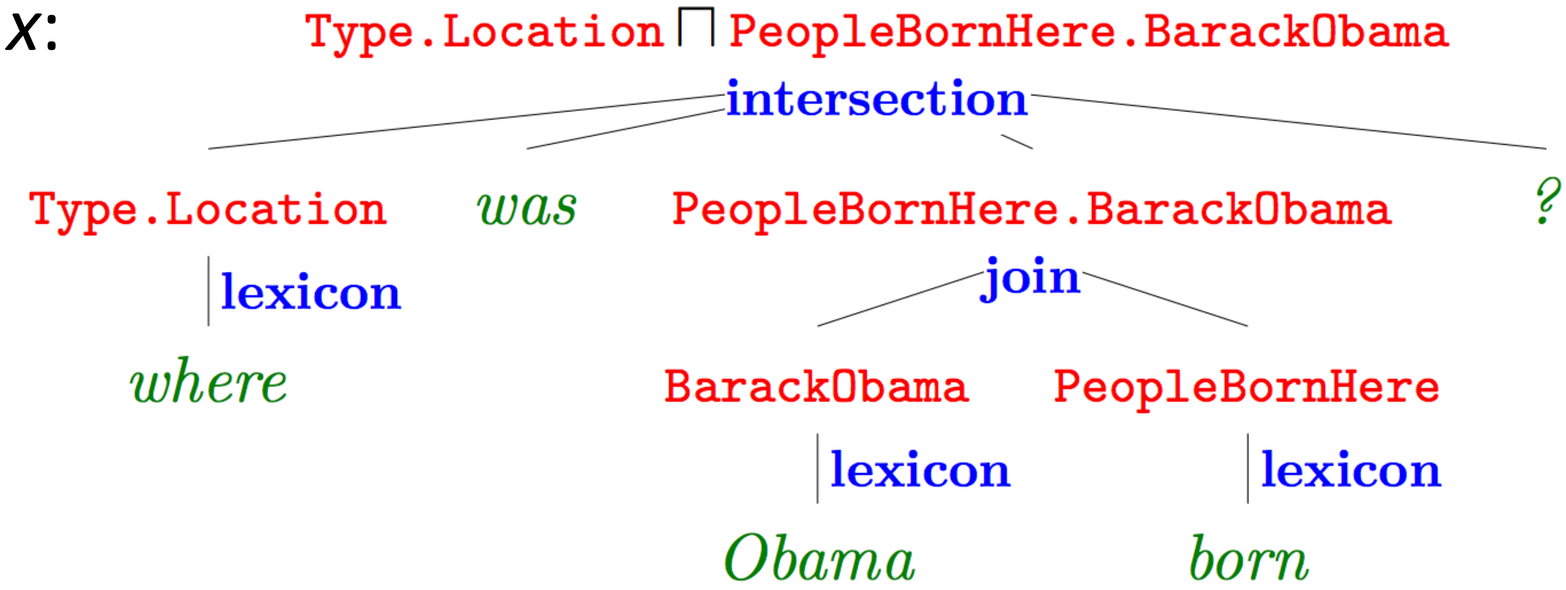
  - Parsing is easy once you have the tags, so we've reduced it to a (hard) tagging problem

CCG parsing

Zettlemoyer and Collins (2005)

Parsing into Lambda-DCS

- Derivation *d* on sentence *x*:



- No more explicit syntax in these derivations like we had in CCG

- Building the lexicon: more sophisticated process than GENLEX, but can handle thousands of predicates

- Log-linear model with features on rules: $P(d|x) \propto \exp w^\top \left(\sum_{r \in d} f(r, x) \right)$

- Similar to CRF parsers

Berant et al. (2013)

This Lecture

- Types of question answering/reading comprehension
- CNN/Daily Mail task: Attentive Reader
- SQuAD task: Bidirectional Attention Flow

Question Answering

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

Q: *where was Barack Obama born*

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

(also Prolog / GeoQuery, etc.)

Question Answering

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

Q: *where was Barack Obama born*

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

(also Prolog / GeoQuery, etc.)

- When do we formulate QA as a semantic parsing task?

QA is very broad

- Factoid QA: *what states border Mississippi?*, *when was Barack Obama born?*
  - Lots of this could be handled by QA over a knowledge base, if we had a big enough knowledge base
- "Question answering": many types; very broad
  - *Is P=NP?*
  - *What is 4+5?*
  - *What is the translation of [sentence] into French?* [McCann et al., 2018]

QA is very broad

- Factoid QA: *what states border Mississippi?*, *when was Barack Obama born?*
  - Lots of this could be handled by QA over a knowledge base, if we had a big enough knowledge base
- From the point of data source's view:
  - Based on knowledge bases
  - Based on texts and tables (unstructured, or semi-structured data)
  - Community QA (such as Stack Overflow, WebMD, etc)

What are the limits of QA?

- Focus on questions where the answer might plausibly appear in text… but this is Still TOO BROAD
  - *What were the main causes of World War II?* — requires summarization
  - *Can you get the flu from a flu shot?* — want IR to provide an explanation of the answer, not just yes/no
  - *What temperature should I cook chicken to?* — could be written down in a KB but probably isn't

What are the limits of QA?

- Focus on questions where the answer might plausibly appear in text… but this is still too broad
  - *What were the main causes of World War II?* — requires summarization
  - *Can you get the flu from a flu shot?* — want IR to provide an explanation of the answer, not just yes/no
  - *What temperature should I cook chicken to?* — could be written down in a KB but probably isn't
  - Today: can we do QA when it requires retrieving the answer from a passage?

Reading Comprehension

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

- Questions per passage explicitly require cross-sentence reasoning

Richardson (2013)

Baselines

- N-gram matching: append question + each answer, return answer which gives highest n-gram overlap with a sentence
- Parsing: find direct object of “pulled” in the document where the subject is James

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.

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

A) pudding
B) fries
C) food
D) splinters

- Don’t need any complex semantic representations

Richardson (2013)

Reading Comprehension

ngram sliding window

	MC160 Test	MC500 Test
Baseline (SW+D)	66.25	56.67
RTE	59.79‡	53.52
Combined	67.60	60.83‡

- Classic textual entailment systems don't work as well as n-grams
- Scores are low partially due to questions spanning multiple sentences
- Unfortunately not much data to train better methods on (2000 questions)

Richardson (2013)

Dataset Explosion

- 30+ QA datasets released since 2015
- Question answering: questions are in natural language
  - Answers: multiple choice, require picking from the passage, or generate freeform answer (last is pretty rare)
  - Require human annotation

Dataset Explosion

- 30+ QA datasets released since 2015
- Question answering: questions are in natural language
  - Answers: multiple choice, require picking from the passage, or generate freeform answer (last is pretty rare)
  - 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

Dataset Properties

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

You can learn to analyze datasets along these axes

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)

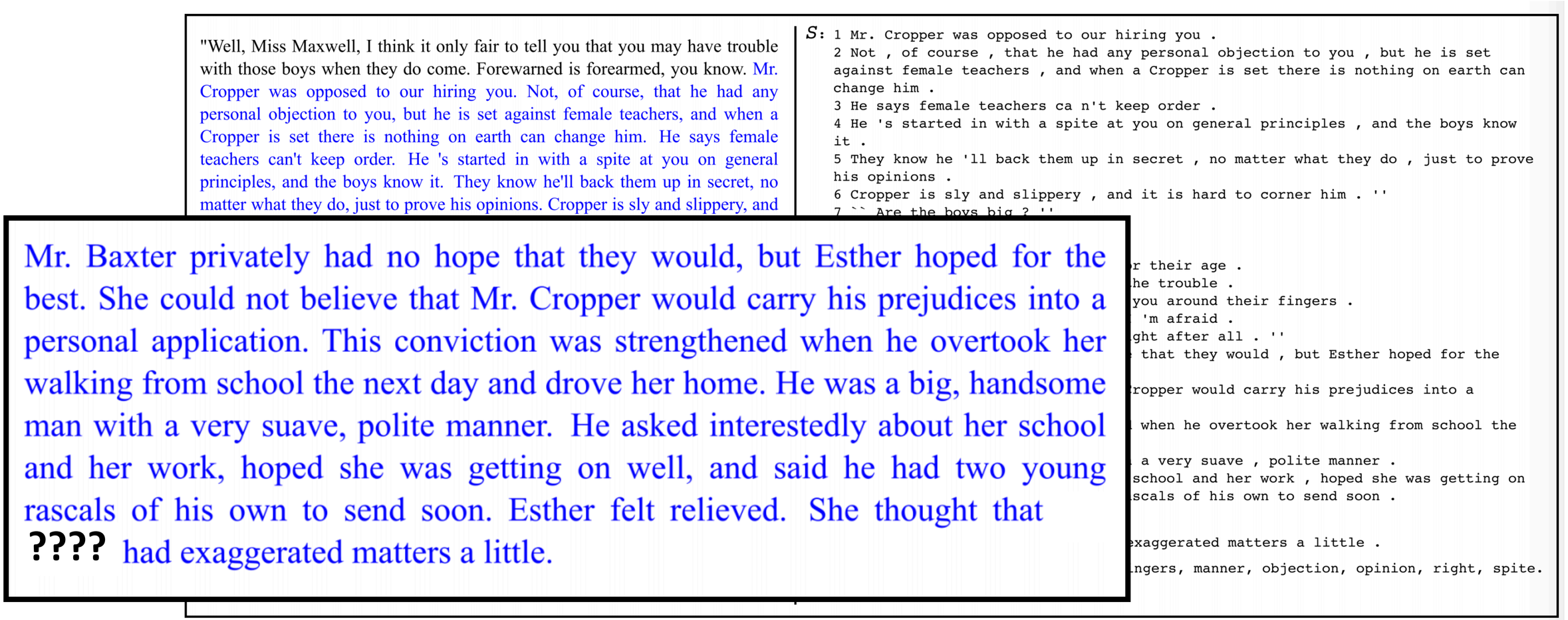
You can learn to analyze datasets along these axes

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?

You can learn to analyze datasets along these axes

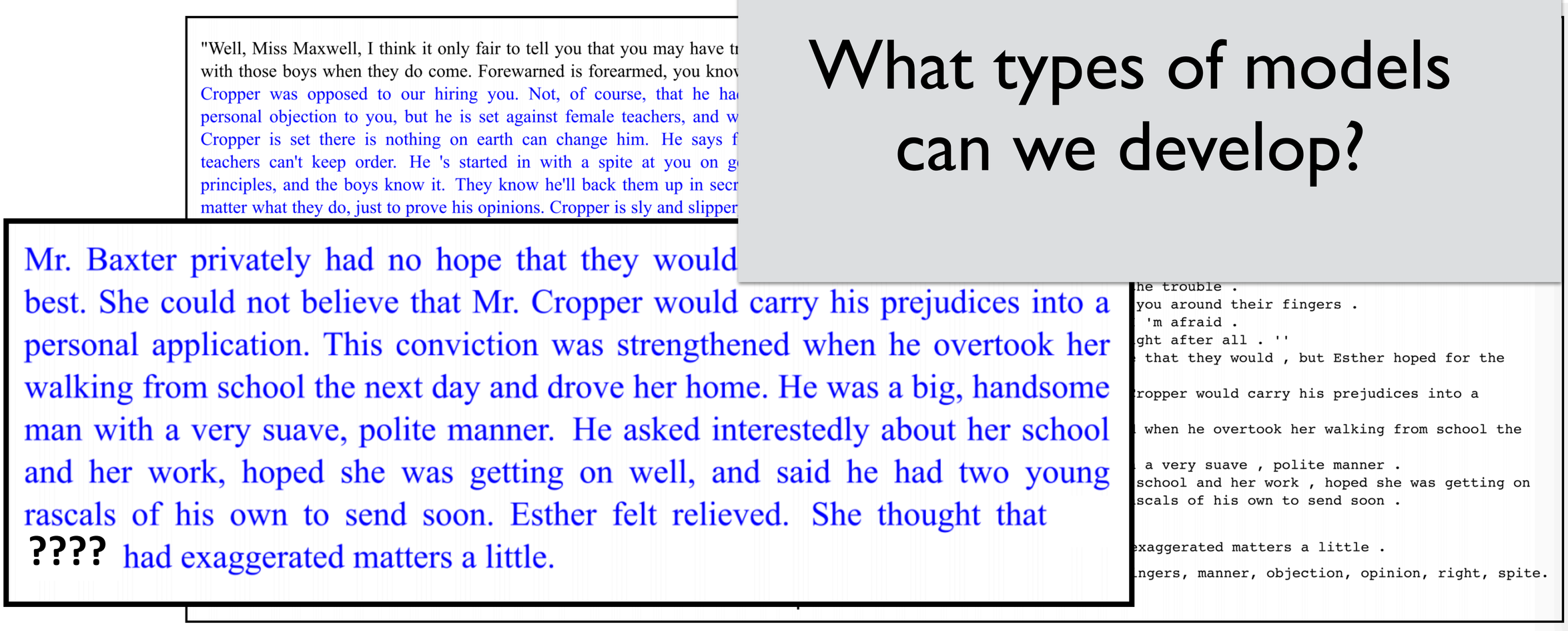
Children's Book Test



- 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)

Children's Book Test



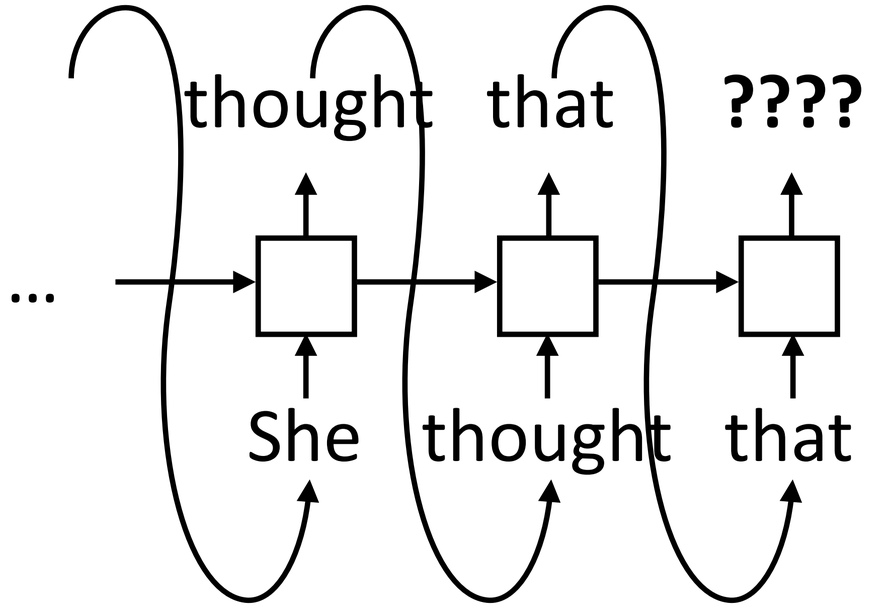
What types of models can we develop?

- 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)

LSTM Language Models

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 **????** had exaggerated matters a little.



- Predict next word with LSTM LM
- Context: either just the current sentence (query) or the whole document up to this point (query+context)

Hill et al. (2015)

LAMBADA

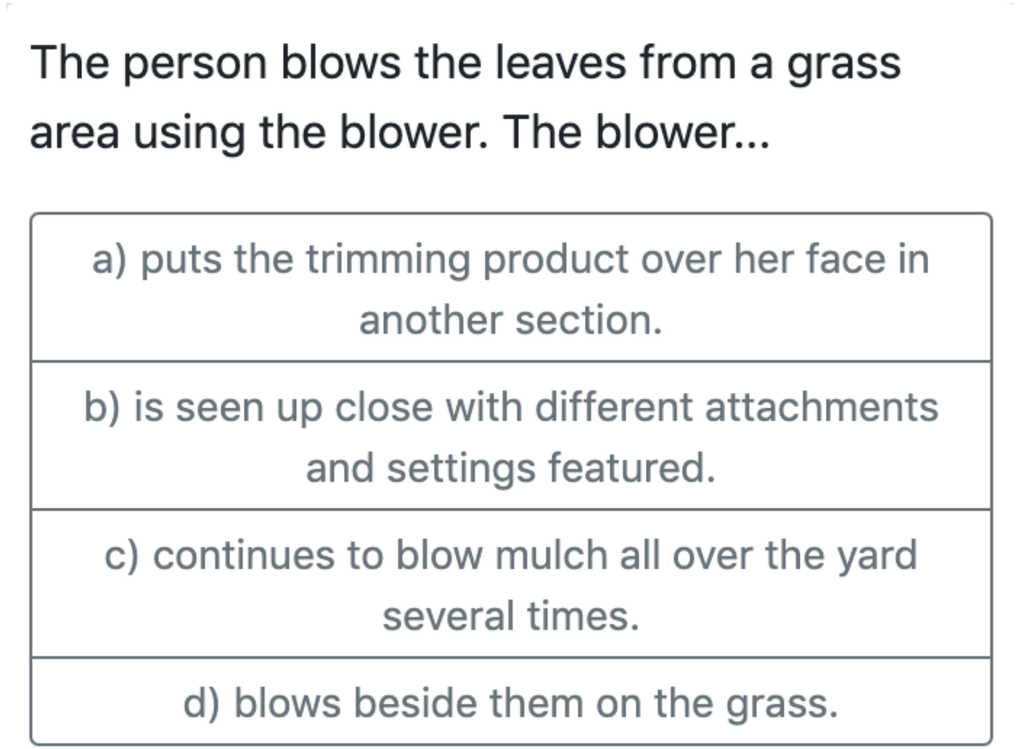
Context: They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I 've always loved _____.
Target word: dancing

- GPT/BERT can in general do very well at cloze tasks because this is what they're trained to do
- Hard to come up with plausible alternatives: "cooking", "drawing", "soccer", etc. don't work in the above context

Paperno et al. (2016)

SWAG

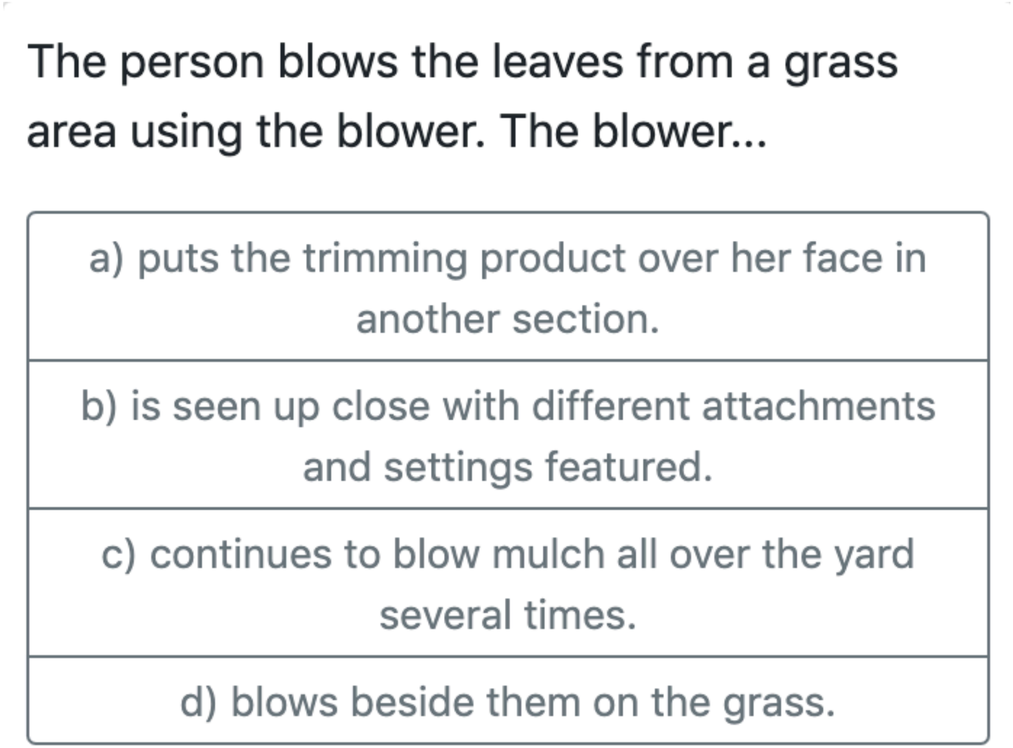
- Dataset was constructed to be difficult for ELMo (pretrained language model)
- BERT subsequently got 20+% accuracy improvements and achieved human-level performance



Zellers et al. (2018)

SWAG

- Dataset was constructed to be difficult for ELMo (pretrained language model)
- BERT subsequently got 20+% accuracy improvements and achieved human-level performance
- Problem: distractors too easy



- Let's look at architectures for retrieval from a passage

Zellers et al. (2018)

CNN/Daily Mail: Attentive Reader

CNN/Daily Mail

- Single-document, (usually) single-sentence cloze task
- Formed based on article summaries — information should mostly be present
- Need to process the question, can't just use LSTM LMs

Passage

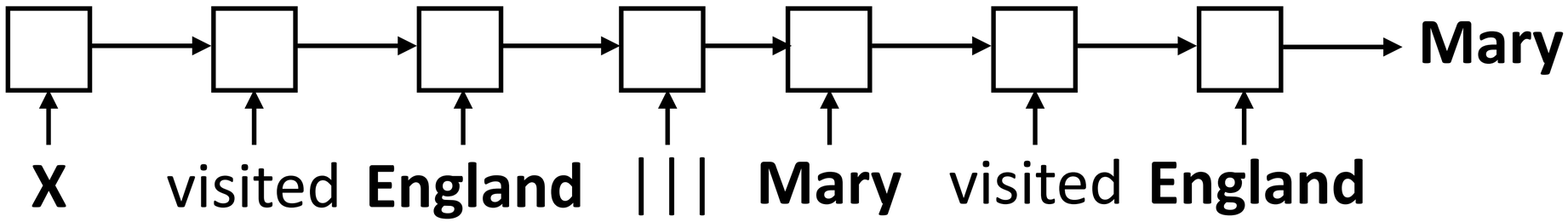
(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question	Answer
characters in " @placeholder " movies have gradually become more diverse	@entity6

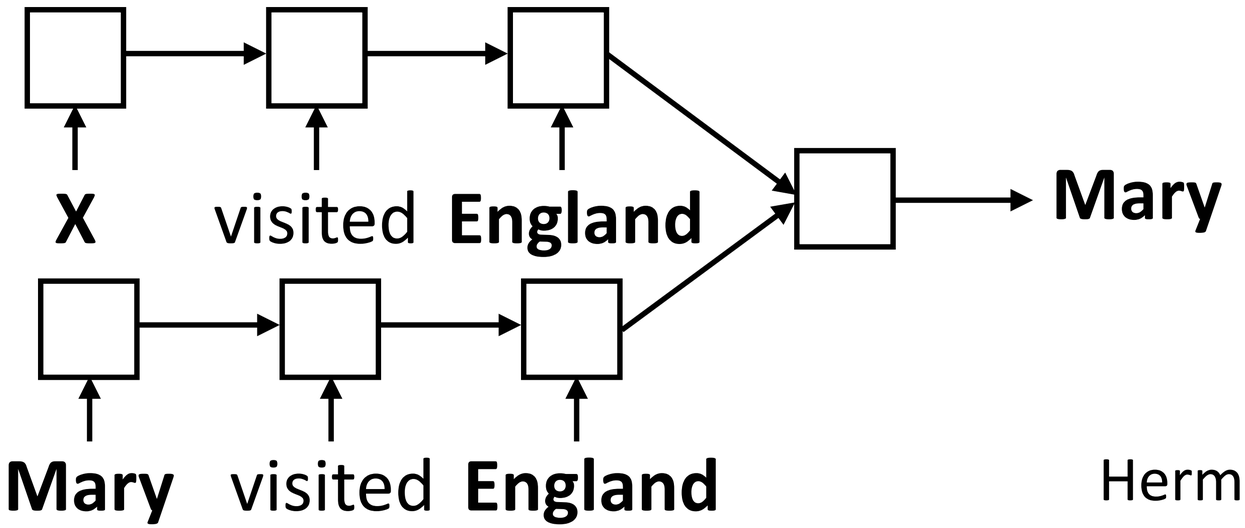
Hermann et al. (2015), Chen et al. (2016)

CNN/Daily Mail

- LSTM reader: encode question, encode passage, predict entity



- Can also use textual entailment-like models

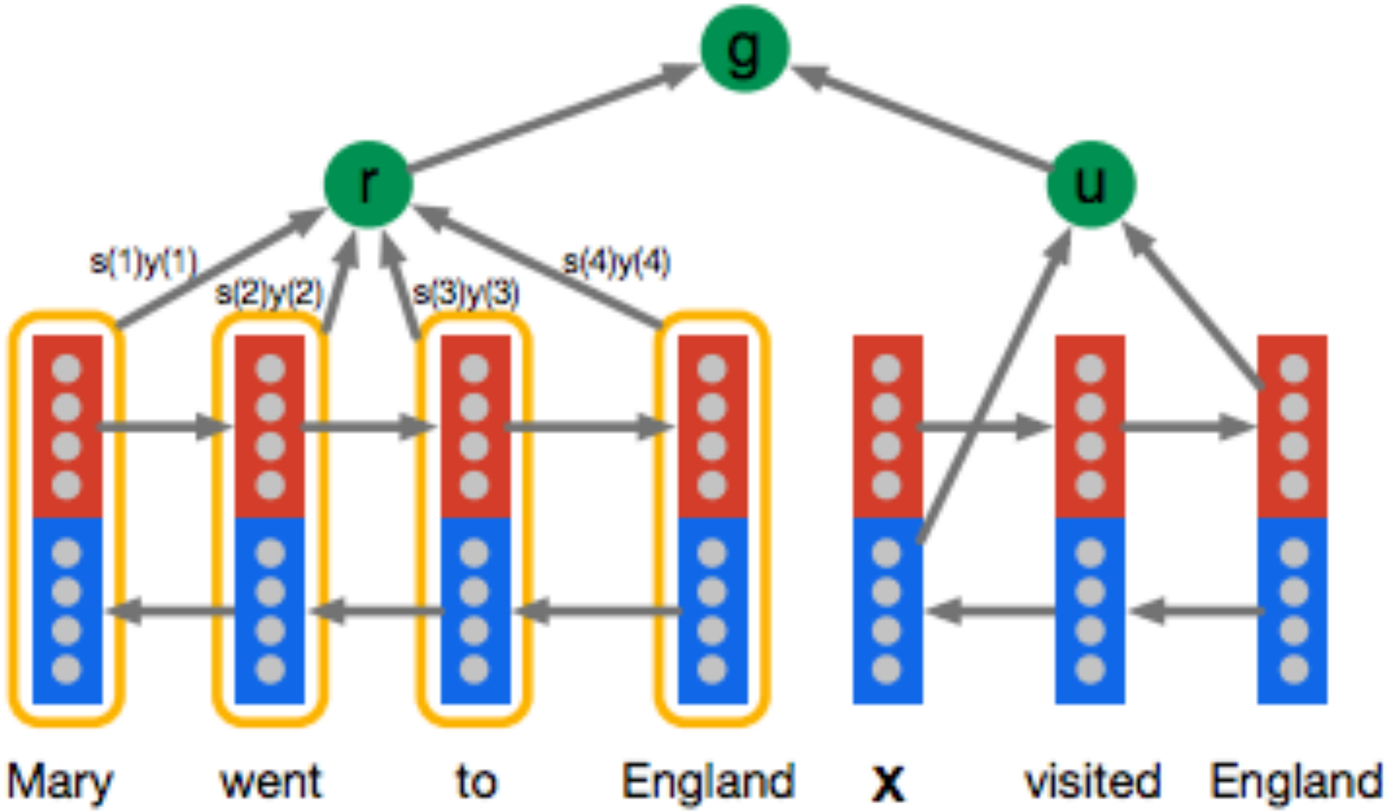


Multiclass classification problem over entities in the document

Hermann et al. (2015), Chen et al. (2016)

CNN/Daily Mail

- Attentive reader:
  u = encode query
  s = encode sentence
  r = attention(u -> s)
  prediction = f(candidate, u, r)
- Uses fixed-size representations for the final prediction, multiclass classification



Hermann et al. (2015)

CNN/Daily Mail

- Chen et al (2016): small changes to the attentive reader
- Additional analysis of the task found that many of the remaining questions were unanswerable or extremely difficult

	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	30.5	33.2	25.6	25.5
Exclusive frequency	36.6	39.3	32.7	32.8
Frame-semantic model	36.3	40.2	35.5	35.5
Word distance model	50.5	50.9	56.4	55.5
Deep LSTM Reader	55.0	57.0	63.3	62.2
Uniform Reader	39.0	39.4	34.6	34.4
Attentive Reader	61.6	63.0	**70.5**	**69.0**
Stanford Attentive Reader	76.2	76.5	**79.5**	**78.7**

Hermann et al. (2015), Chen et al. (2016)

SQuAD

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

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

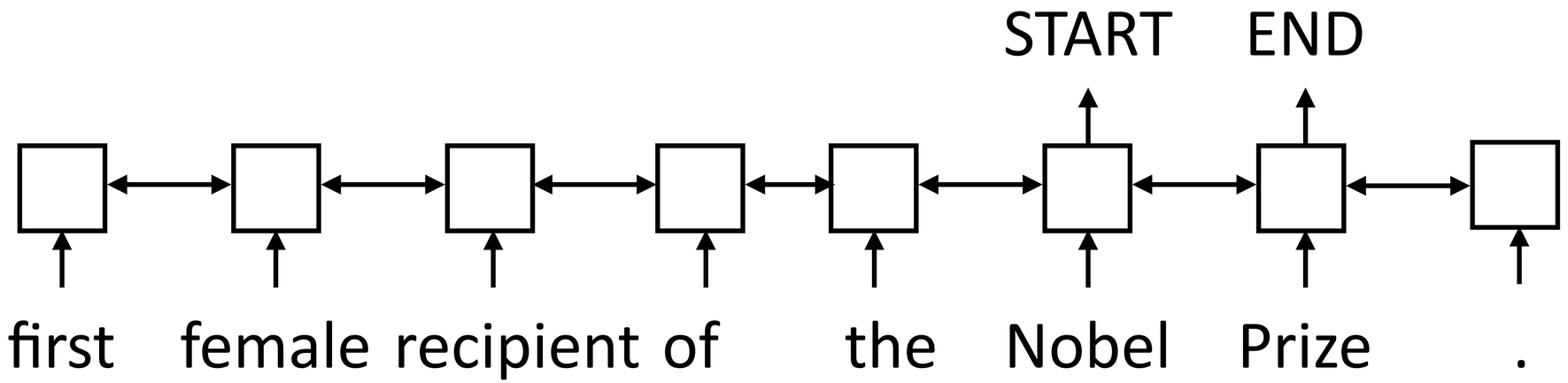
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Rajpurkar et al. (2016)

SQuAD

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)

Bidirectional Attention Flow (Offline Reading)

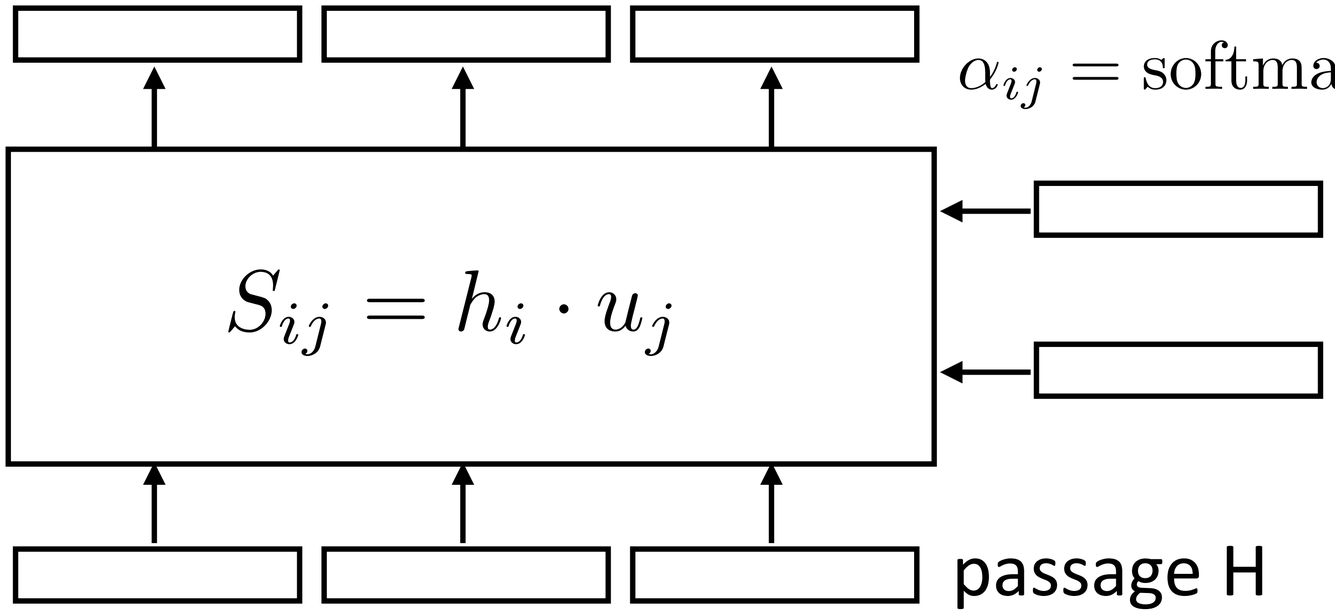
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

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

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

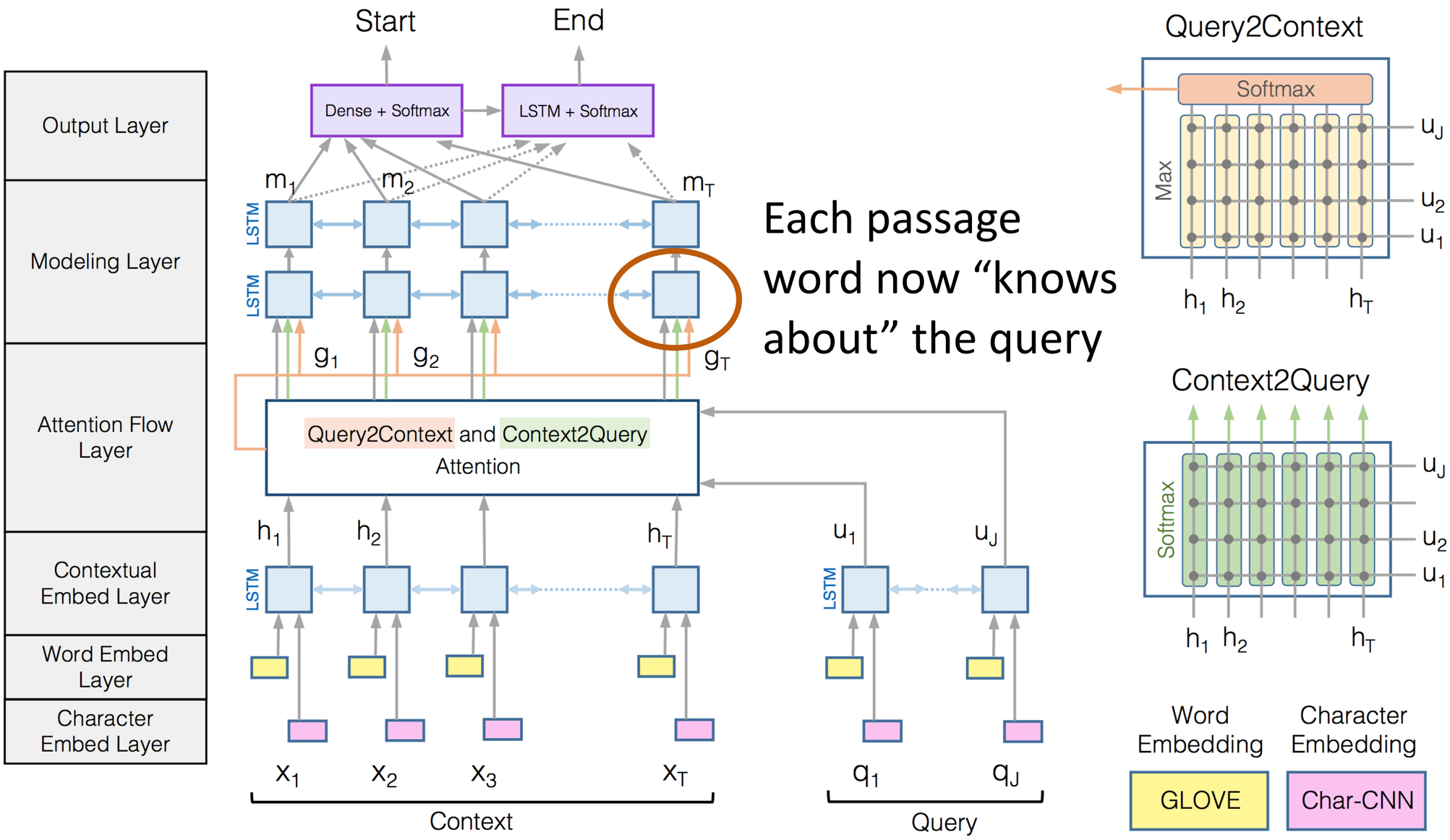
$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- dist over query



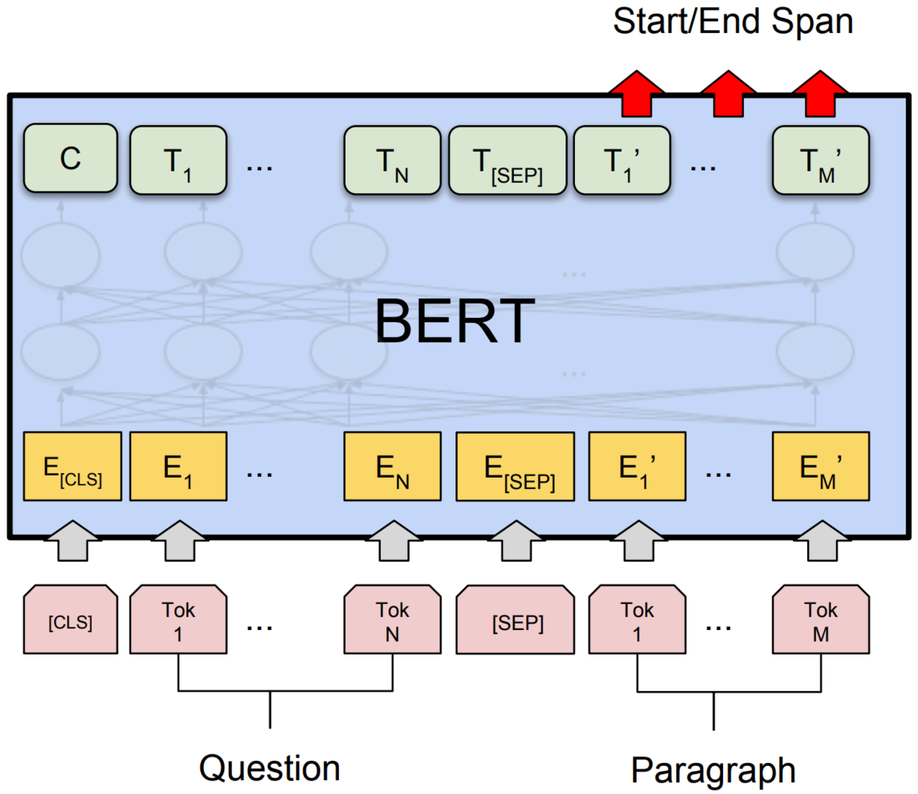
Seo et al. (2016)

Bidirectional Attention Flow (Offline Reading)



Seo et al. (2016)

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 in passage
- No need for cross-attention mechanisms!

Devlin et al. (2019)

SQuAD SOTA: Fall 18

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: Spring 19

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 SOTA: more recently

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!

https://rajpurkar.github.io/SQuAD-explorer/

Takeaways

- Many flavors of reading comprehension tasks: cloze or actual questions, single or multi-sentence

- Complex attention schemes can match queries against input texts and identify answers

Question Answering 2

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.

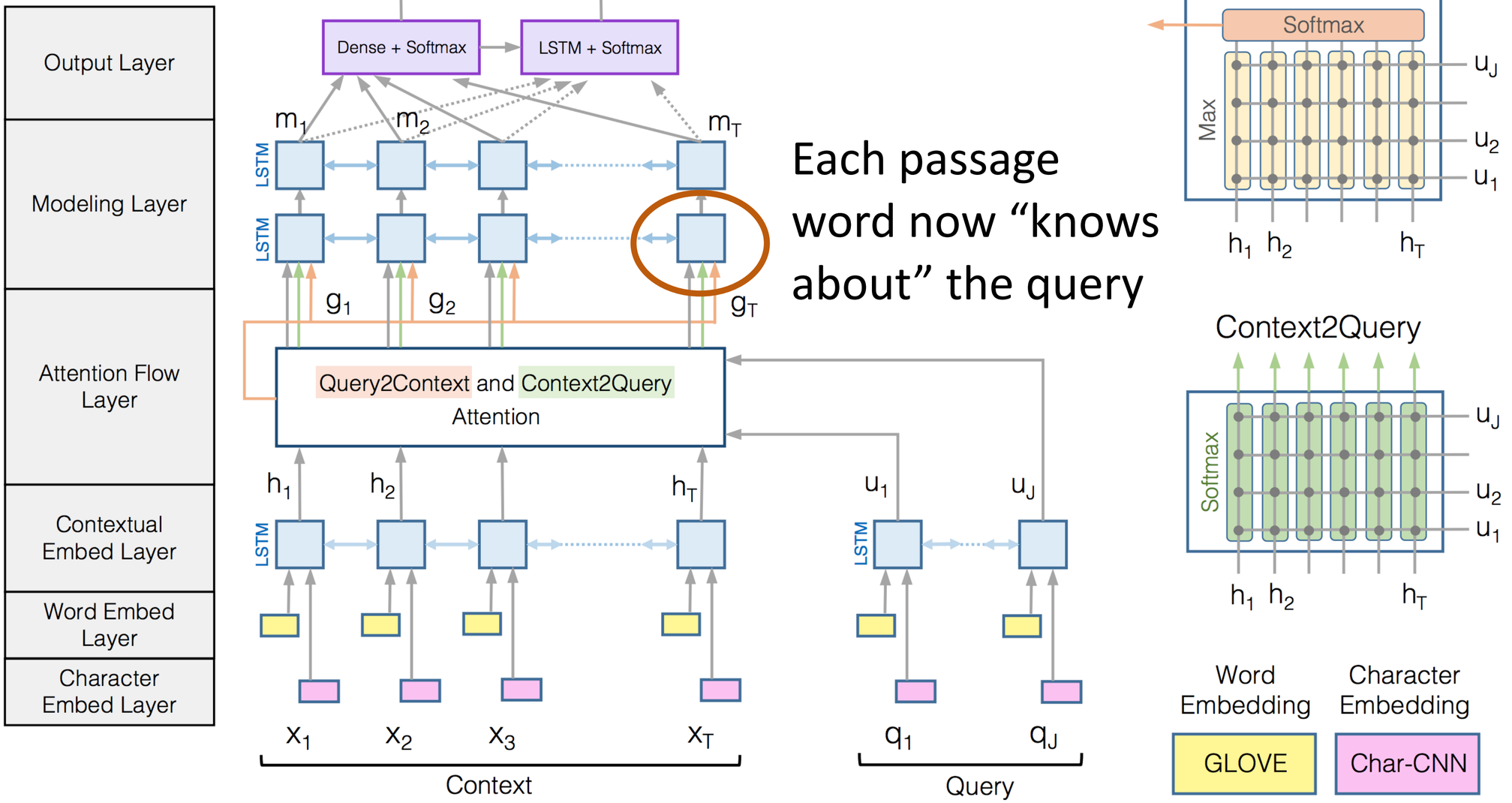
What was Maria Curie the first female recipient of?
Ground Truth Answers: Nobel Prize | Nobel Prize | Nobel Prize

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

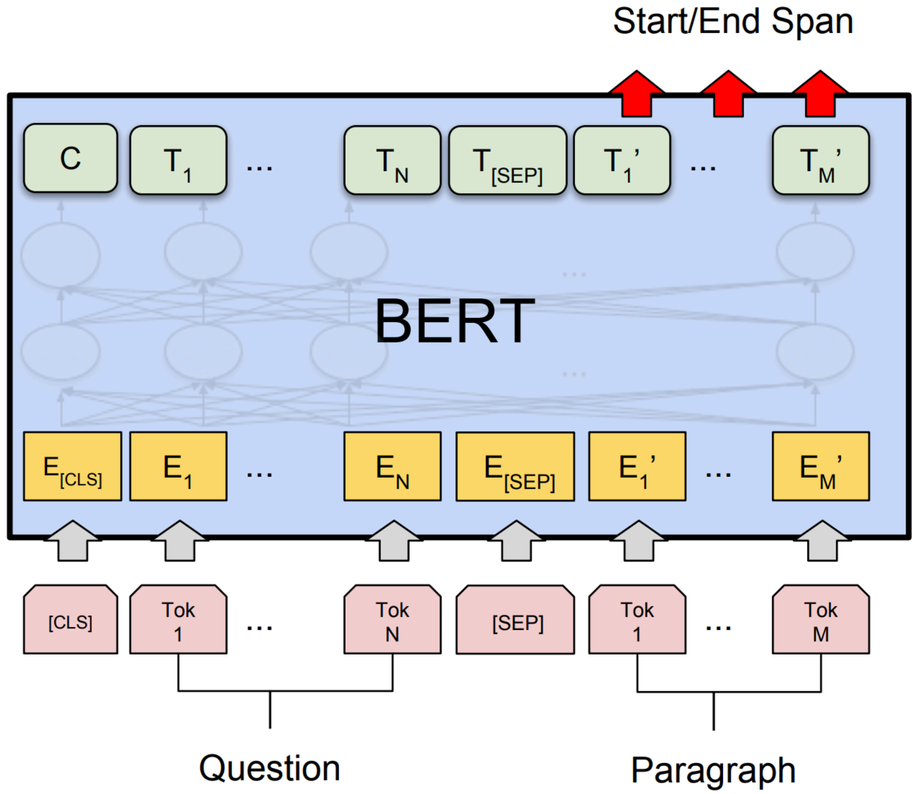
Rajpurkar et al. (2016)

Recall: Bidirectional Attention Flow



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: SQuAD SOTA

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

This Part

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

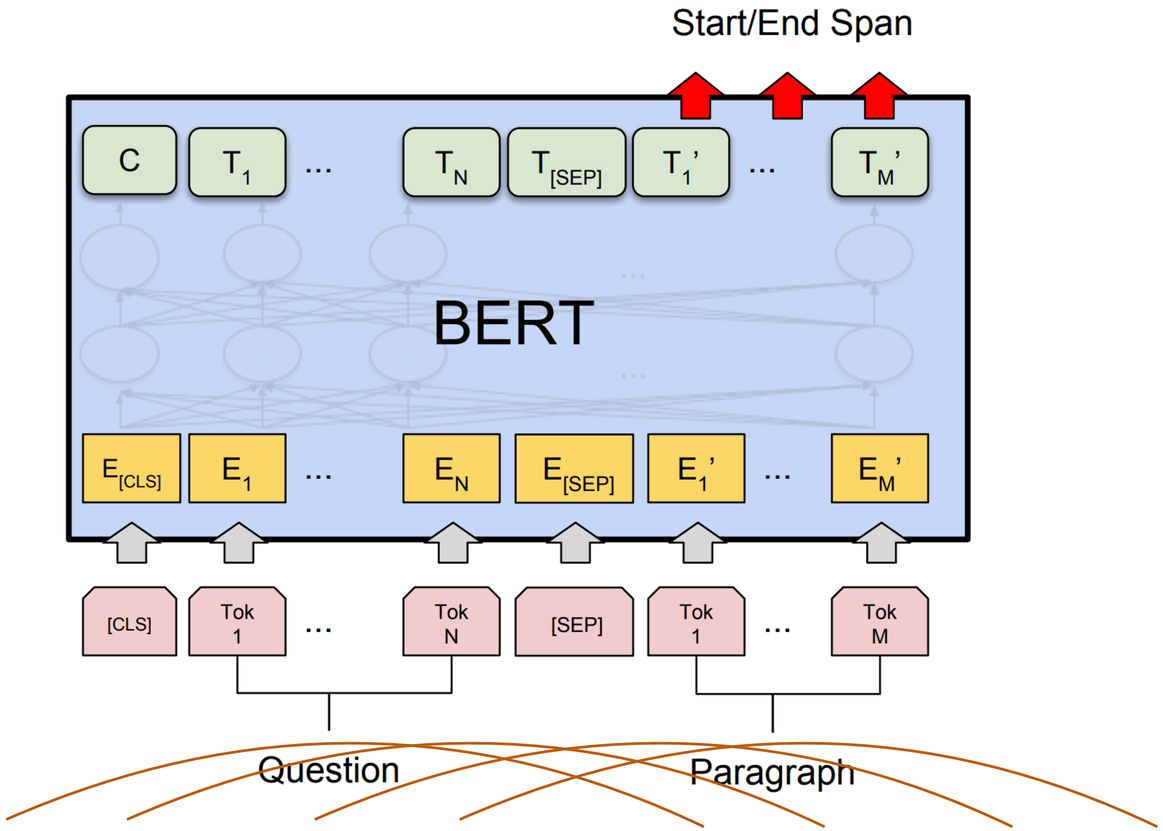
Problems in QA

Adversarial SQuAD

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

Jia and Liang (2017)

Adversarial SQuAD

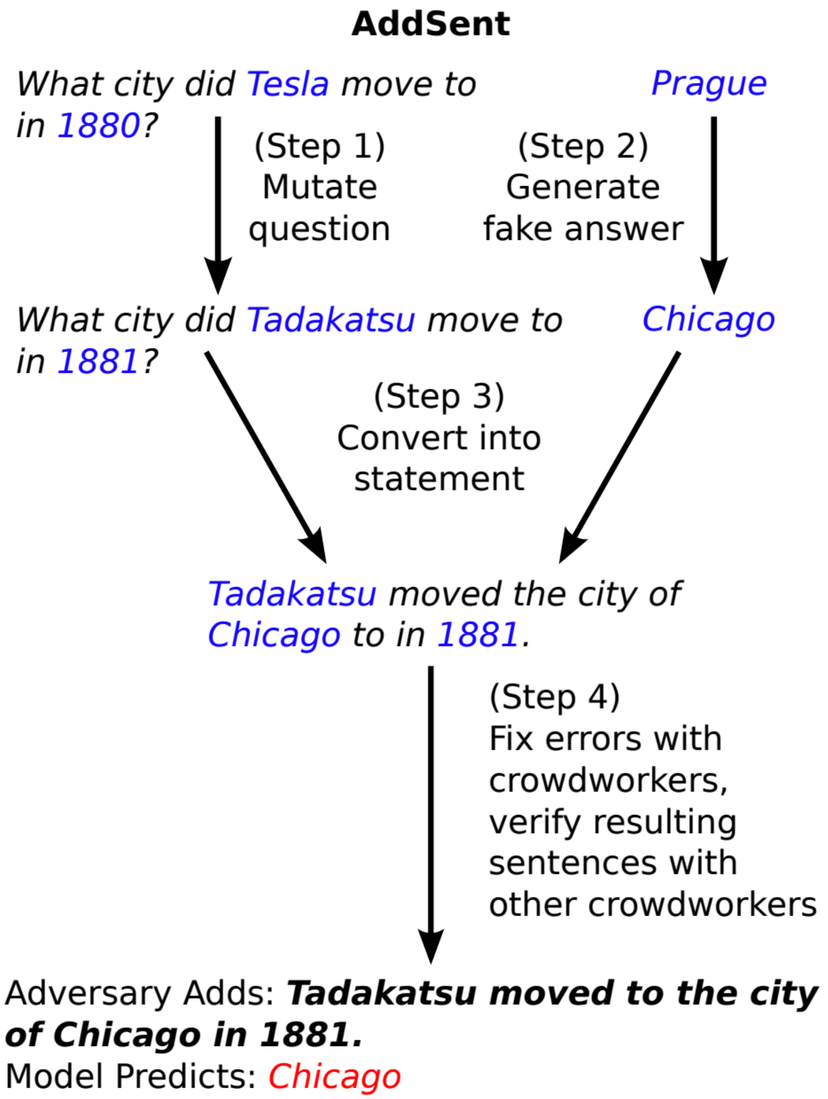


What was Marie Curie the first female recipient of ? [SEP] … first female recipient of **the Nobel Prize** …

- 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)

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

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

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

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
  - Find 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	+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

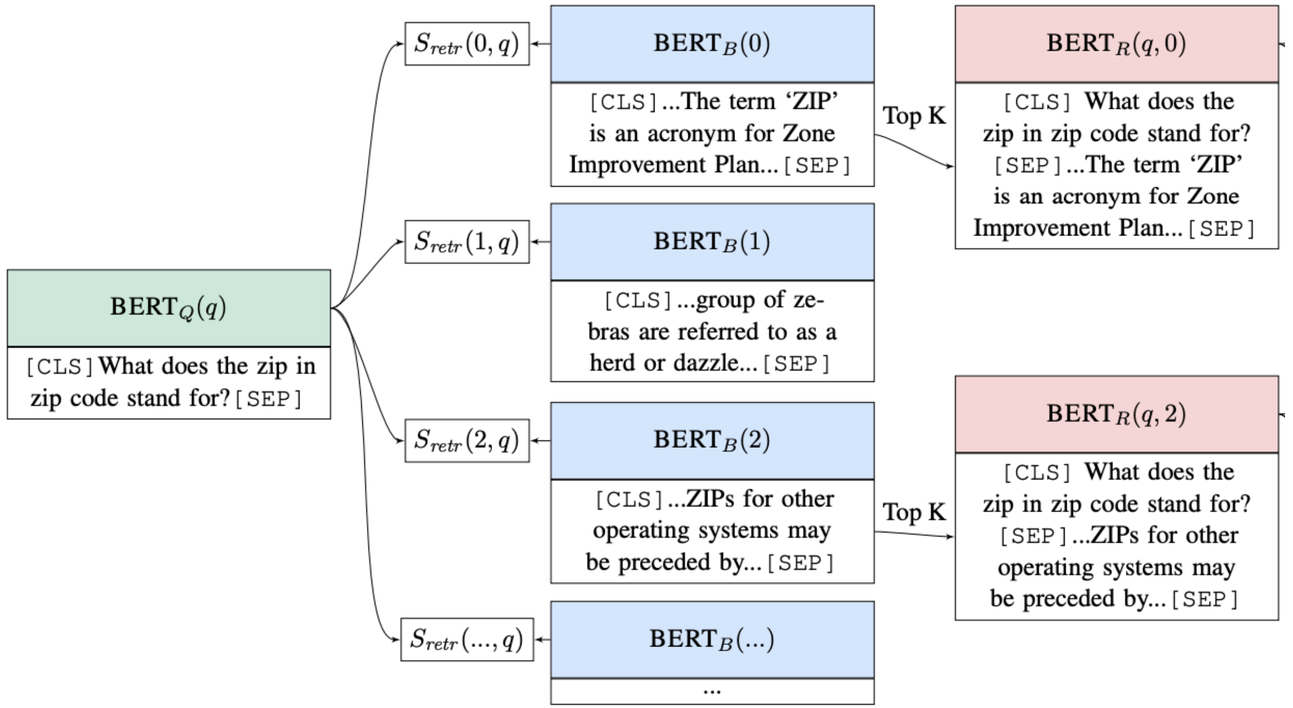
Dataset	SQuAD
SQuAD *(All Wikipedia)*	27.1
CuratedTREC	19.7
WebQuestions	11.8
WikiMovies	24.5

Chen et al. (2017)

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)

Problems

- Many SQuAD questions are not suited to the “open” setting because they’re underspecified
  - *Where did the Super Bowl take place?*

- 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)

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

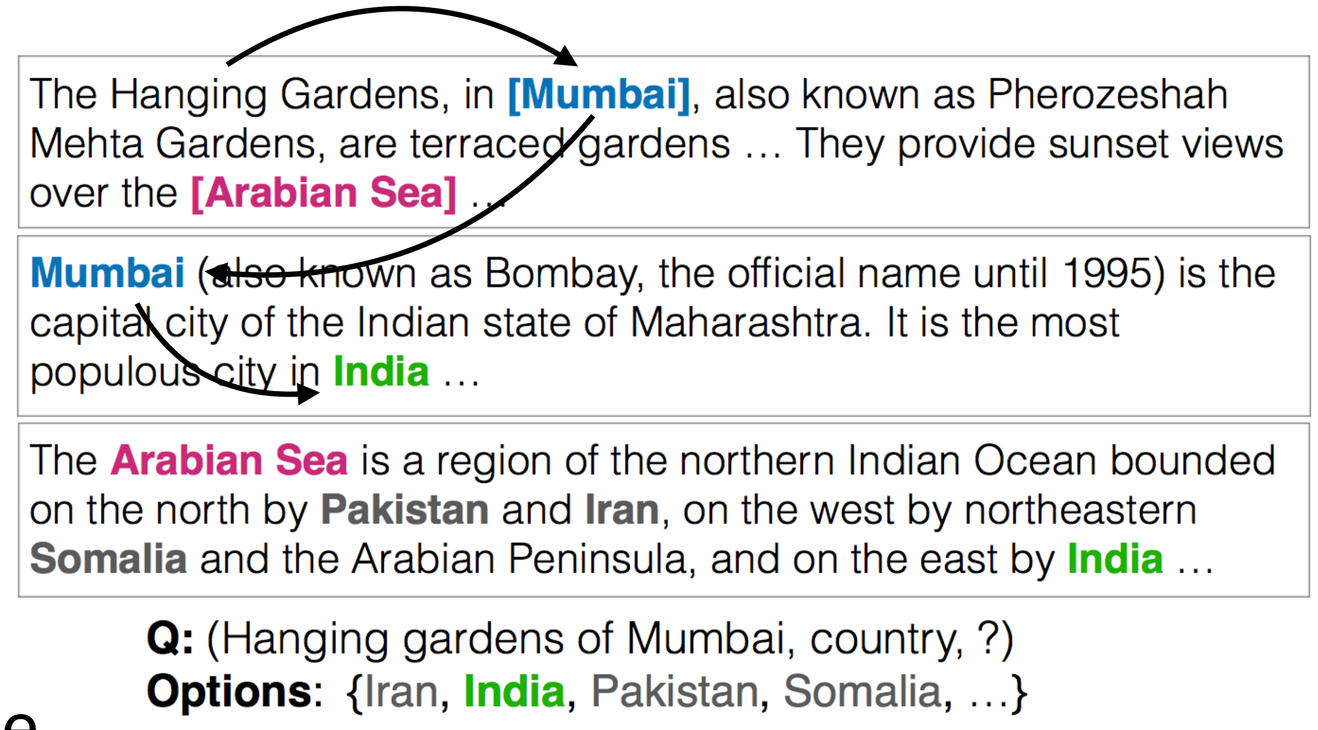


Figure from Welbl et al. (2018)

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 family is an Indian famil[illegible]hat 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**: 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]

Investigation

Can a model identify the answer with only a set of candidates?

Government position → *Chief of Protocol, actress, singer*

Can a model identify where the answer is in a single hop?

Oberoi Family → *Delhi*

Chen and Durrett (2019)

Finding the answer directly

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 7-year-old Shirley Temple acts as Corliss Archer .*

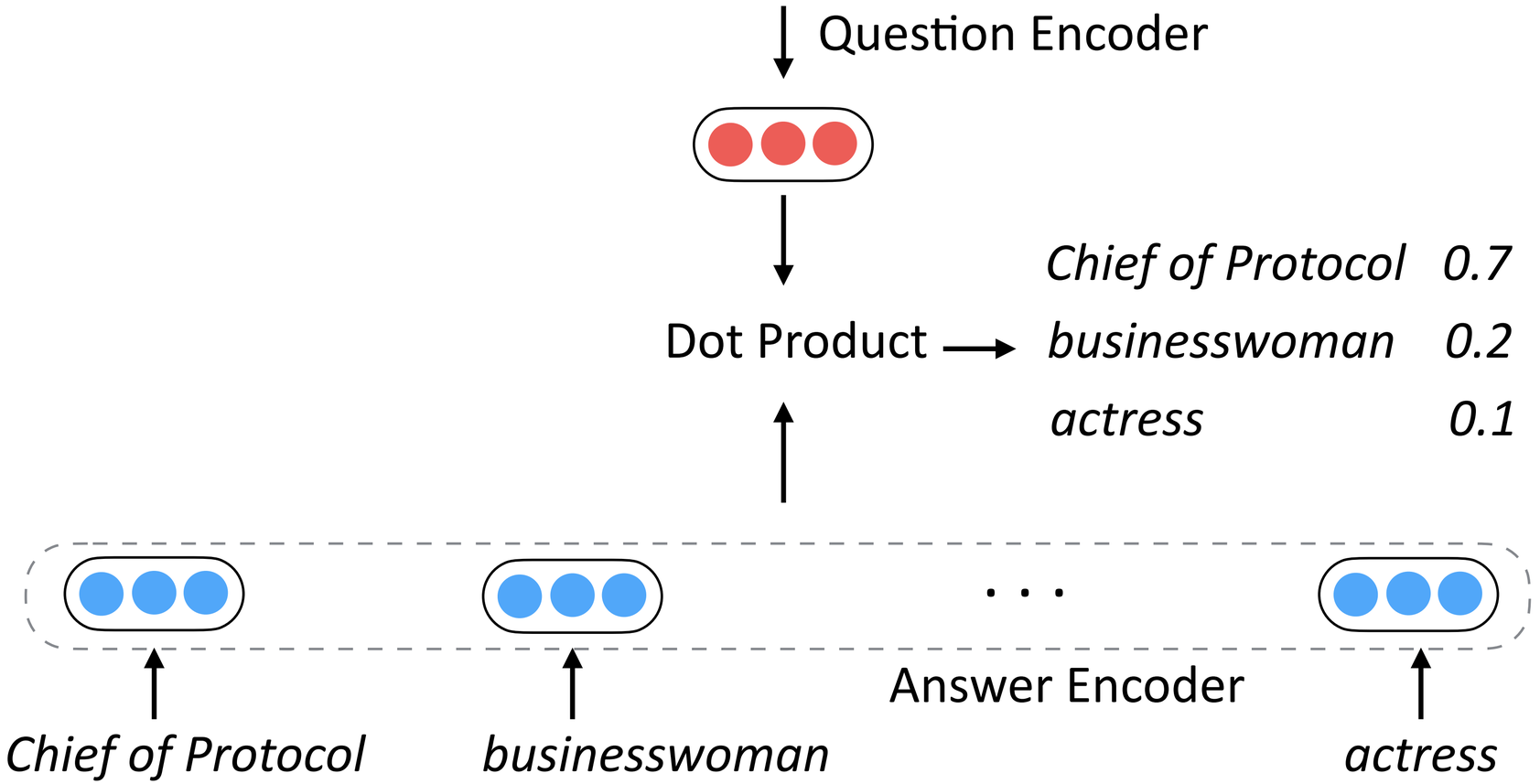
Doc 3 *Meet Corliss Archer is an American television sitcom that aired on CBS …*

Chief of Protocol **businesswoman** **…** **actress**

Kaushik and Lipton (2018)

No Context Baseline

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



Chen and Durrett (2019)

Results on WikiHop

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)

Investigation

Can a model identify the answer with only a set of candidates?

Government position → *Chief of Protocol, actress, singer*

Can a model identify where the answer is in a single hop?

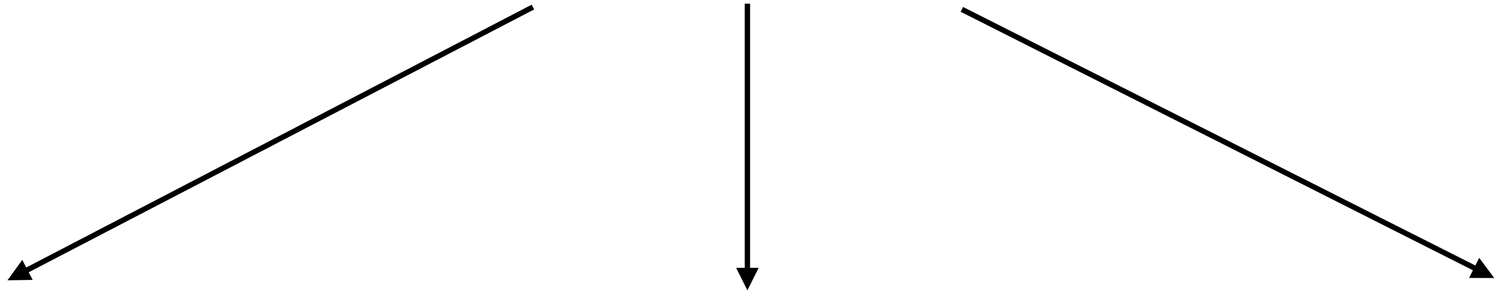
Oberoi Family → *Delhi*

Chen and Durrett (2019)

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 …

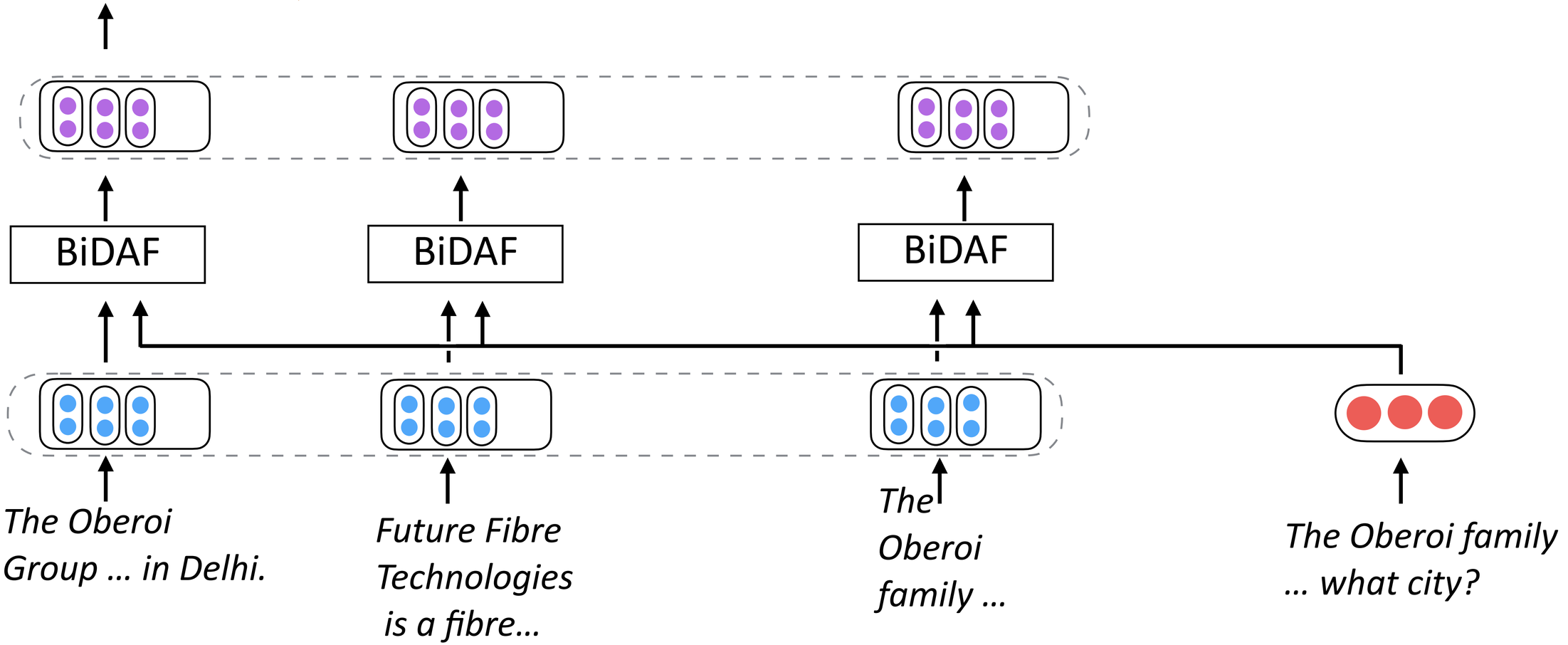
Chen and Durrett (2019)

Sentence Factored Model

Answer prediction:

Delhi

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



Chen and Durrett (2019)

Results on HotpotQA

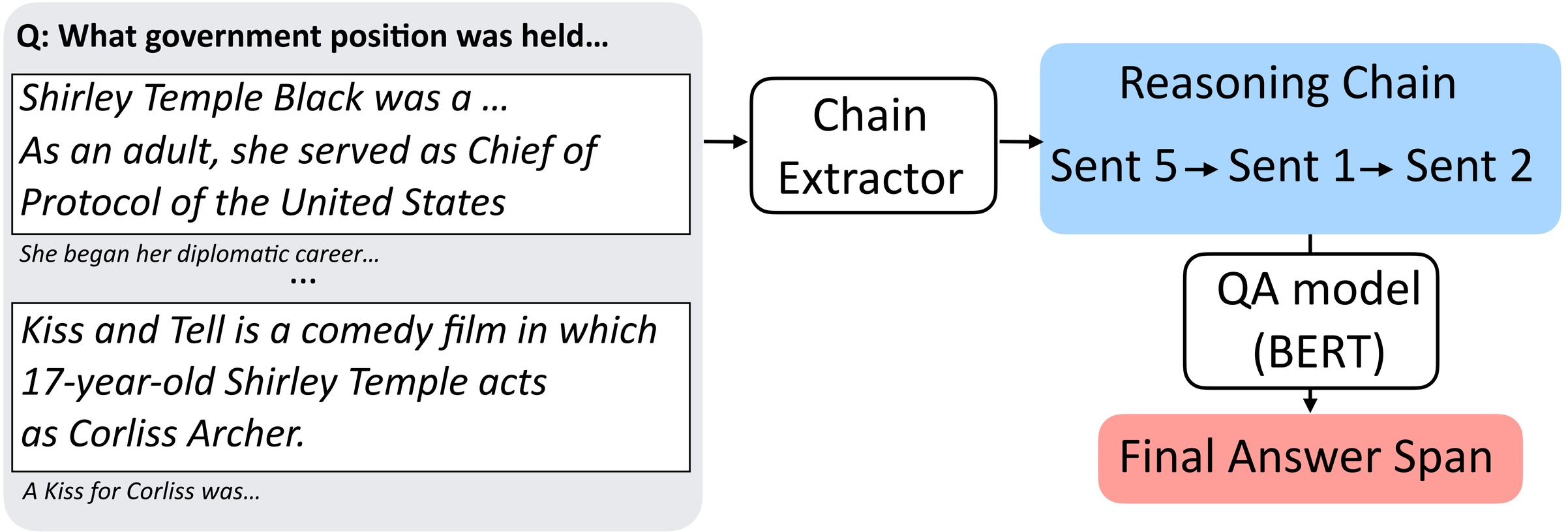
A simple single sentence reasoning model can solve more than half questions on HotpotQA.

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

Question Answering with Chains



- Maybe we can strengthen our models to avoid these weaknesses. Force them to explicitly extract a reasoning chain to make them better

Chen et al. (2019)

Question Answering with Chains

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

Doc 1

Shirley Temple Black was an American actress, businesswoman, and diplomat …

As an adult, she served as the Chief of Protocol of the United States …

She began her diplomatic career in 1969, when she represented …

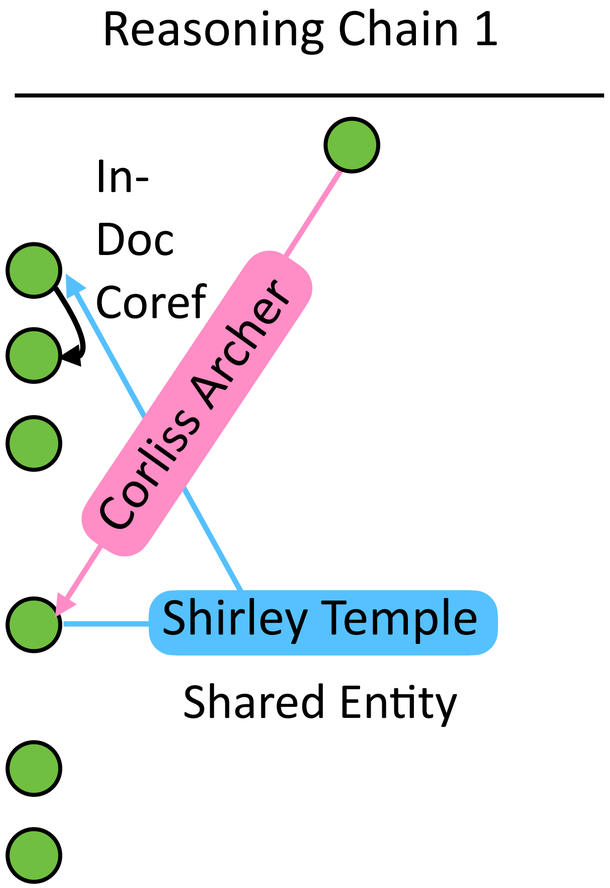
Doc 2

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

Doc 3

"A Kiss for Corliss" is a sequel to the film "Kiss and Tell".

It stars Shirley Temple in her final starring role …



- Strong connection between the entities used here

Chen et al. (2019)

Question Answering with Chains

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

Doc 1

Shirley Temple Black was an American actress, businesswoman, and diplomat …

As an adult, she served as the Chief of Protocol of the United States …

She began her diplomatic career in 1969, when she represented …

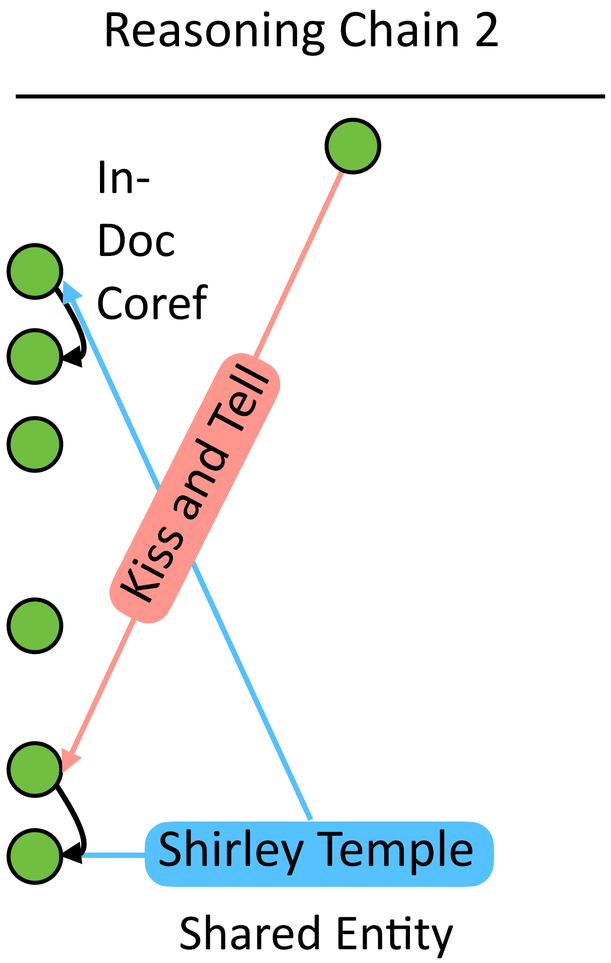
Doc 2

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

Doc 3

"A Kiss for Corliss" is a sequel to the film "Kiss and Tell".

It stars Shirley Temple in her final starring role …

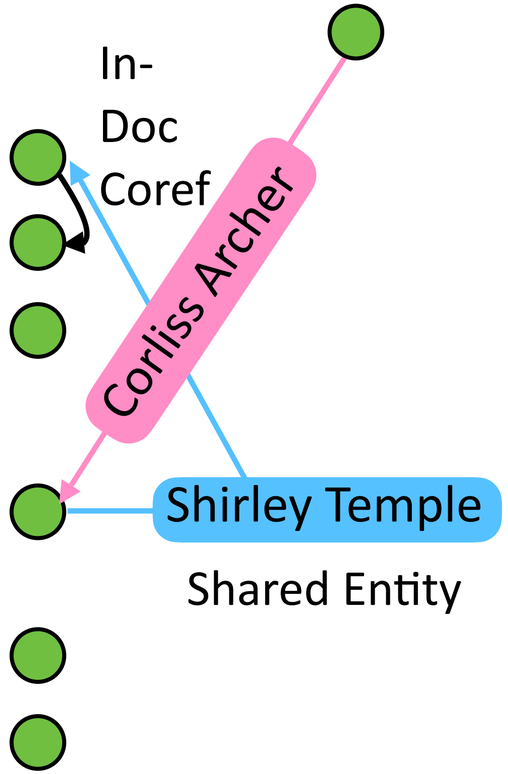


- More speculative than the other chain but still leads to the answer

Chen et al. (2019)

Chain Supervision

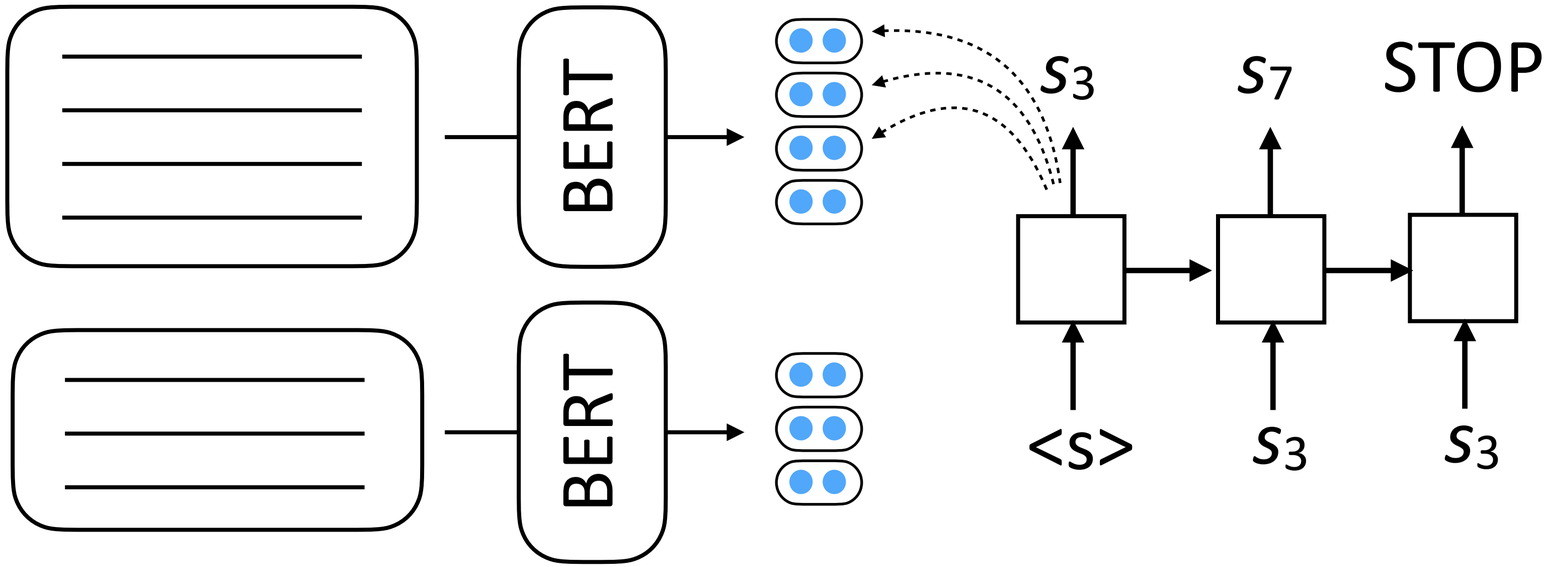
- Extract pseudogold chains based on:
  - Within-document coreference: we don't run a coreference system but instead link all sentences within a paragraph
  - Shared entities: enable connections between different sources
- Given these chains, we learn a model to extract them. **At test time, no annotations are needed**

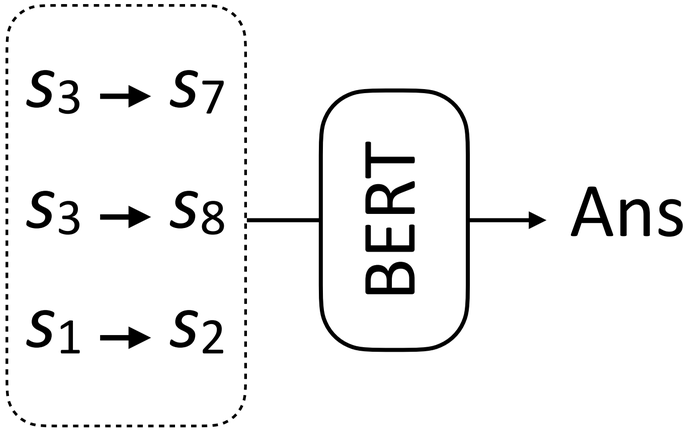


Chen et al. (2019)

Chain Extraction and QA

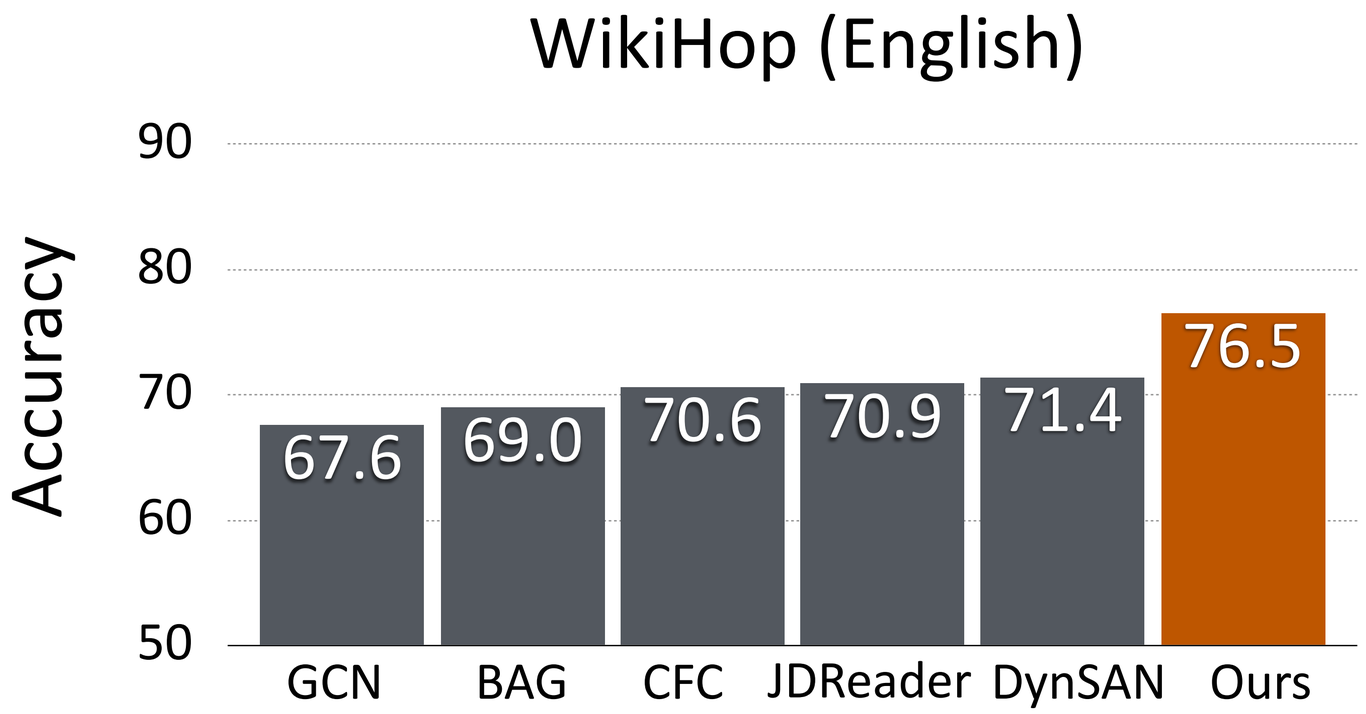
- Paragraphs are encoded with BERT to compute sentence representations
- A pointer network selects a sequence of sentences
- A final BERT model then extracts an answer span from one or more chains

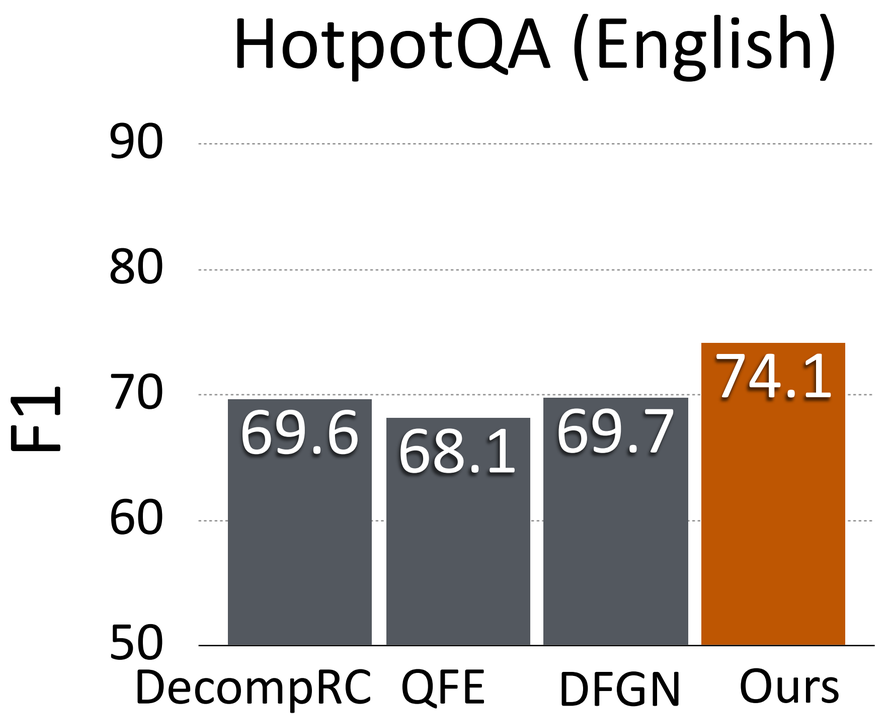




Chen et al. (2019)

QA Results





- High performance on WikiHop (*past systems didn't use BERT) and Hotpot
- **Also large gains on hard examples in HotpotQA** (our model from part 1 could not find answers in a single hop)
- Ongoing work: how can reasoning chains be taken below the sentence level and be more strongly tied to interpretable logical inference?

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