Lecture 14: Reading Comprehension

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”

```
λx. type(x, Location) ∧ 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”

```
λx. 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 from Open IE

(a) **CCG parse** builds an underspecified semantic representation of the sentence.

Former	municipalities	in	Brandenburgh
N/N	N	$N\backslash N/NP$	NP
$\lambda f \lambda x.f(x) \wedge former(x)$	$\lambda x.municipalities(x)$	$\lambda f \lambda x \lambda y.f(y) \wedge in(y,x)$	$Brandenburg$
	N (>)		$N\backslash N$ (>)
	$\lambda x.former(x) \wedge municipalities(x)$		$\lambda f \lambda y.f(y) \wedge in(y, Brandenburg)$
	N (<)		

$$l_0 = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$

(b) **Constant matches** replace underspecified constants with Freebase concepts

$$\mathbf{l_0} = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$

$$\mathbf{l_1} = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, \texttt{Brandenburg})$$

$$\mathbf{l_2} = \lambda x.former(x) \wedge municipalities(x) \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$

$$\mathbf{l_3} = \lambda x.former(x) \wedge \texttt{OpenRel}(x, \texttt{Municipality}) \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$

$$\mathbf{l_4} = \lambda x.\texttt{OpenType}(x) \wedge \boxed{\texttt{OpenRel}(x, \texttt{Municipality})} \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$

Choi et al. (2015)

QA from Open IE

(a) **CCG parse** builds an underspecified semantic representation of the sentence.

Former	municipalities	in	Brandenburgh
N/N	N	$N\backslash N/NP$	NP
$\lambda f \lambda x.f(x) \wedge former(x)$	$\lambda x.municipalities(x)$	$\lambda f \lambda x \lambda y.f(y) \wedge in(y,x)$	$Brandenburg$
	N (>)		$N\backslash N$ (>)
	$\lambda x.former(x) \wedge municipalities(x)$		$\lambda f \lambda y.f(y) \wedge in(y, Brandenburg)$
	N (<)		

$$l_0 = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$

(b) **Constant matches** replace underspecified constants with Freebase concepts

$$\mathbf{l_0} = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$
$$\mathbf{l_1} = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, \texttt{Brandenburg})$$
$$\mathbf{l_2} = \lambda x.former(x) \wedge municipalities(x) \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$
$$\mathbf{l_3} = \lambda x.former(x) \wedge \texttt{OpenRel}(x, \texttt{Municipality}) \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$
$$\mathbf{l_4} = \lambda x.\texttt{OpenType}(x) \wedge \boxed{\texttt{OpenRel}(x, \texttt{Municipality})} \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$

- Why use the KB at all? Why not answer questions directly from text?
  Like information retrieval!

Choi et al. (2015)

What can’t KB QA systems do?

What can’t KB QA systems do?

- What were the main causes of World War II? — requires summarization

What can’t KB QA systems do?

- 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

What can't KB QA systems do?

- 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
- What temperature should I cook chicken to? — could be written down in a KB but probably isn't

What can't KB QA systems do?

- 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
- 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

- "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
- Recognizing Textual Entailment (2006)

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
- Recognizing Textual Entailment (2006)
- 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)

Baselines

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

Richardson (2013)

Baselines

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

Richardson (2013)

Baselines

- N-gram matching: append question + each answer, return answer which gives highest n-gram overlap with a sentence

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

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

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

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

Richardson (2013)

Reading Comprehension

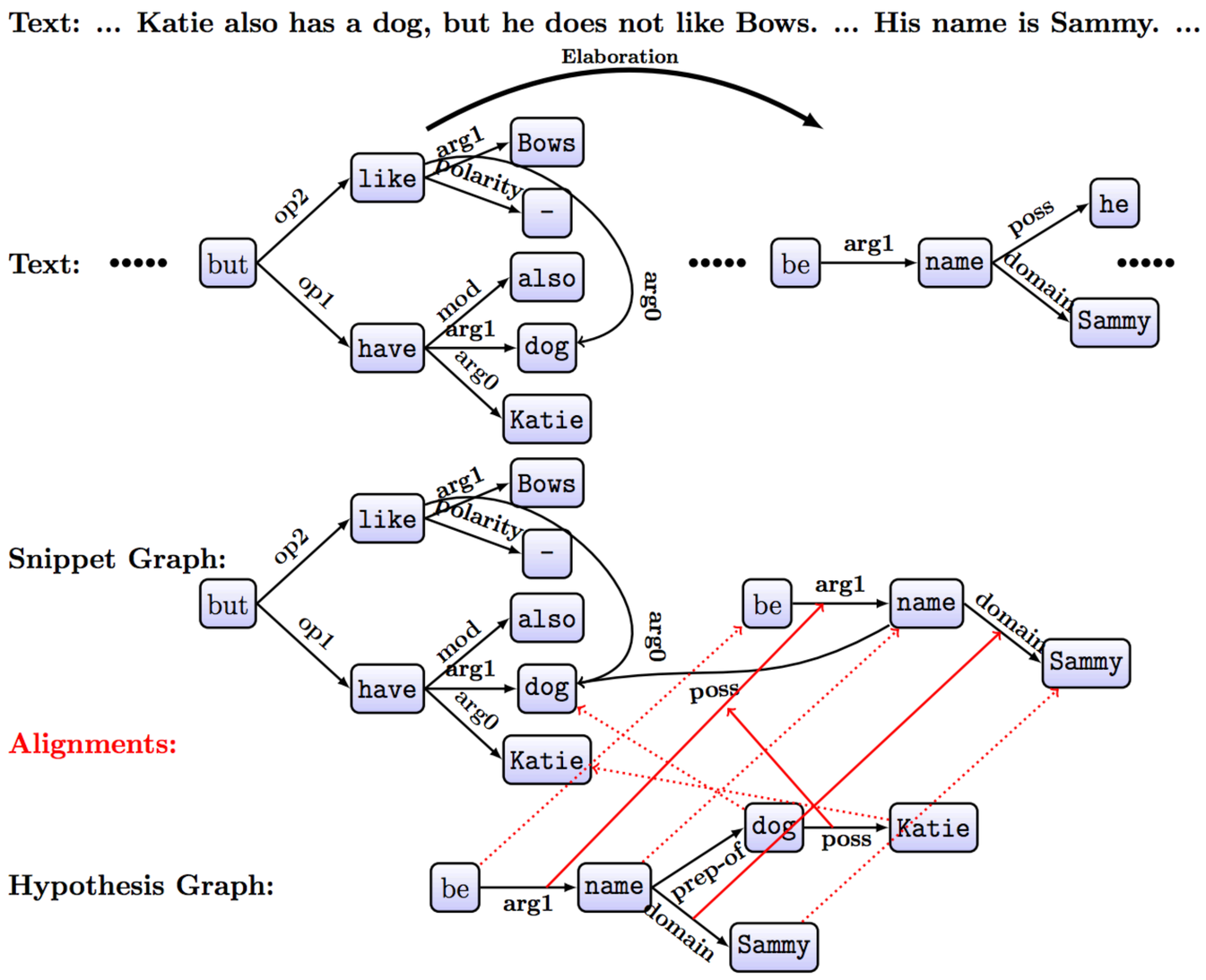
ngram sliding window

	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)

MCTest State of the Art



Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. Answer: Sammy

- Match an AMR (abstract meaning representation) of the question against the original text
- 70% accuracy (roughly 10% better than baseline)

Sachan and Xing (2016)

Dataset Explosion

- 10+ QA datasets released since 2015
  - Children's Book Test, CNN/Daily Mail, SQuAD, TriviaQA are most well-known (others: SearchQA, MS Marco, RACE, WikiHop, …)

Dataset Explosion

- 10+ QA datasets released since 2015
  - Children's Book Test, CNN/Daily Mail, SQuAD, TriviaQA are most well-known (others: SearchQA, MS Marco, RACE, WikiHop, …)
- Question answering: questions are in natural language
  - Answers: multiple choice or require picking from the passage
  - Require human annotation

Dataset Explosion

- 10+ QA datasets released since 2015
  - Children's Book Test, CNN/Daily Mail, SQuAD, TriviaQA are most well-known (others: SearchQA, MS Marco, RACE, WikiHop, …)
- 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

Dataset Properties

- Axis 1: QA vs. cloze

Dataset Properties

- Axis 1: QA vs. cloze
- Axis 2: single-sentence vs. passage
  - Often shallow methods work well because most answers are in a single sentence (SQuAD, MCTest)
  - Some explicitly require linking between multiple sentences (MCTest)

Dataset Properties

- Axis 1: QA vs. cloze
- Axis 2: single-sentence vs. passage
  - Often shallow methods work well because most answers are in a single sentence (SQuAD, MCTest)
  - Some explicitly require linking between multiple sentences (MCTest)
- Axis 3: single-document (datasets in this lecture) vs. multi-document (TriviaQA, WikiHop, HotPotQA, …)

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 .

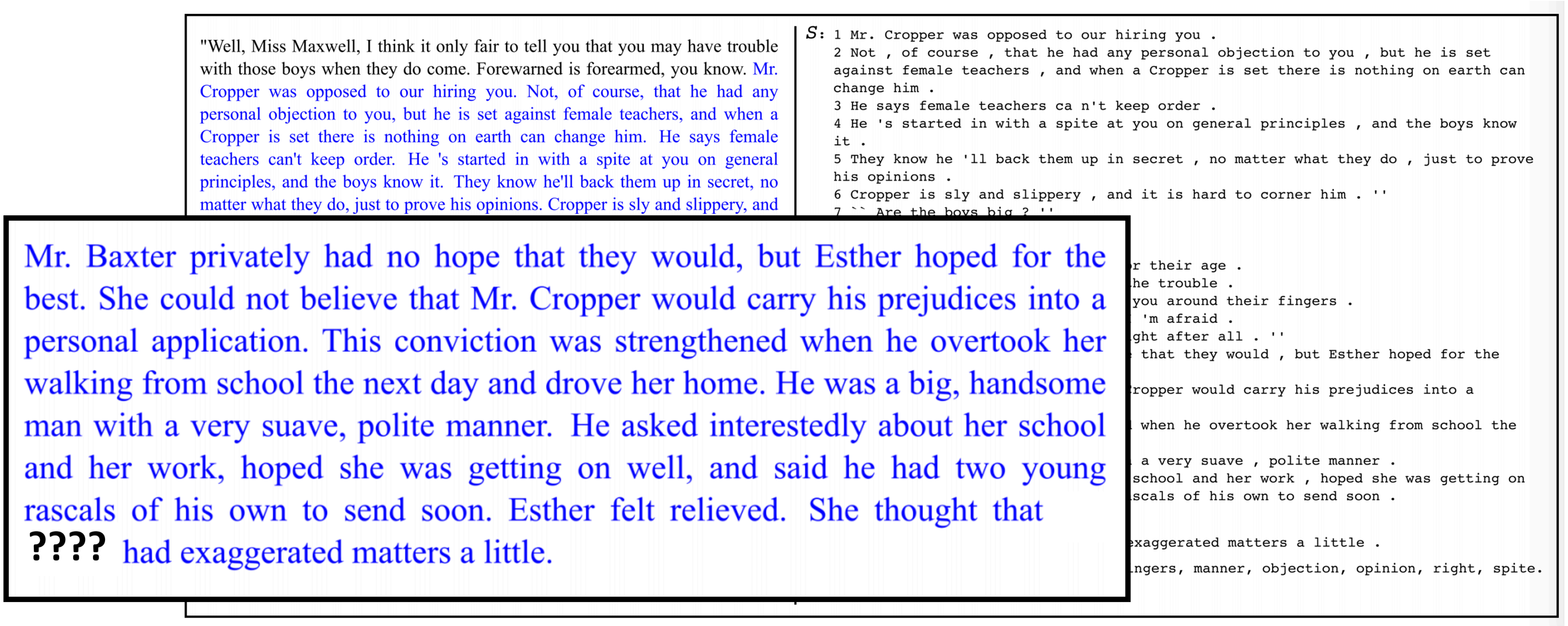
C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

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

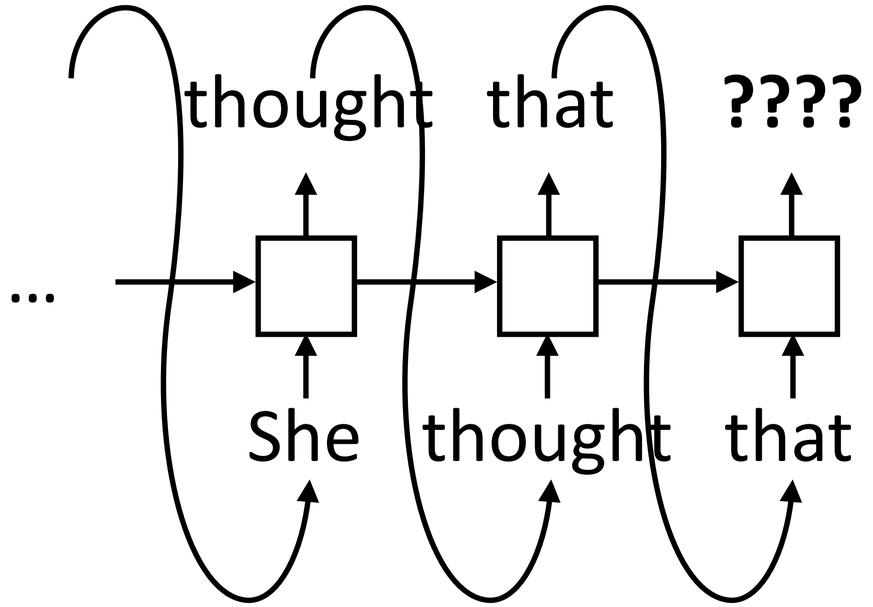


- Children's Book Test: take a section of a children's story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

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.

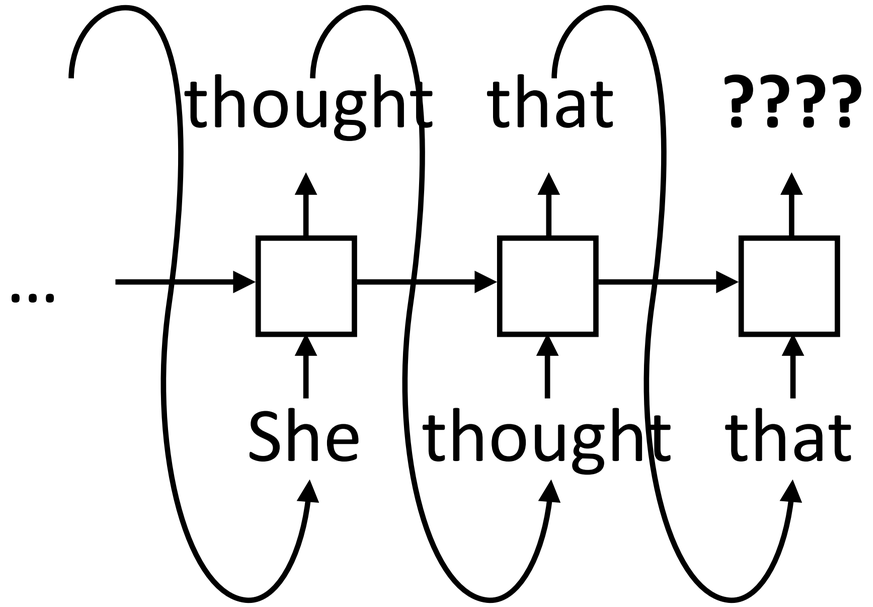


- Predict next word with LSTM LM

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
- 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)

Memory Networks

Memory Networks

- Memory networks let you reference input with attention

Sukhbaatar et al. (2015)

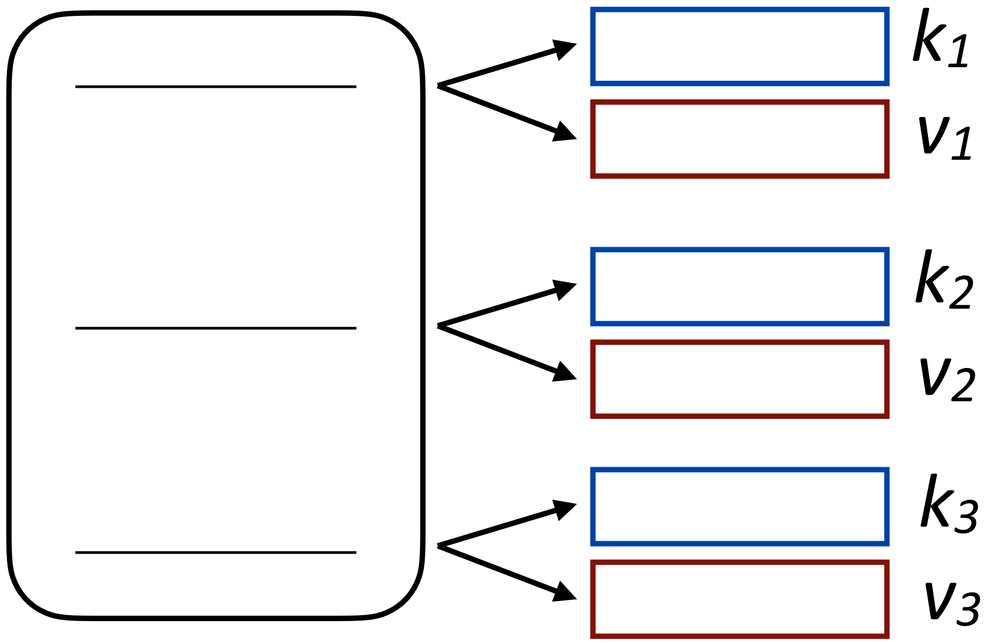
Memory Networks

- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value

Sukhbaatar et al. (2015)

Memory Networks

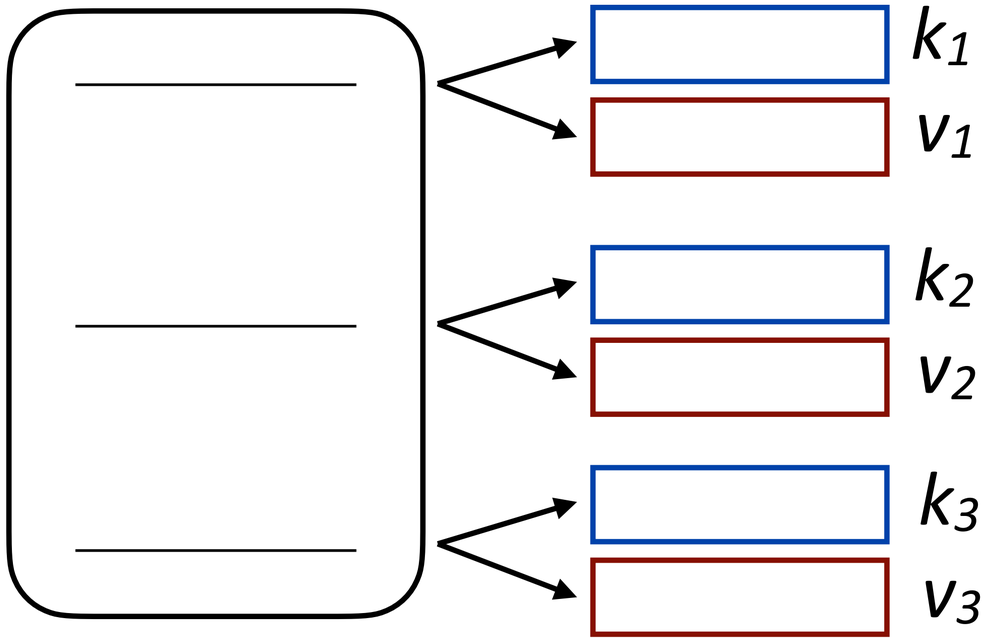
- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value



Sukhbaatar et al. (2015)

Memory Networks

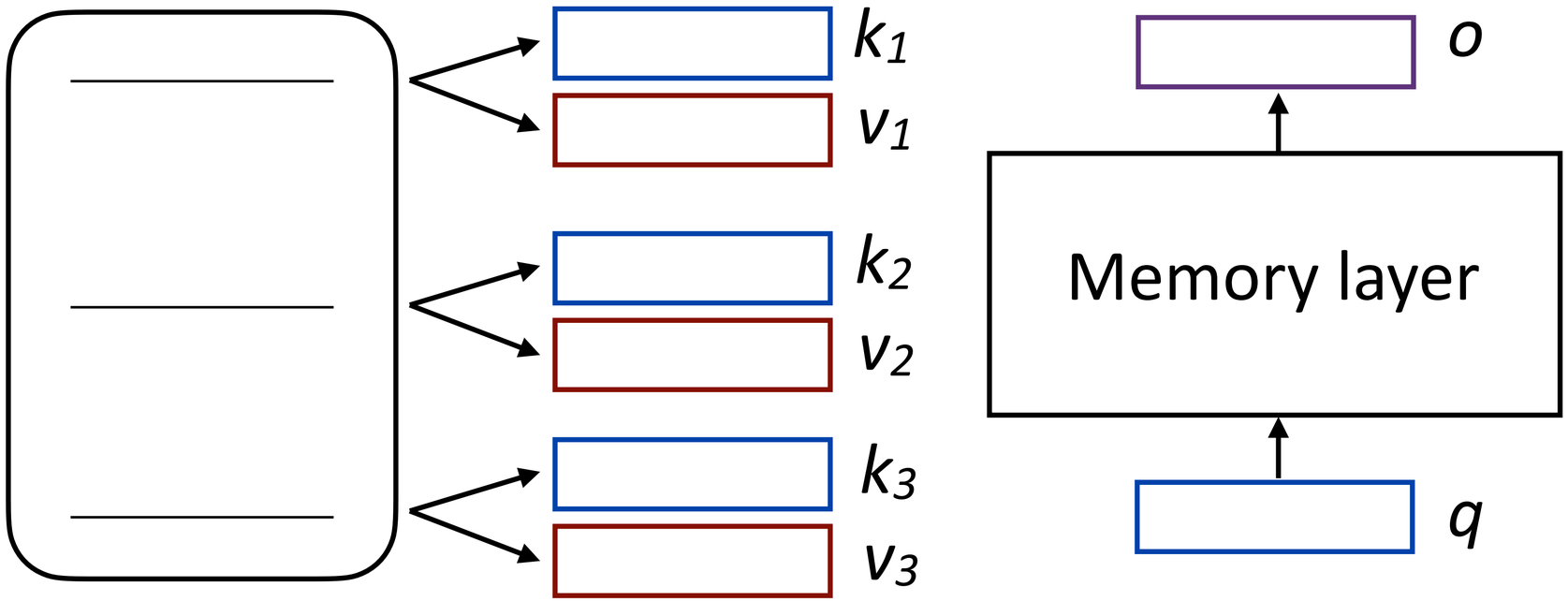
- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output



Sukhbaatar et al. (2015)

Memory Networks

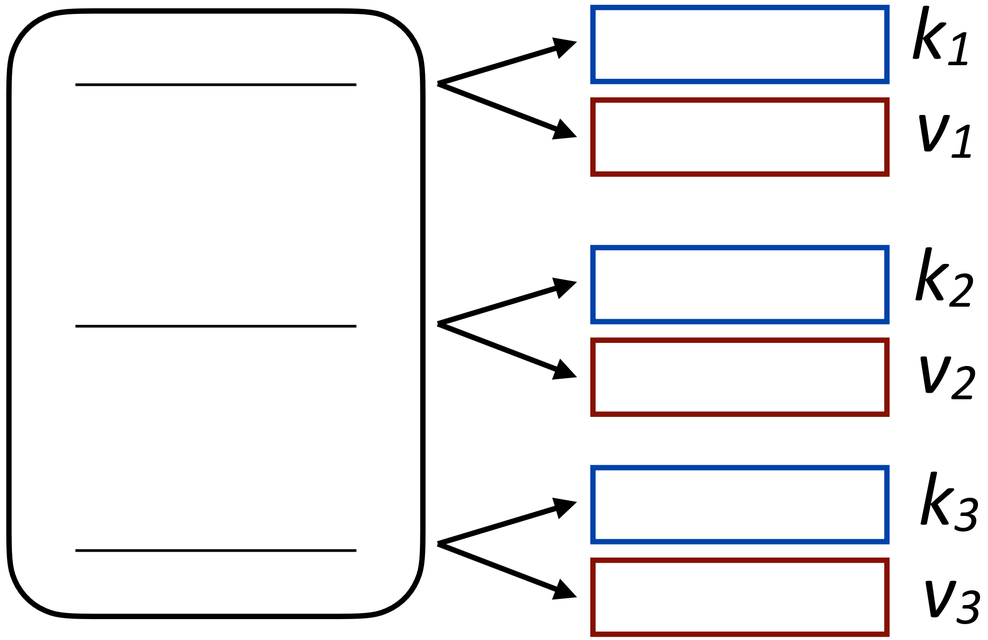
- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output

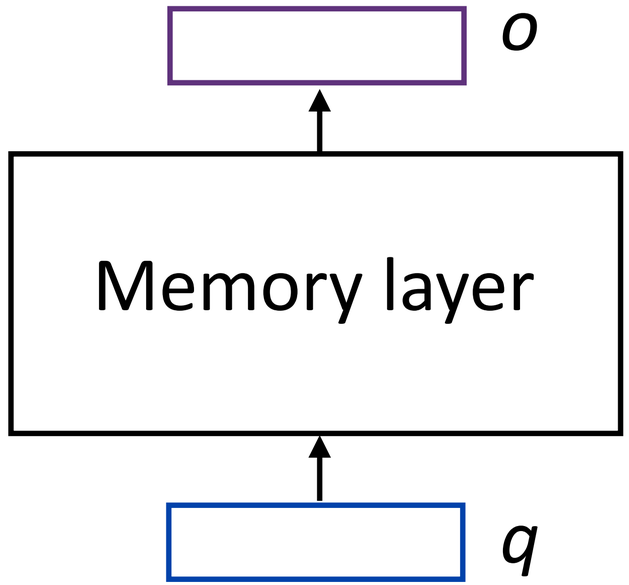


Sukhbaatar et al. (2015)

Memory Networks

- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output



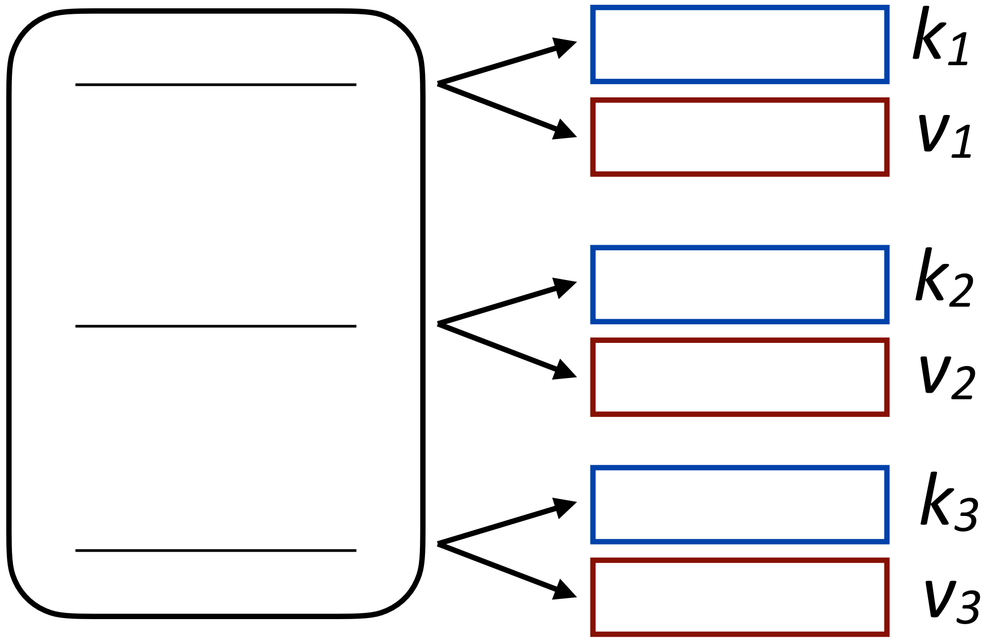


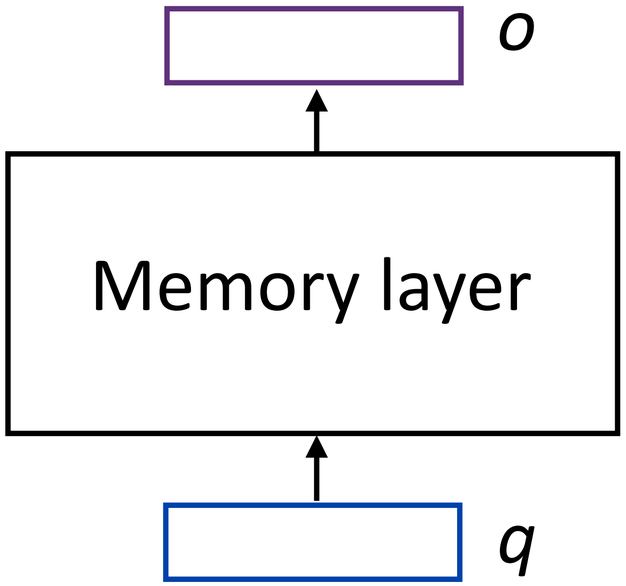
$$e_i = q \cdot k_i$$

Sukhbaatar et al. (2015)

Memory Networks

- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output





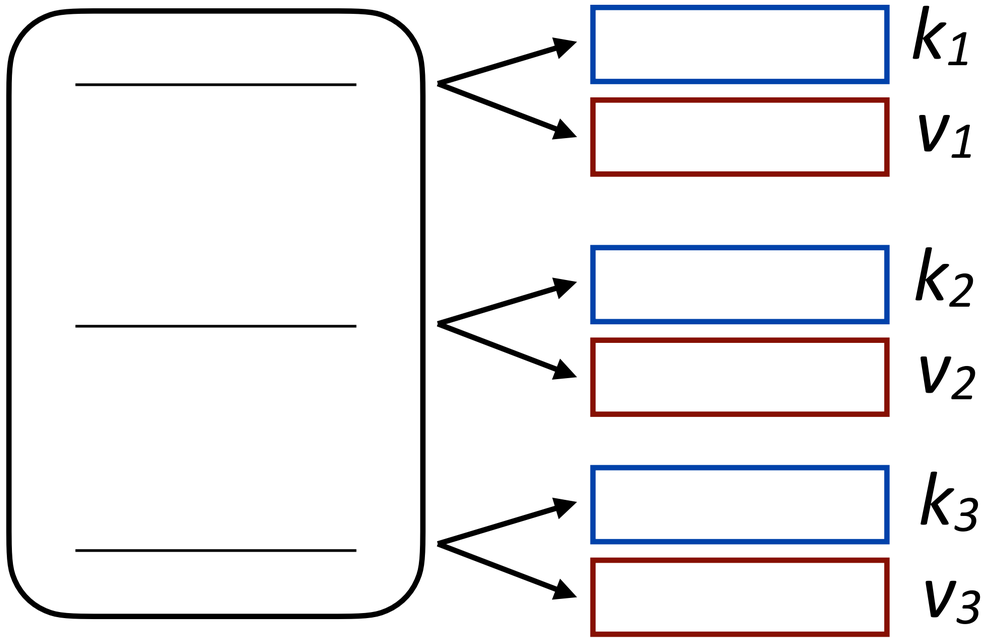
$$\alpha = \mathrm{softmax}(e)$$

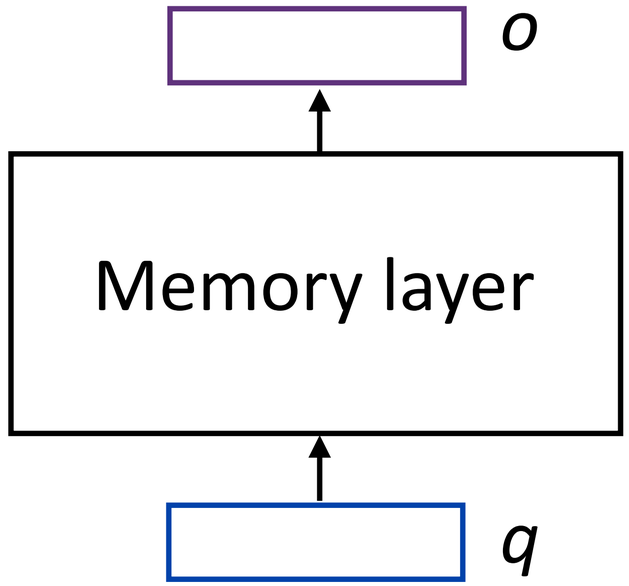
$$e_i = q \cdot k_i$$

Sukhbaatar et al. (2015)

Memory Networks

- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output





$$o = \sum_i \alpha_i v_i$$

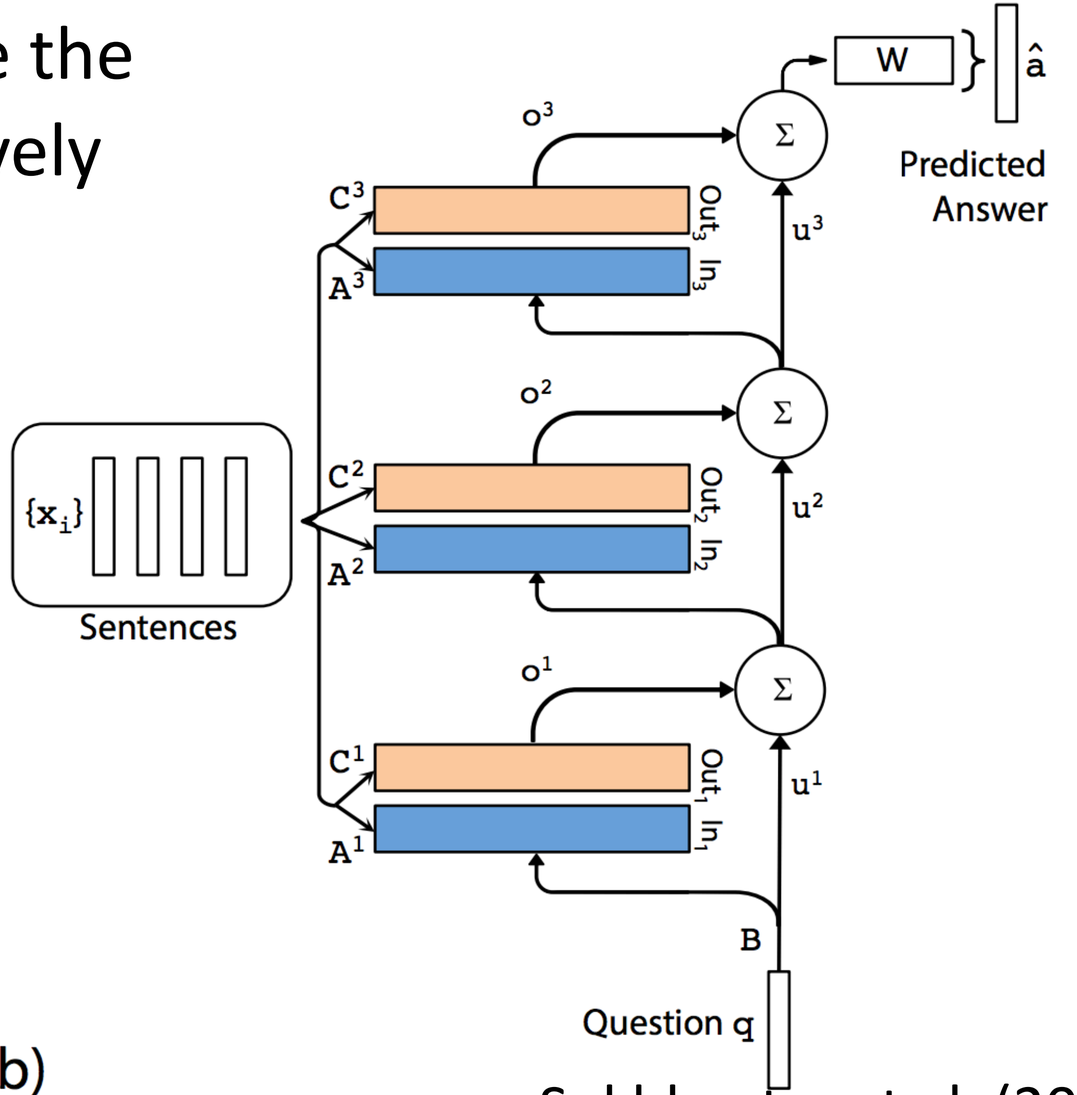
$$\alpha = \mathrm{softmax}(e)$$

$$e_i = q \cdot k_i$$

Sukhbaatar et al. (2015)

Memory Networks

- Three layers of memory network where the query representation is updated additively based on the memories at each step



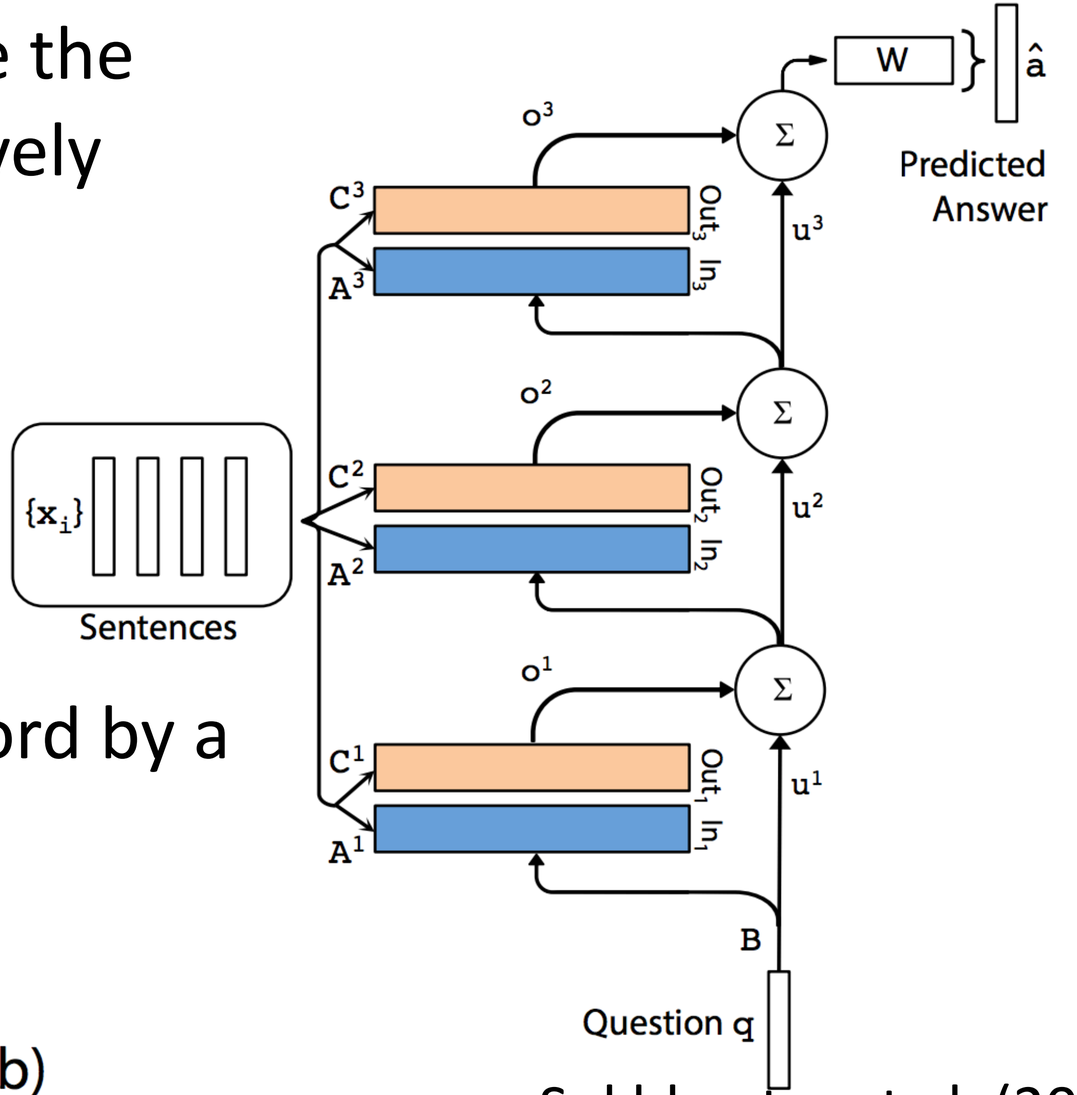
(b)

Sukhbaatar et al. (2015)

Memory Networks

- Three layers of memory network where the query representation is updated additively based on the memories at each step

- How to encode the sentences?
  - Bag of words (average embeddings)
  - Positional encoding: multiply each word by a vector capturing position in sentence



(b)

Sukhbaatar et al. (2015)

bAbI

- Evaluation on 20 tasks proposed as building blocks for building "AI-complete" systems

Task 1: Single Supporting Fact

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 13: Compound Coreference

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning

In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Weston et al. (2014)

bAbI

- Evaluation on 20 tasks proposed as building blocks for building "AI-complete" systems
- Various levels of difficulty, exhibit different linguistic phenomena

Task 1: Single Supporting Fact

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 13: Compound Coreference

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning

In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Weston et al. (2014)

bAbI

- Evaluation on 20 tasks proposed as building blocks for building "AI-complete" systems
- Various levels of difficulty, exhibit different linguistic phenomena
- Small vocabulary, language isn't truly "natural"

Task 1: Single Supporting Fact

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 13: Compound Coreference

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning

In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Weston et al. (2014)

Evaluation: bAbI

	Baseline			MemN2N				
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint
Mean error (%)	6.7	51.3	40.2	25.1	20.3	25.8	15.6	13.3
Failed tasks (err. $>$ 5%)	4	20	18	15	13	17	11	11

Evaluation: bAbI

	Baseline			MemN2N				
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint
Mean error (%)	6.7	51.3	40.2	25.1	20.3	25.8	15.6	13.3
Failed tasks (err. $>$ 5%)	4	20	18	15	13	17	11	11

- 3-hop memory network does pretty well, better than LSTM at processing these types of examples

Evaluation: bAbI

Task	Baseline			MemN2N				
	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint
Mean error (%)	6.7	51.3	40.2	25.1	20.3	25.8	15.6	13.3
Failed tasks (err. $>$ 5%)	4	20	18	15	13	17	11	11

- 3-hop memory network does pretty well, better than LSTM at processing these types of examples

Story (16: basic induction)	**Support**	**Hop 1**	**Hop 2**	**Hop 3**
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

Evaluation: Children's Book Test

Methods	Named Entities
Humans (query)(*)	0.520
Humans (context+query)(*)	***0.816***
Maximum frequency (corpus)	0.120
Maximum frequency (context)	0.335
Sliding window	0.168
Word distance model	0.398
Kneser-Ney language model	0.390
Kneser-Ney language model + cache	0.439
LSTMs (query)	0.408
LSTMs (context+query)	0.418
Contextual LSTMs (window context)	0.436
MemNNs (lexical memory)	0.431
MemNNs (window memory)	0.493
MemNNs (sentential memory + PE)	0.318
MemNNs (window memory + self-sup.)	**0.666**

Evaluation: Children's Book Test

Methods	Named Entities
Humans (query)(*)	0.520
Humans (context+query)(*)	***0.816***
Maximum frequency (corpus)	0.120
Maximum frequency (context)	0.335
Sliding window	0.168
Word distance model	0.398
Kneser-Ney language model	0.390
Kneser-Ney language model + cache	0.439
LSTMs (query)	0.408
LSTMs (context+query)	0.418
Contextual LSTMs (window context)	0.436
MemNNs (lexical memory)	0.431
MemNNs (window memory)	0.493
MemNNs (sentential memory + PE)	0.318
MemNNs (window memory + self-sup.)	**0.666**

- Outperforms LSTMs substantially with the right supervision

Memory Network Takeaways

- Memory networks provide a way of attending to abstractions over the input
- Useful for cloze tasks where far-back context is necessary
- What can we do with more basic attention?

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, makes it easier than Children's Book Test

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

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

Passage

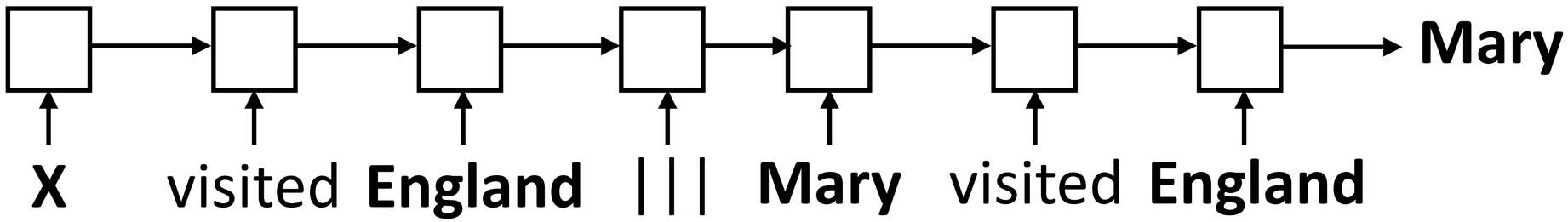
(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to 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

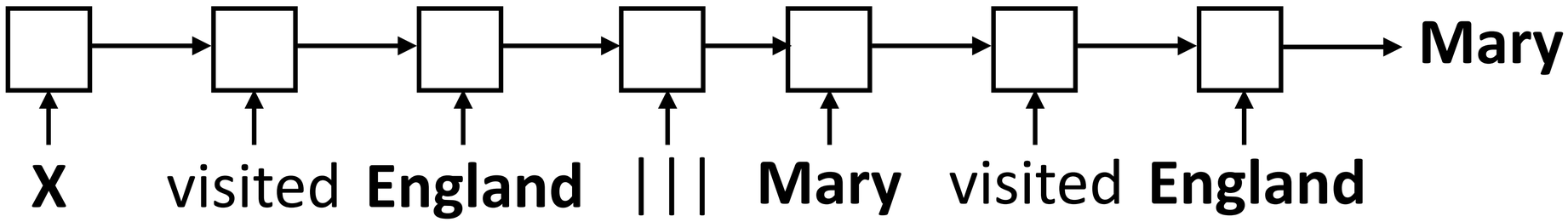
- LSTM reader: encode question, encode passage, predict entity



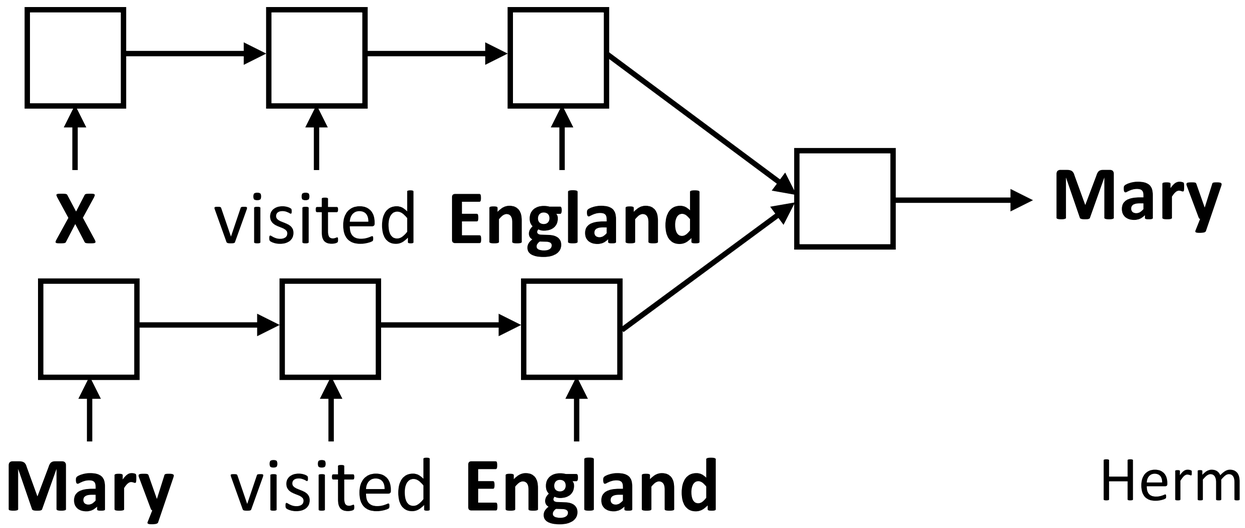
Hermann et al. (2015), Chen et al. (2016)

CNN/Daily Mail

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



- Can also use textual entailment-like models

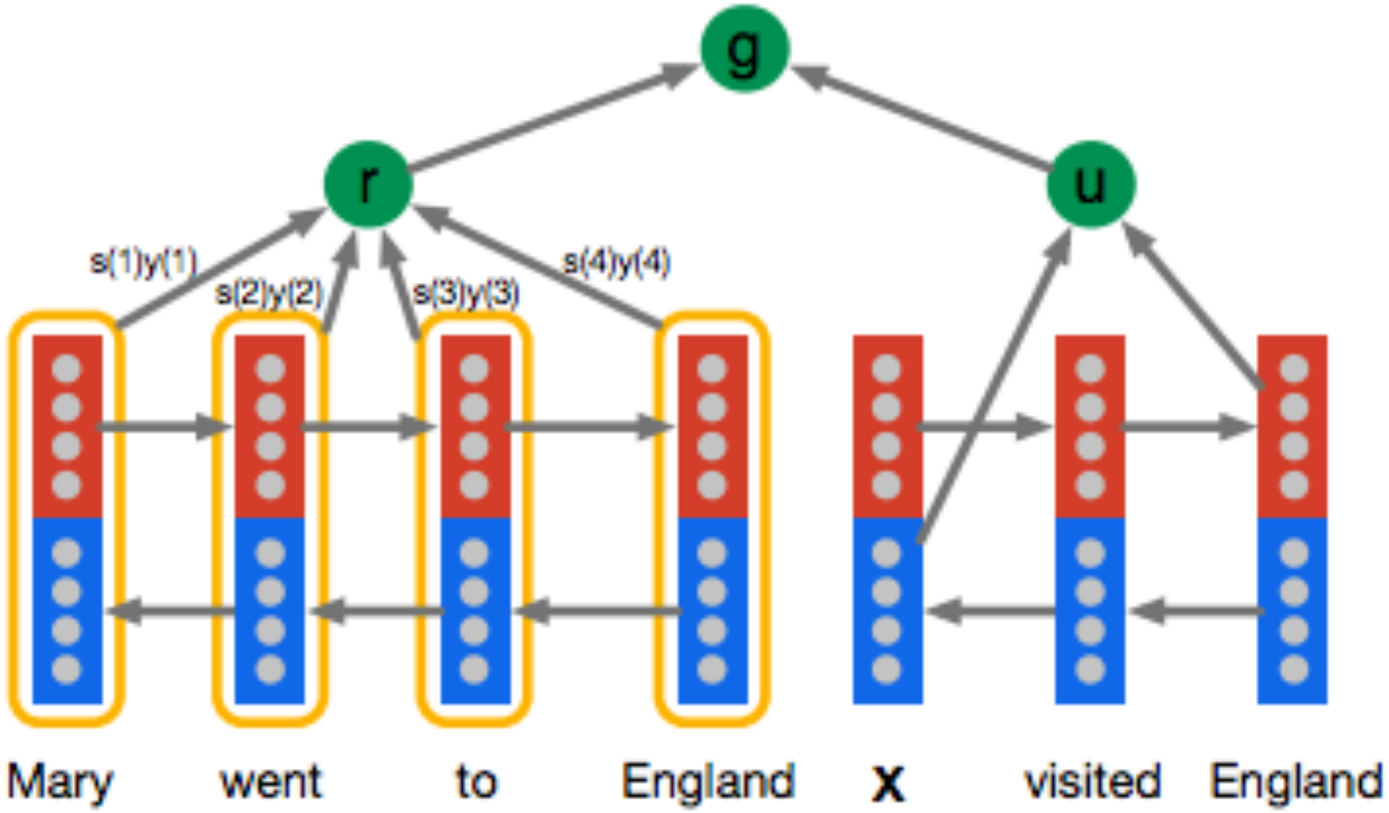


Multiclass classification problem over entities in the document

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

CNN/Daily Mail

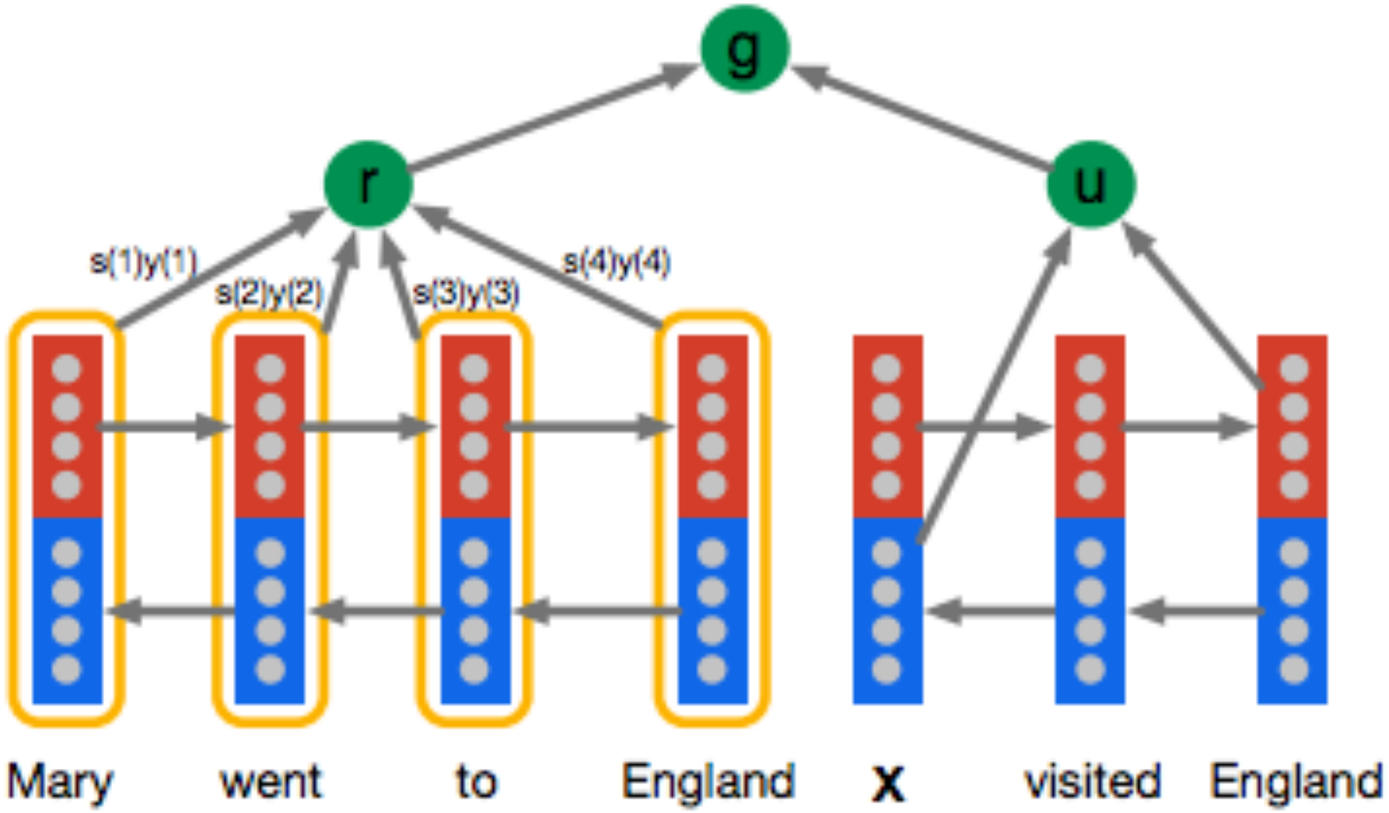
- Attentive reader:
  u = encode query
  s = encode sentence
  r = attention(u -> s)
  prediction = f(candidate, u, r)



Hermann et al. (2015)

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

	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)

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: Bidirectional Attention Flow

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

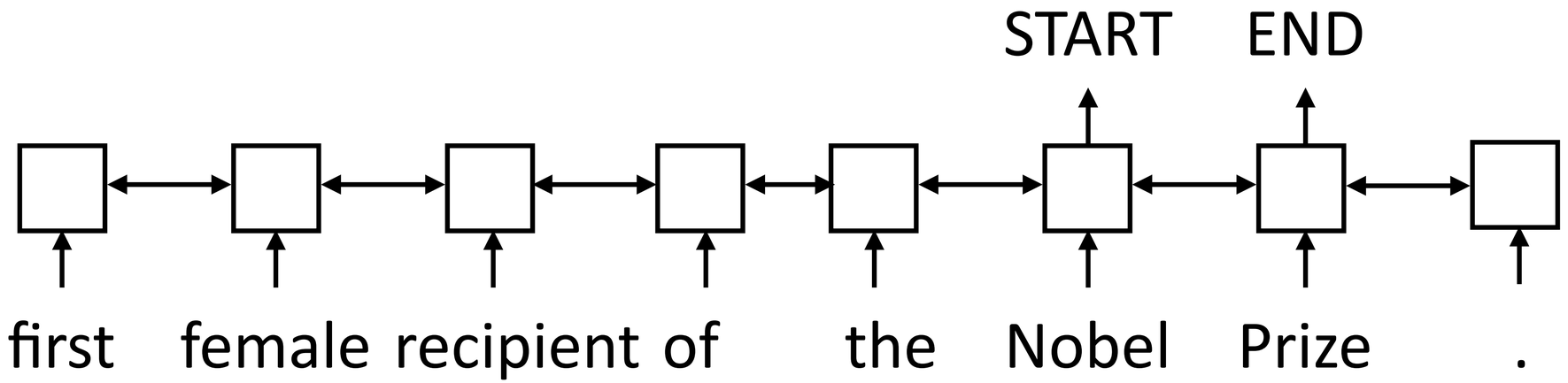
What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 | 1745 | 1745

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)

SQuAD

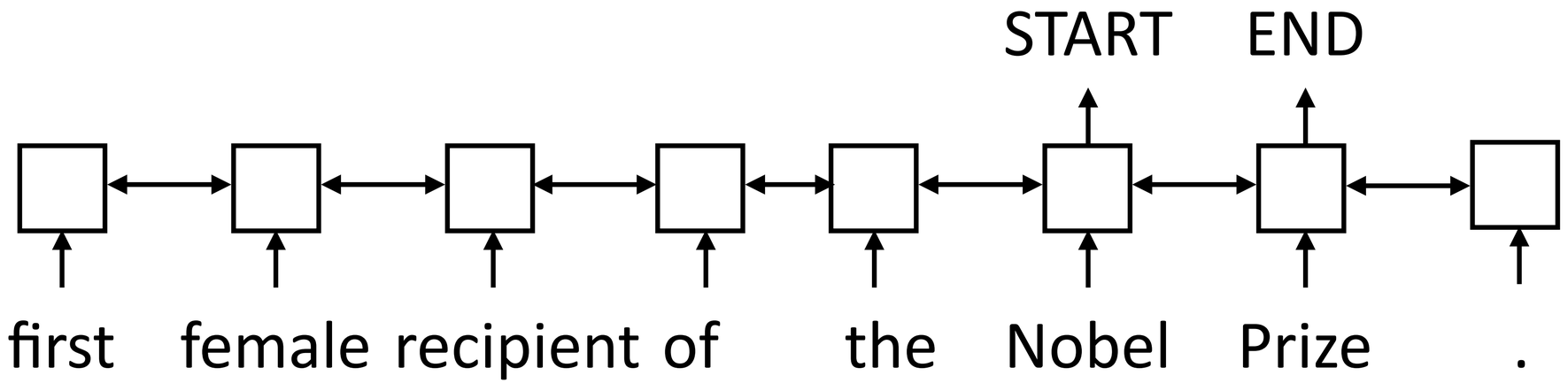
What was Marie Curie the first female recipient of?



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)

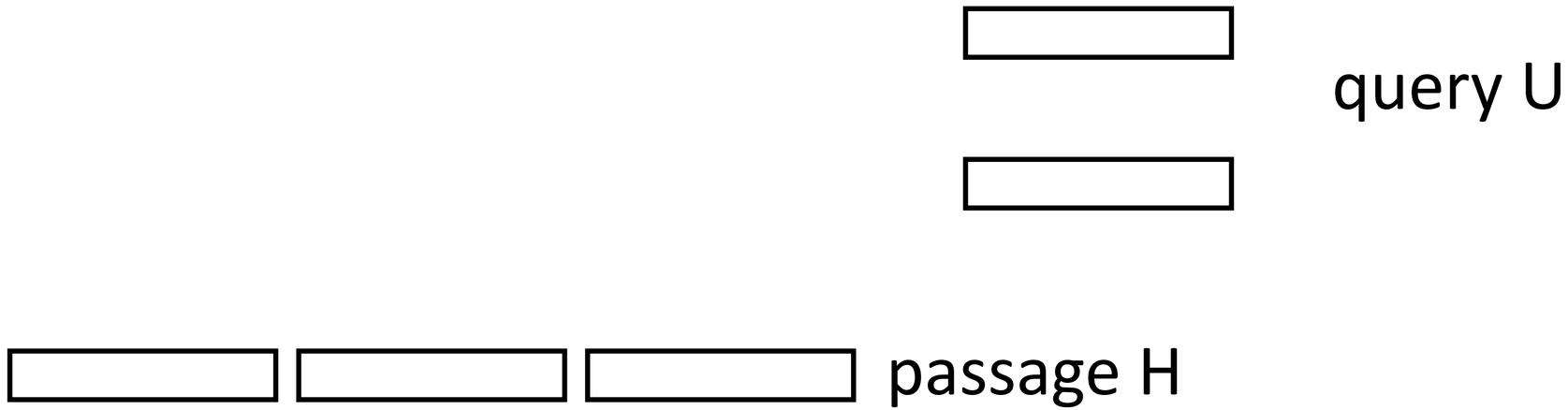
Bidirectional Attention Flow

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

Seo et al. (2016)

Bidirectional Attention Flow

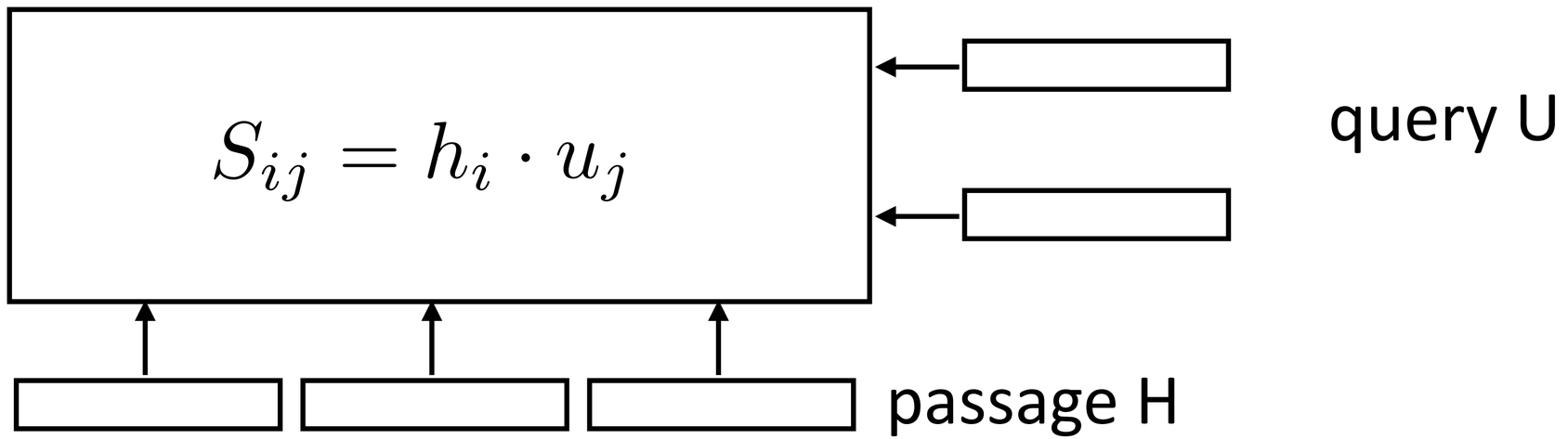
- Passage (context) and query are both encoded with BiLSTMs



Seo et al. (2016)

Bidirectional Attention Flow

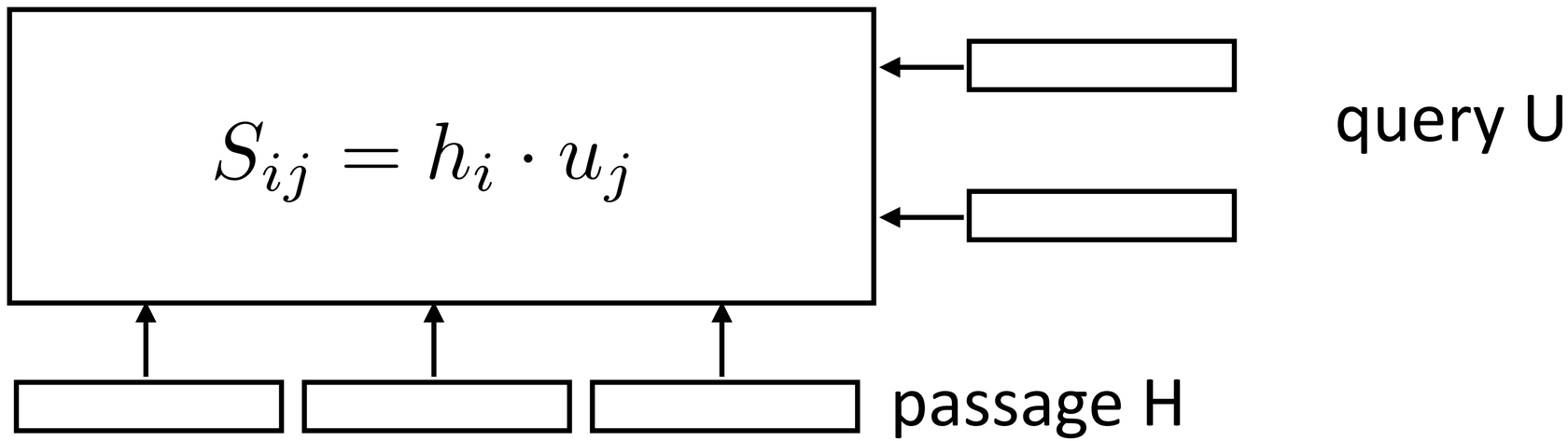
- Passage (context) and query are both encoded with BiLSTMs



Seo et al. (2016)

Bidirectional Attention Flow

- 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

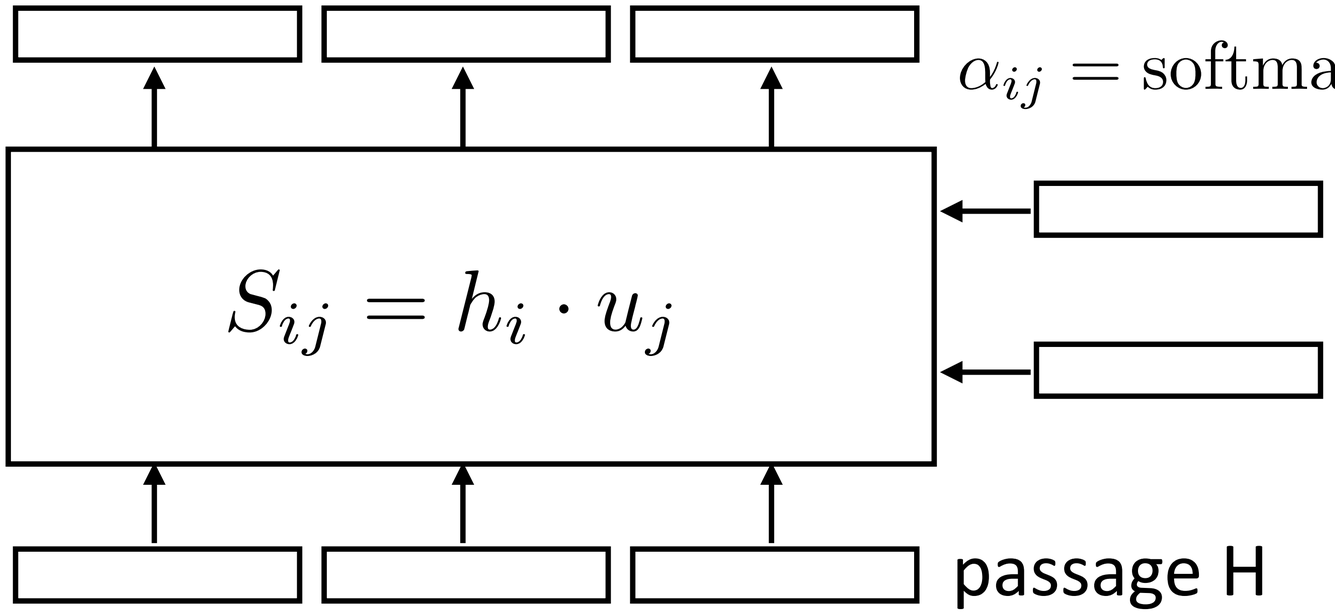
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

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

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

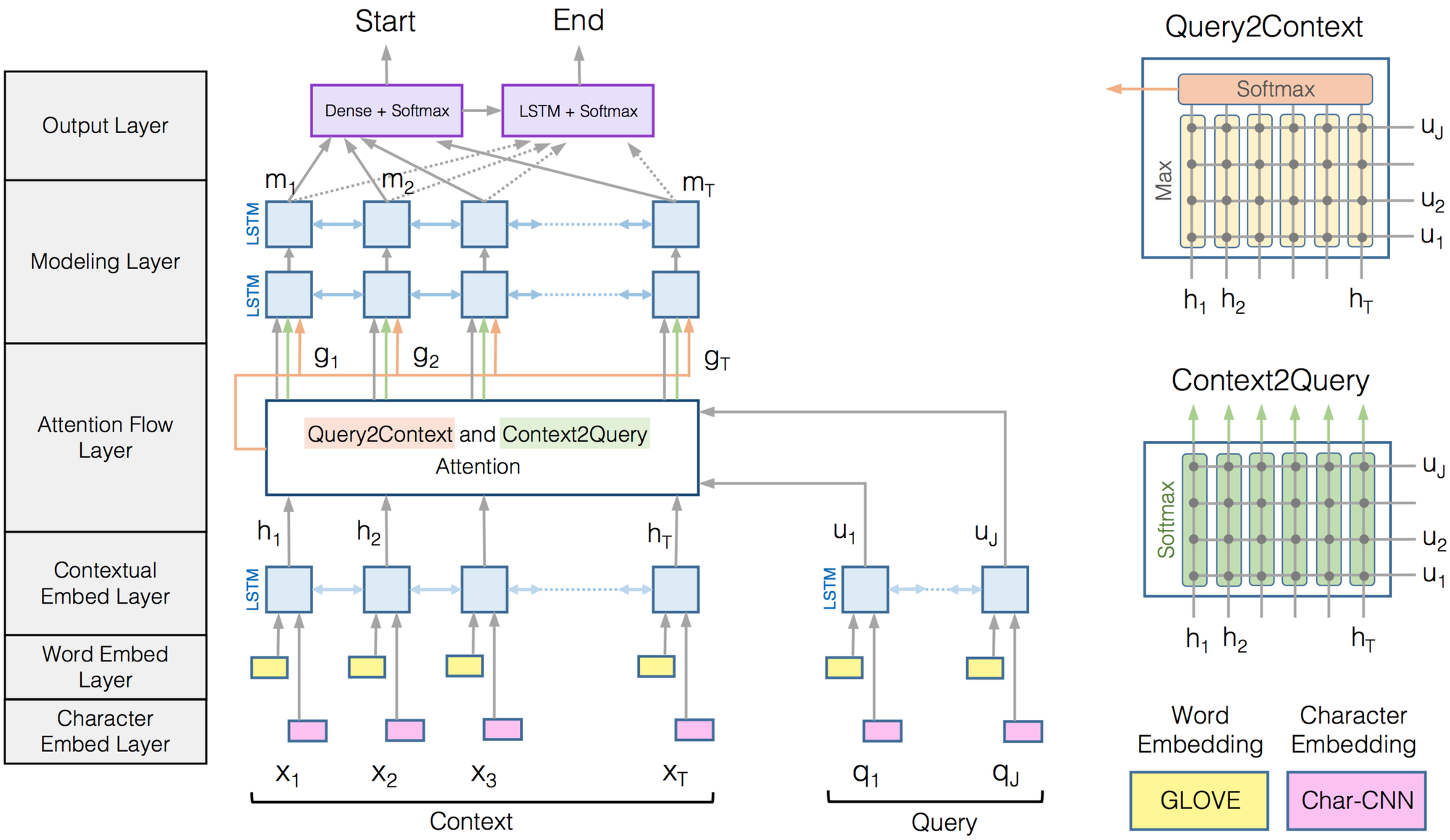
$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- dist over query



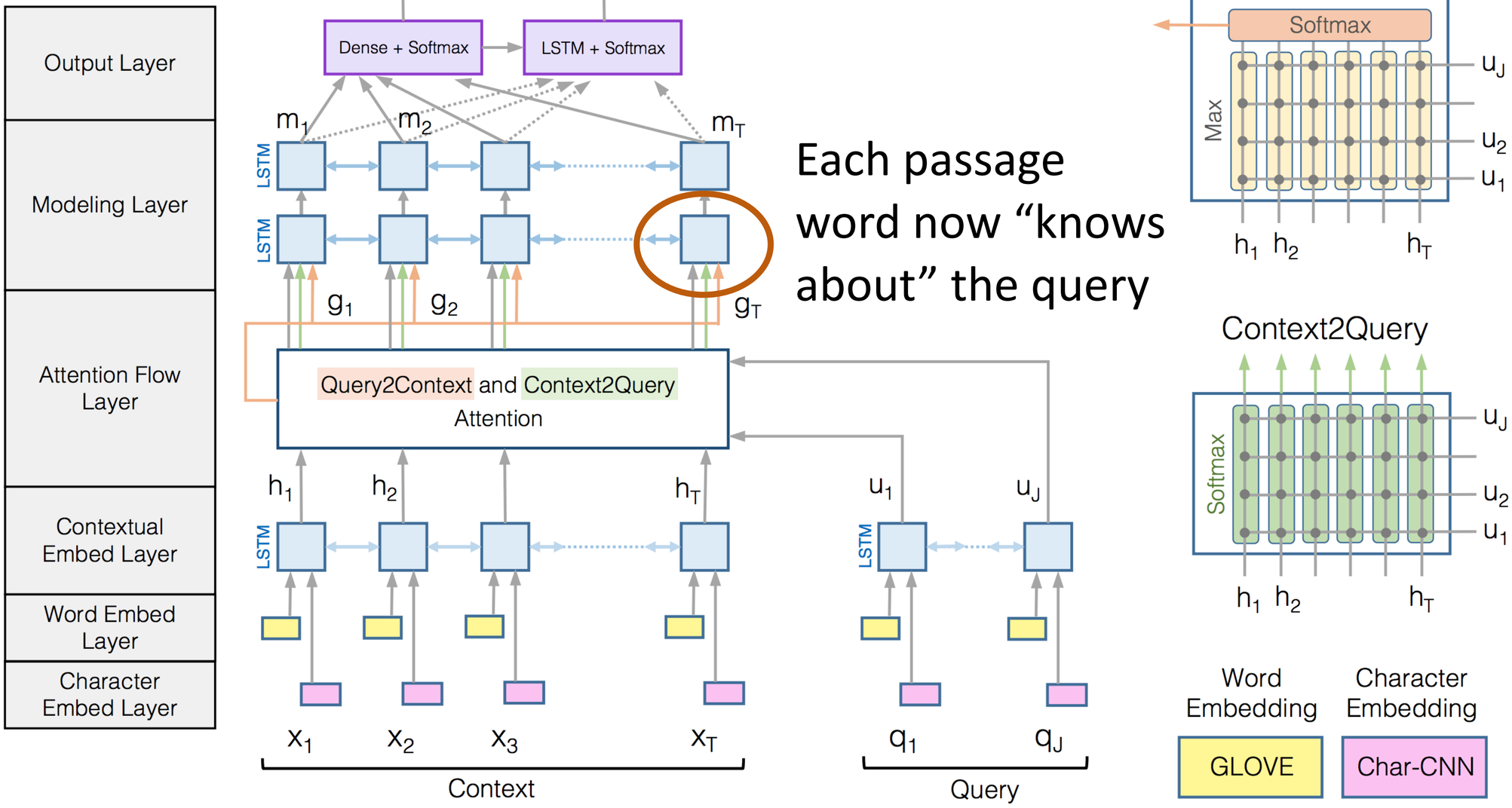
Seo et al. (2016)

Bidirectional Attention Flow



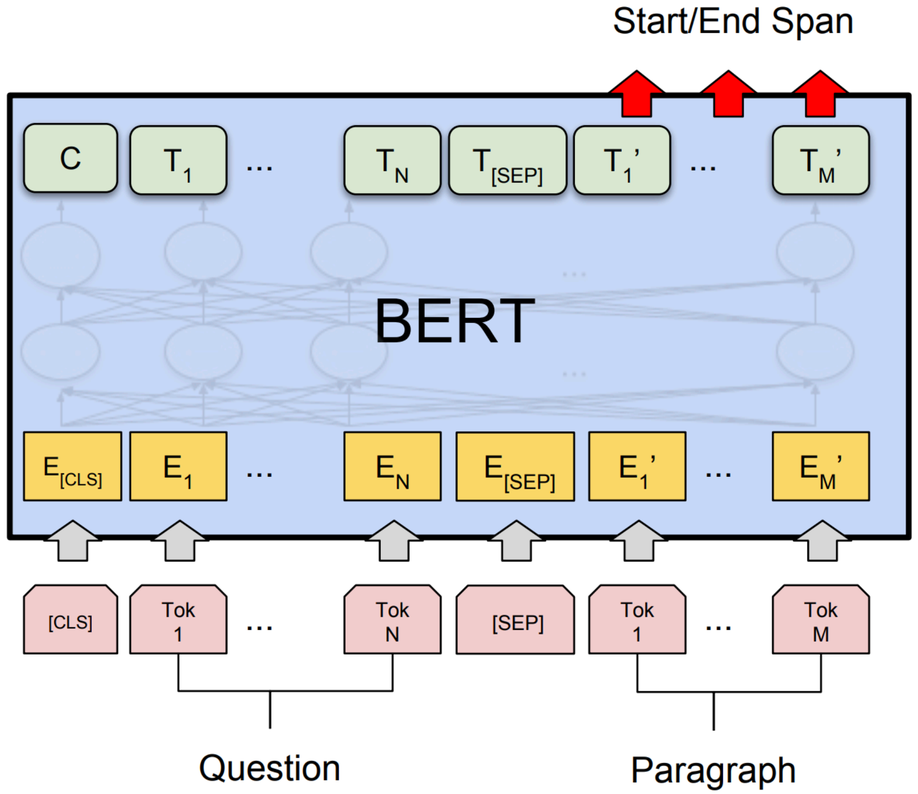
Seo et al. (2016)

Bidirectional Attention Flow



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 Today

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
4 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
4 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899
5 Dec 01, 2020	EntitySpanFocusV2 (ensemble) *RICOH_SRCB_DML*	90.521	92.824
5 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
5 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839

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)

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

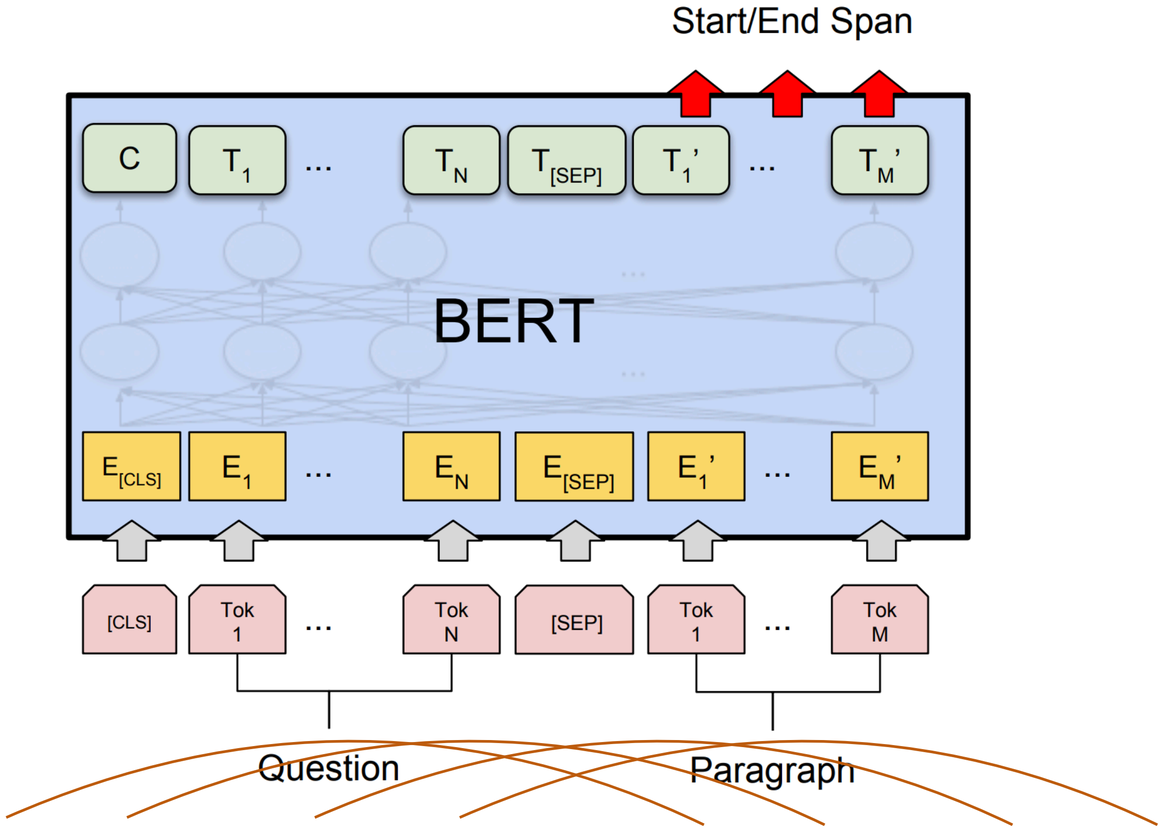
Problems in QA

Adversarial SQuAD

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

Jia and Liang (2017)

Adversarial SQuAD

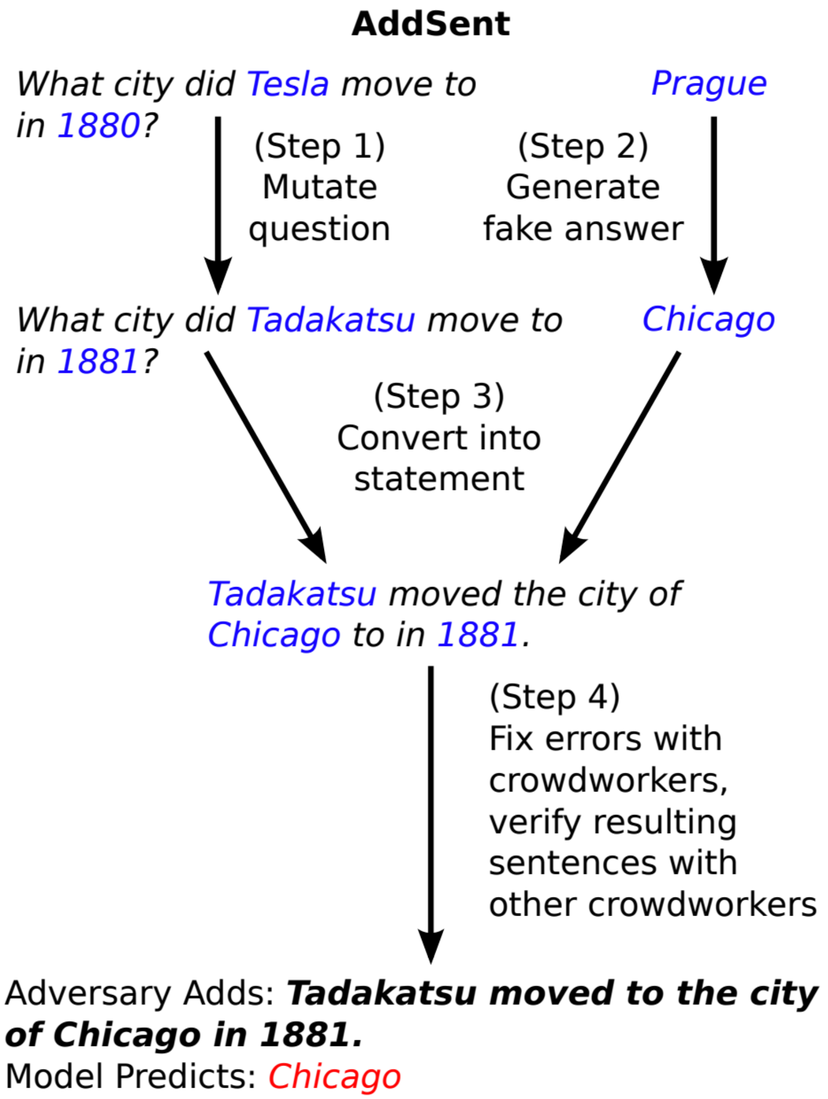


What was Marie Curie the first female recipient of ? [SEP] … first female recipient of **the Nobel Prize** …

- 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	+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

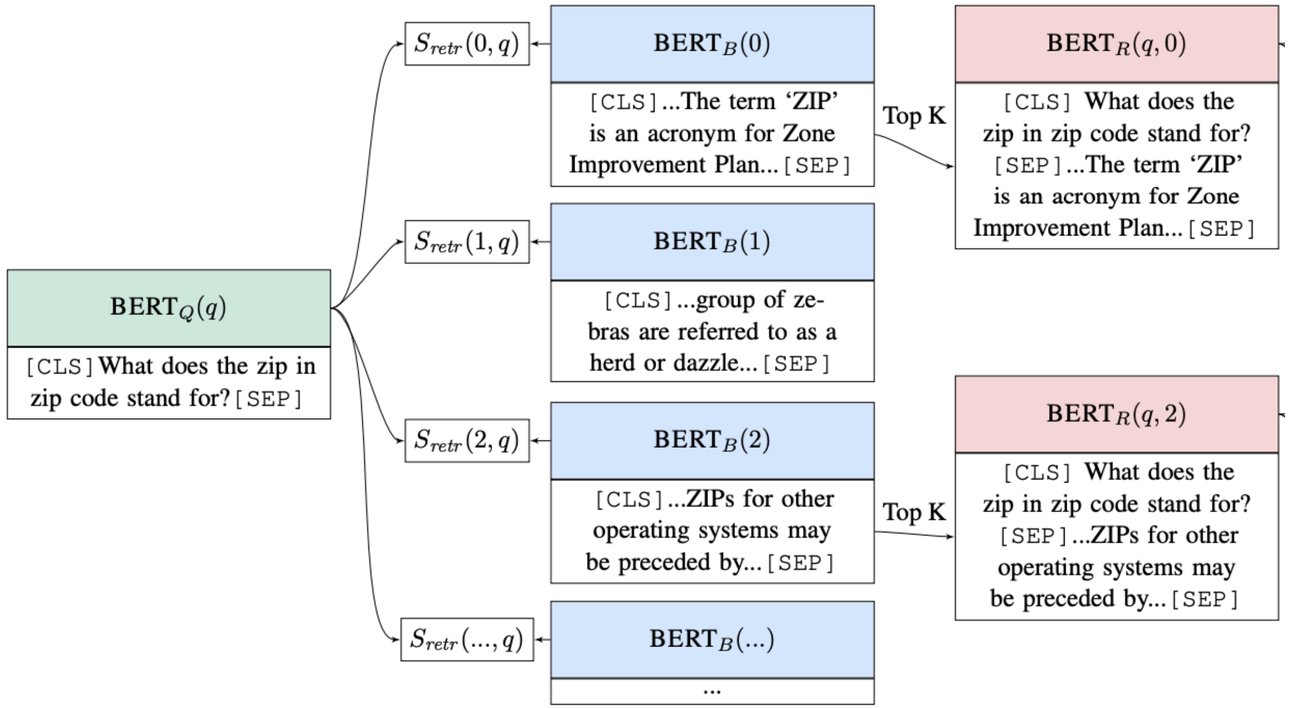
Dataset	SQuAD
SQuAD *(All Wikipedia)*	27.1
CuratedTREC	19.7
WebQuestions	11.8
WikiMovies	24.5

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

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



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?*
  - *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)

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

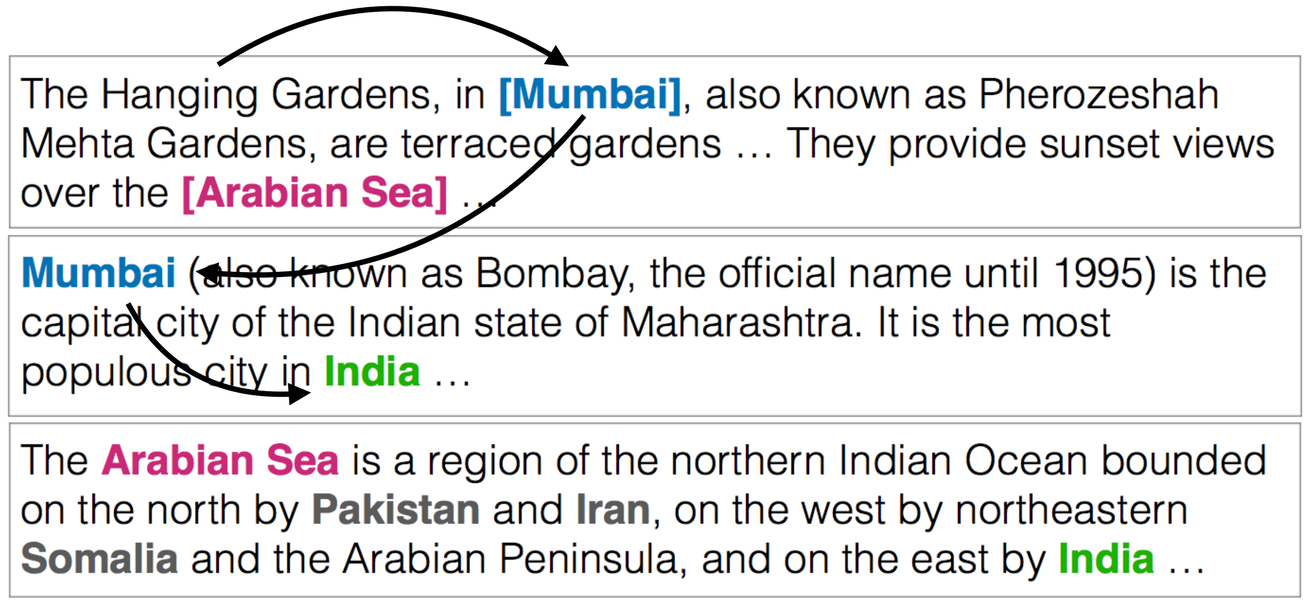


Figure from Welbl et al. (2018)

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.

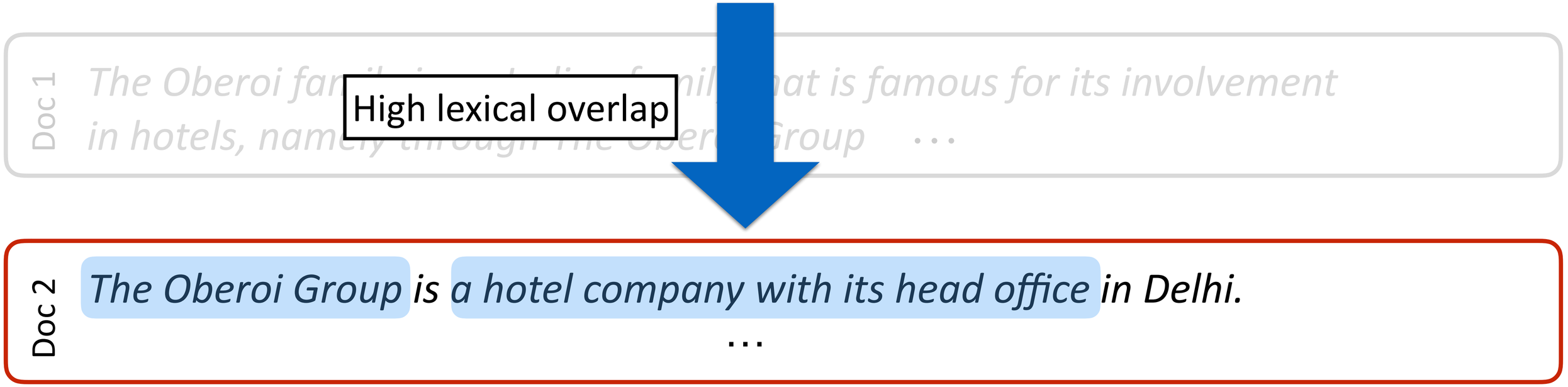
...

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?*



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