# 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"

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

(other representations like SQL possible too...)

# Classical Question Answering

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

Q: “where was Barack Obama born”

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

(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 |
|---|---|---|---|
| $N/N$ | $N$ | $N\backslash N/NP$ | $NP$ |
| $\lambda f \lambda x. f(x) \wedge former(x)$ | $\lambda x. municipalities(x)$ | $\lambda f \lambda x \lambda y. f(y) \wedge in(y, x)$ | $Brandenburg$ |

$$\xrightarrow{>} \quad N \quad \lambda x. former(x) \wedge municipalities(x)$$

$$\xrightarrow{>} \quad N\backslash N \quad \lambda f \lambda y. f(y) \wedge in(y, Brandenburg)$$

$$\xrightarrow{<} \quad N$$

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

(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.

$$\dfrac{\dfrac{\dfrac{\text{Former}}{N/N \quad \lambda f \lambda x.f(x) \wedge former(x)} \quad \dfrac{\text{municipalities}}{N \quad \lambda x.municipalities(x)}}{N \quad \lambda x.former(x) \wedge municipalities(x)}> \quad \dfrac{\dfrac{\text{in}}{N\backslash N/NP \quad \lambda f \lambda x \lambda y.f(y) \wedge in(y,x)} \quad \dfrac{\text{Brandenburgh}}{NP \quad Brandenburg}}{N\backslash N \quad \lambda f \lambda y.f(y) \wedge in(y, Brandenburg)}>}{N \quad l_0 = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)}<$$

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

$$\mathbf{l}_0 = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$
$$\mathbf{l}_1 = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, \mathtt{Brandenburg})$$
$$\mathbf{l}_2 = \lambda x.former(x) \wedge municipalities(x) \wedge \mathtt{location.containedby}(x, \mathtt{Brandenburg})$$
$$\mathbf{l}_3 = \lambda x.former(x) \wedge \mathtt{OpenRel}(x, \mathtt{Municipality}) \wedge \mathtt{location.containedby}(x, \mathtt{Brandenburg})$$
$$\mathbf{l}_4 = \lambda x.\mathtt{OpenType}(x) \wedge \boxed{\mathtt{OpenRel}(x, \mathtt{Municipality})} \wedge \mathtt{location.containedby}(x, \mathtt{Brandenburg})$$

- 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. <u>He went to the grocery store</u> and pulled all the pudding off the shelves and ate two jars. Then <u>he walked to the fast food restaurant</u> 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 <u>pulled all the pudding off the shelves</u> 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 <u>pull off of the shelves</u> 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

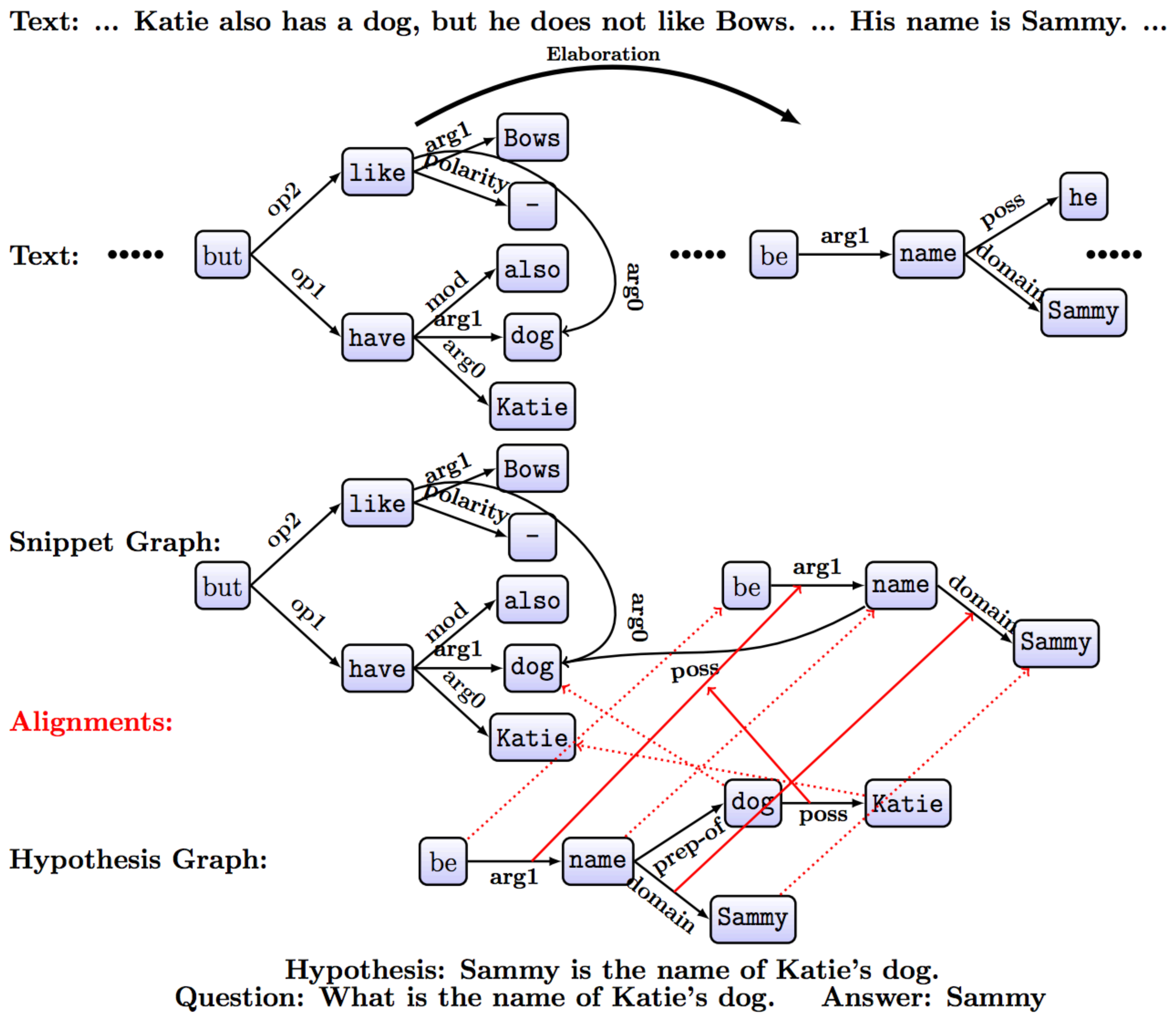
ngram sliding window

| | 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




- 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 .

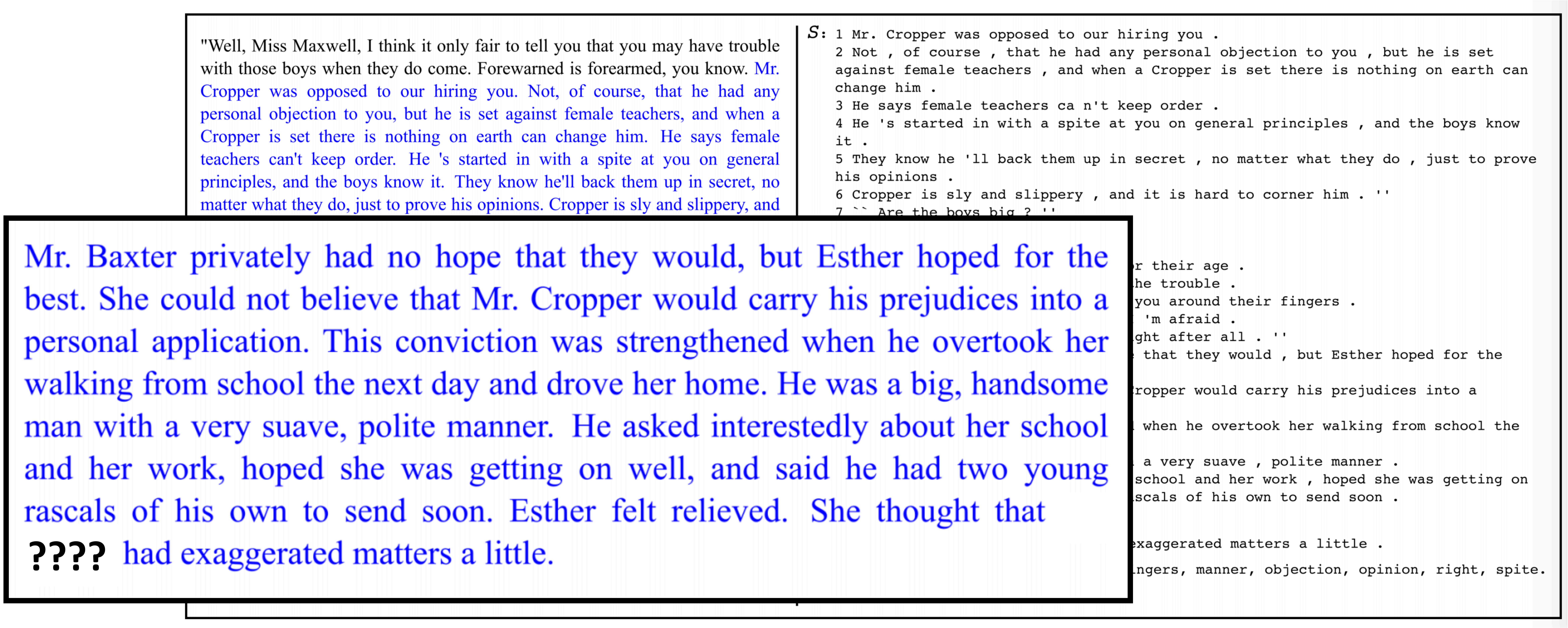
*C*: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

*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



"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

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 . ''

r their age .
he trouble .
you around their fingers .
'm afraid .
ght after all . ''
that they would , but Esther hoped for the
ropper would carry his prejudices into a
when he overtook her walking from school the
a very suave , polite manner .
school and her work , hoped she was getting on
scals of his own to send soon .
xaggerated matters a little .
ngers, manner, objection, opinion, right, spite.

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

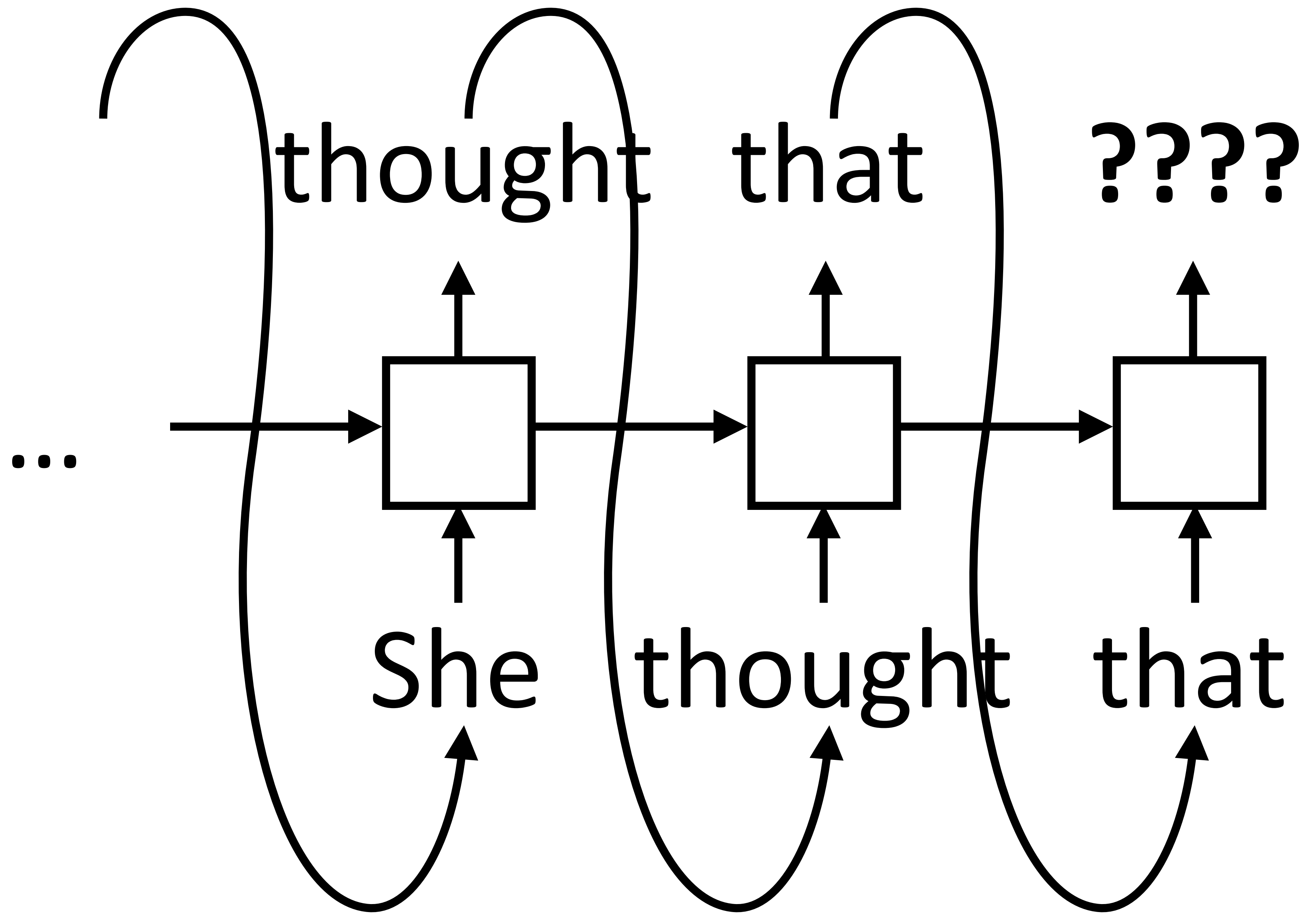
**????** had exaggerated matters a little.

- 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.

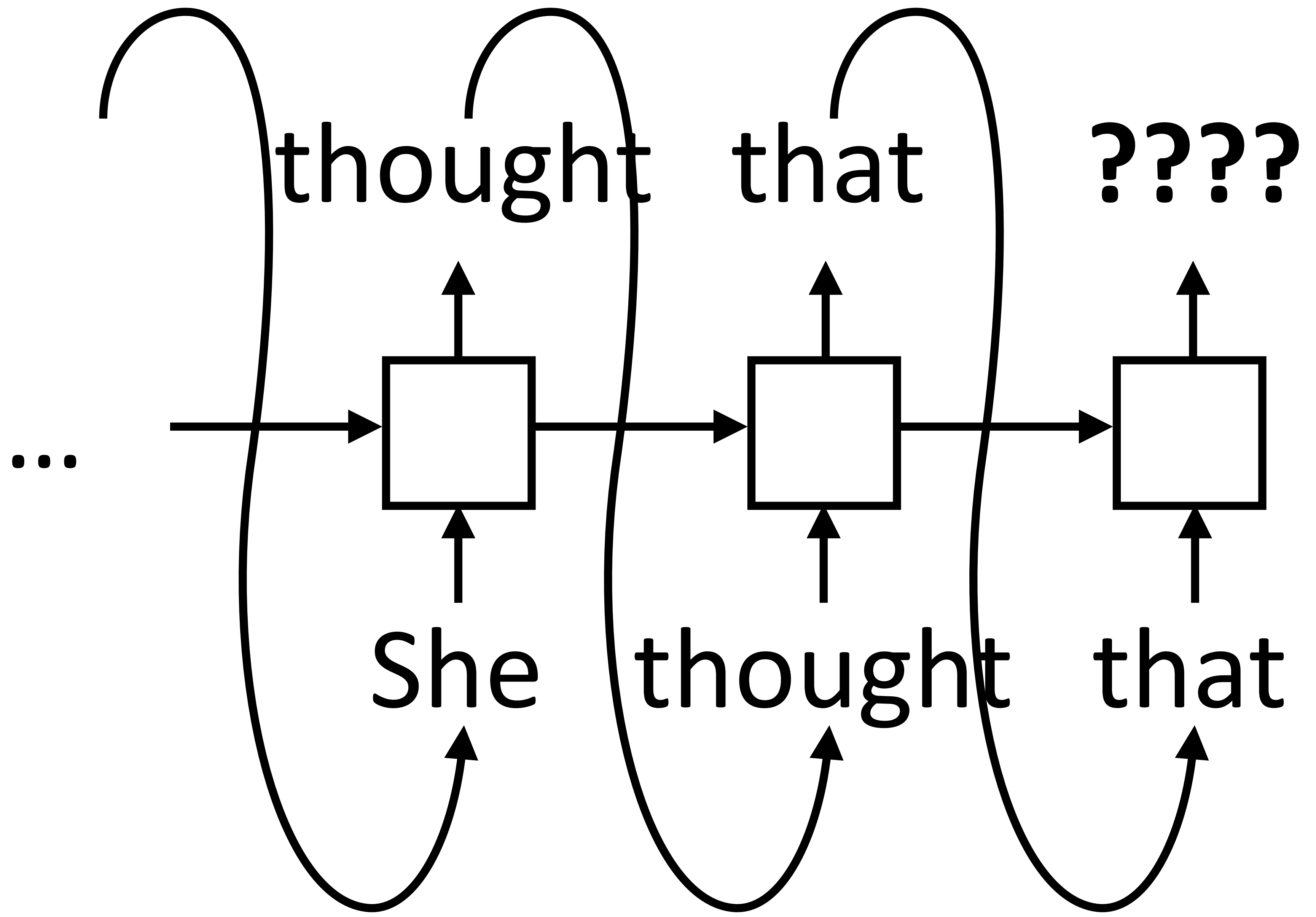



- Predict next word with LSTM LM

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)

# Children's Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

Hill et al. (2015)

# Children's Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

| 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 |

Hill et al. (2015)

# Children's Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

| 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 |

- Neural LMs aren't better than n-gram LMs

Hill et al. (2015)

# Children's Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

| Methods | Named Entities | Common Nouns | Verbs | Prepositions |
|---|---|---|---|---|
| Humans (query)(*) | 0.520 | 0.644 | 0.716 | 0.676 |
| Humans (context+query)(*) | ***0.816*** | ***0.816*** | ***0.828*** | 0.708 |
| Maximum frequency (corpus) | 0.120 | 0.158 | 0.373 | 0.315 |
| Maximum frequency (context) | 0.335 | 0.281 | 0.285 | 0.275 |
| Sliding window | 0.168 | 0.196 | 0.182 | 0.101 |
| Word distance model | 0.398 | 0.364 | 0.380 | 0.237 |
| Kneser-Ney language model | 0.390 | 0.544 | 0.778 | 0.768 |
| Kneser-Ney language model + cache | 0.439 | 0.577 | 0.772 | 0.679 |
| LSTMs (query) | 0.408 | 0.541 | 0.813 | 0.802 |
| LSTMs (context+query) | 0.418 | 0.560 | **0.818** | 0.791 |

Hill et al. (2015)

# Children's Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

| Methods | Named Entities | Common Nouns | Verbs | Prepositions |
|---|---|---|---|---|
| Humans (query)(*) | 0.520 | 0.644 | 0.716 | 0.676 |
| Humans (context+query)(*) | ***0.816*** | ***0.816*** | ***0.828*** | 0.708 |
| Maximum frequency (corpus) | 0.120 | 0.158 | 0.373 | 0.315 |
| Maximum frequency (context) | 0.335 | 0.281 | 0.285 | 0.275 |
| Sliding window | 0.168 | 0.196 | 0.182 | 0.101 |
| Word distance model | 0.398 | 0.364 | 0.380 | 0.237 |
| Kneser-Ney language model | 0.390 | 0.544 | 0.778 | 0.768 |
| Kneser-Ney language model + cache | 0.439 | 0.577 | 0.772 | 0.679 |
| LSTMs (query) | 0.408 | 0.541 | 0.813 | 0.802 |
| LSTMs (context+query) | 0.418 | 0.560 | **0.818** | 0.791 |

- Why are these results so low?

Hill et al. (2015)

# Memory Networks

# Memory Networks

- Memory networks let you reference input with attention

Sukhbaatar et al. (2015)

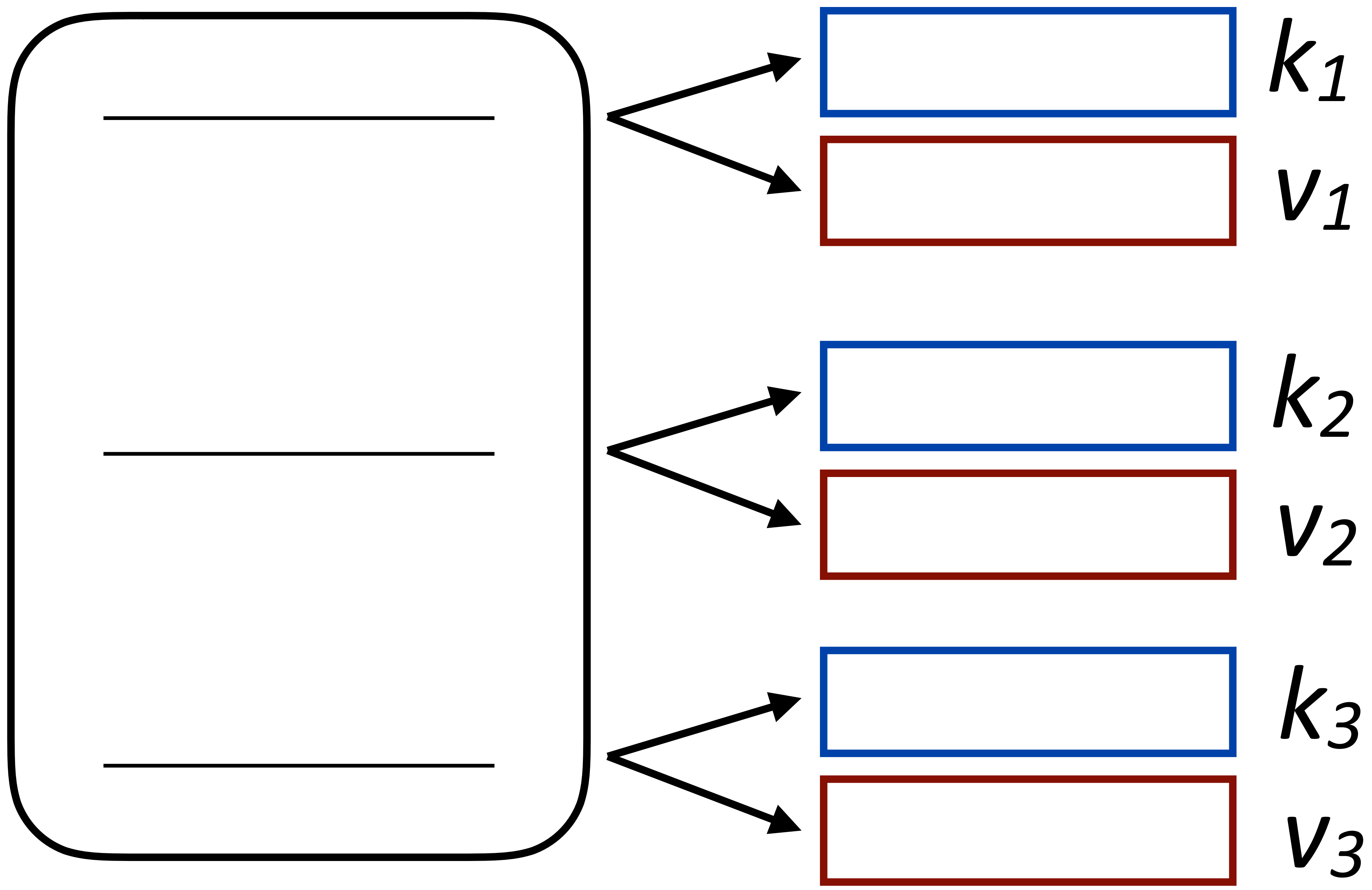
# Memory Networks

- 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

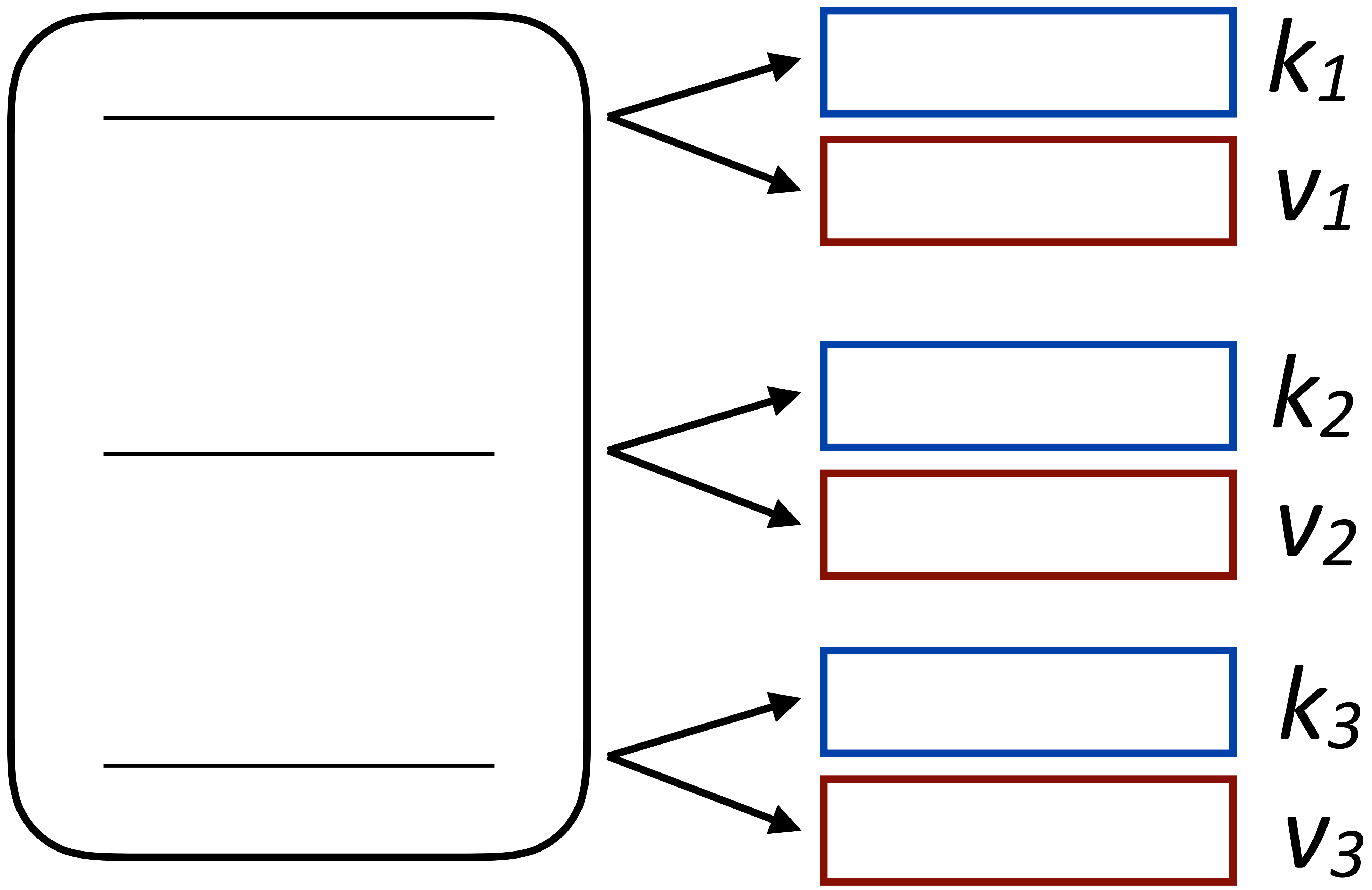
- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value




Sukhbaatar et al. (2015)

# Memory Networks

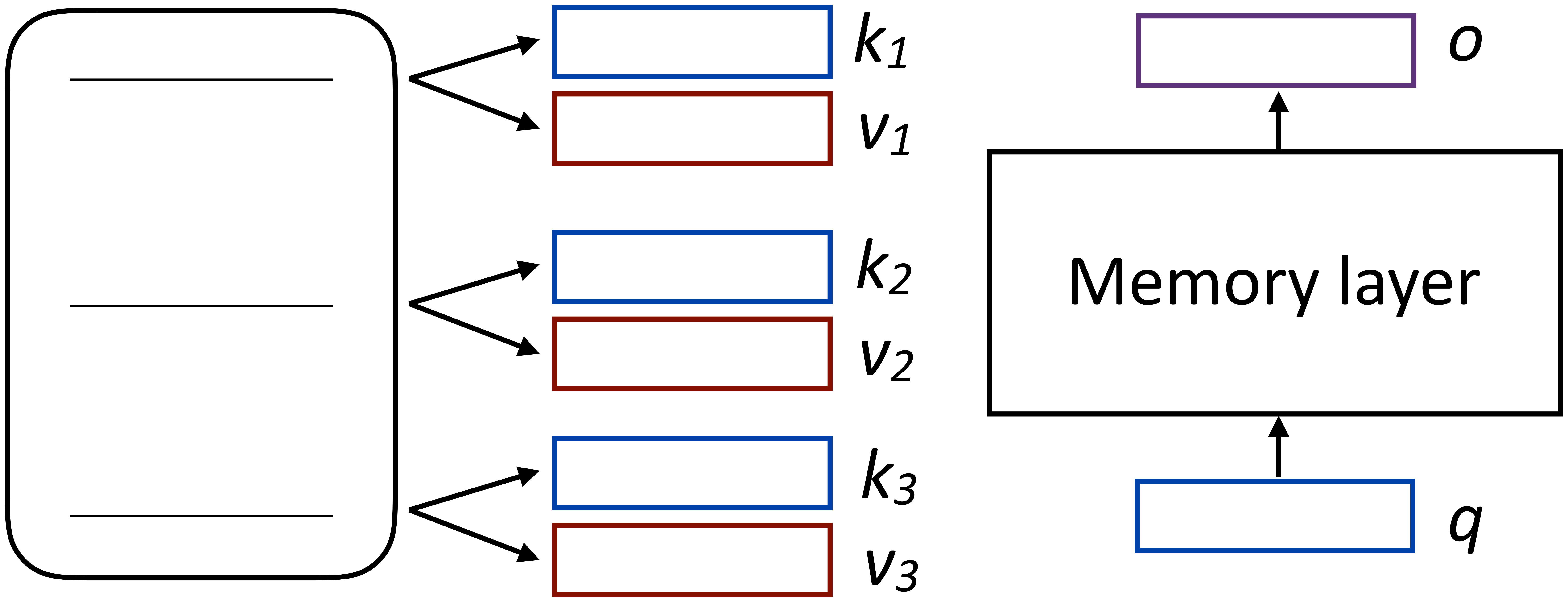
- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output




Sukhbaatar et al. (2015)

# Memory Networks

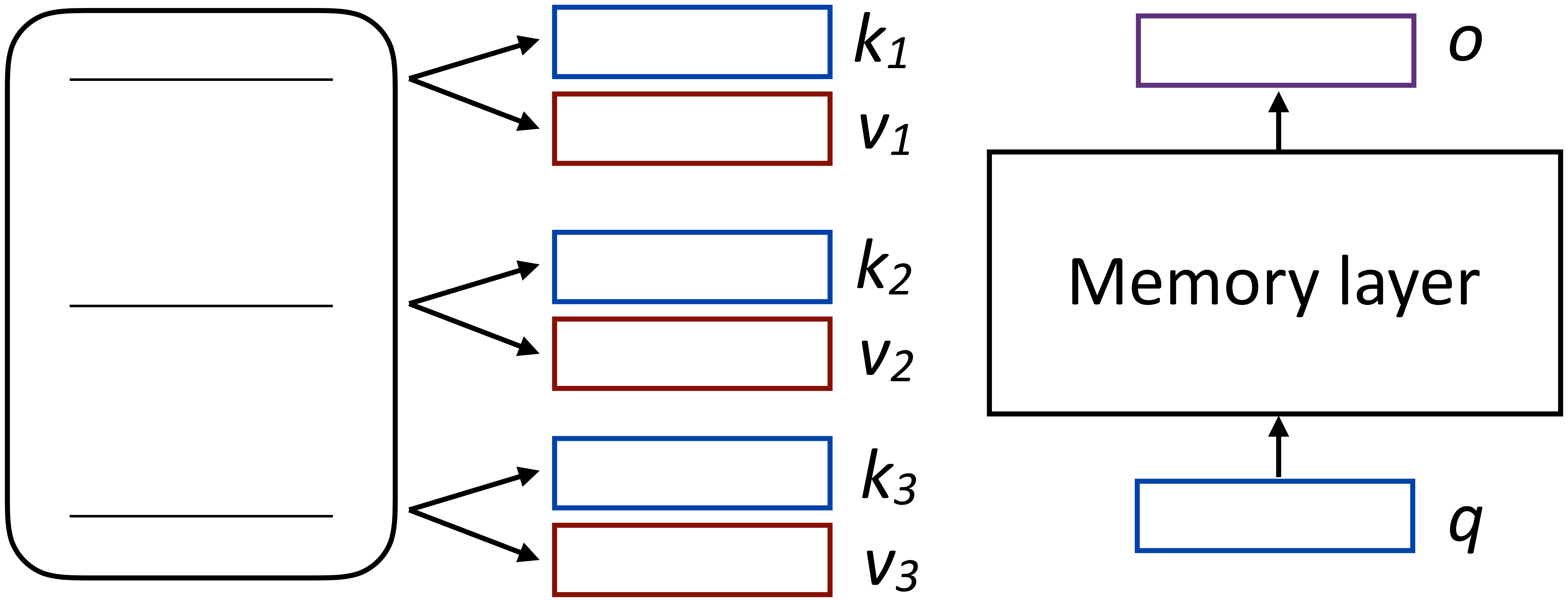
- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output




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

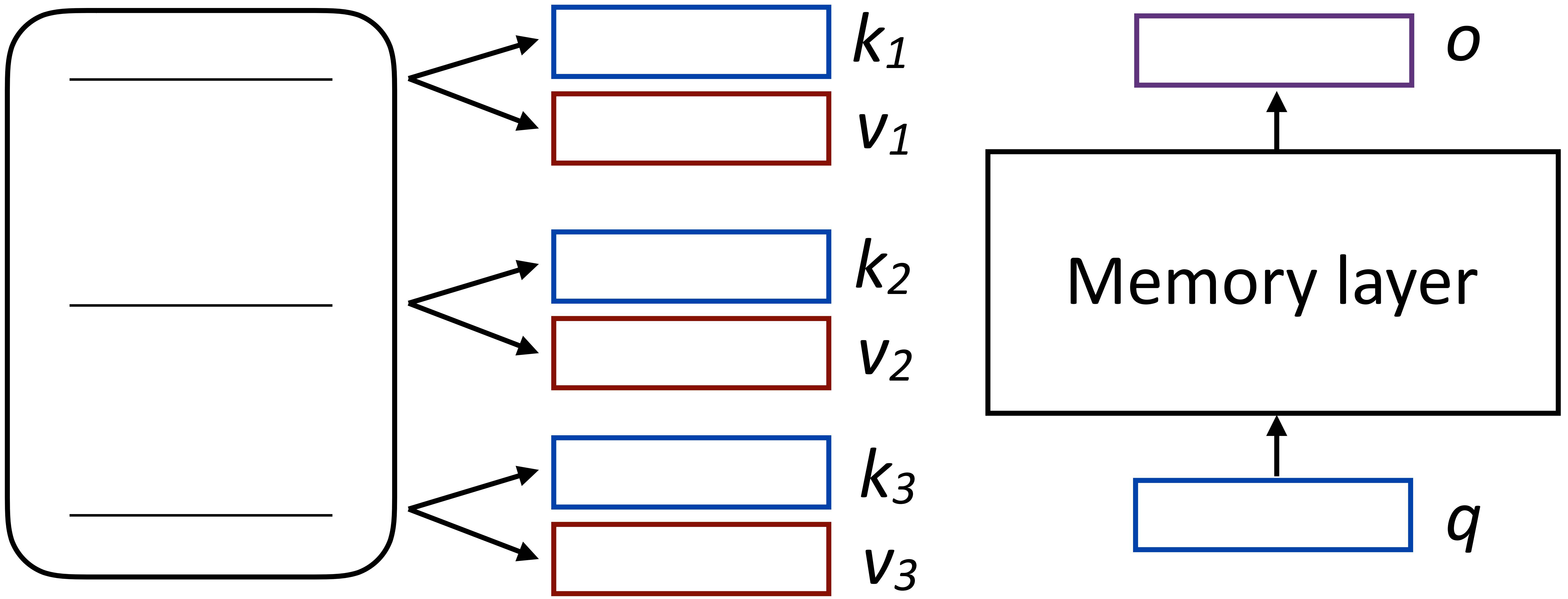



$$e_i = q \cdot k_i$$

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



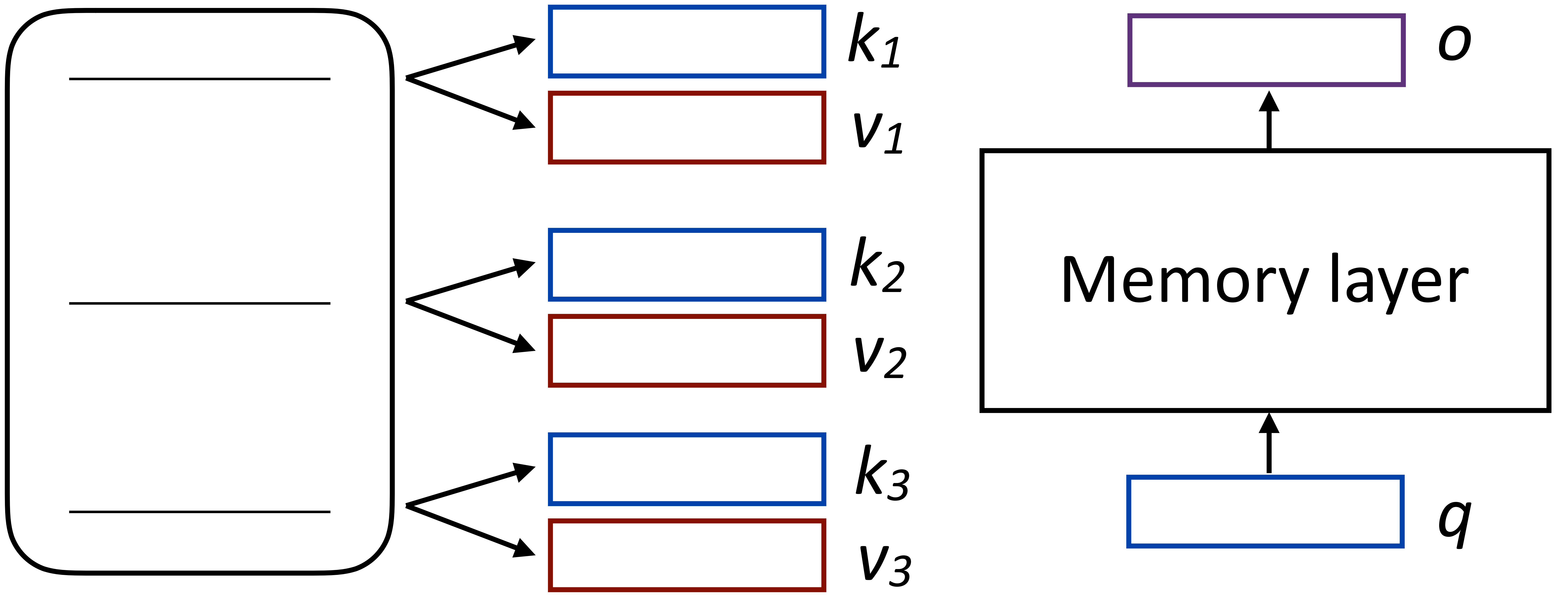

$$\alpha = \mathrm{softmax}(e)$$

$$e_i = q \cdot k_i$$

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




$$o = \sum_i \alpha_i v_i$$

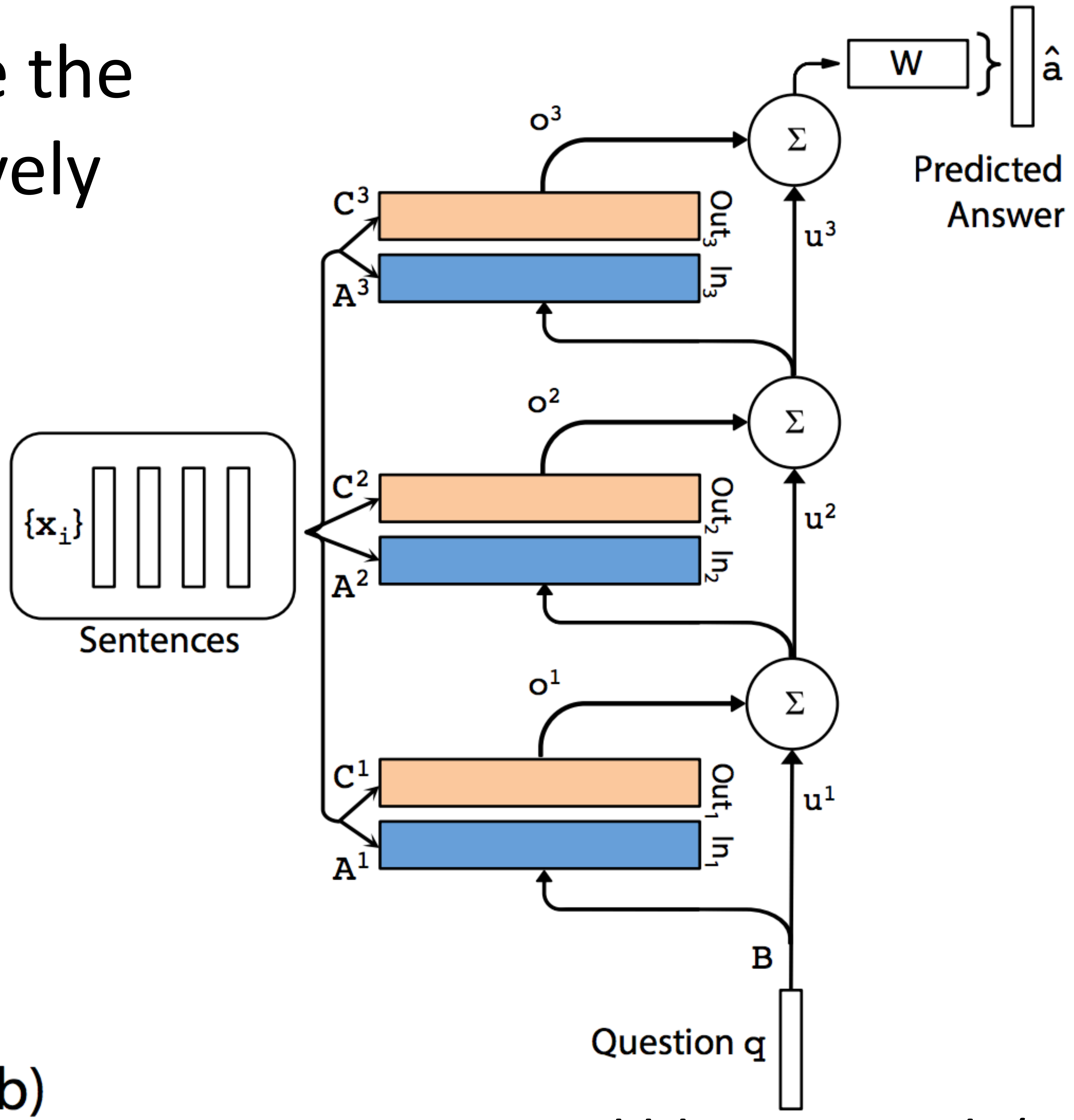
$$\alpha = \mathrm{softmax}(e)$$

$$e_i = q \cdot k_i$$

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



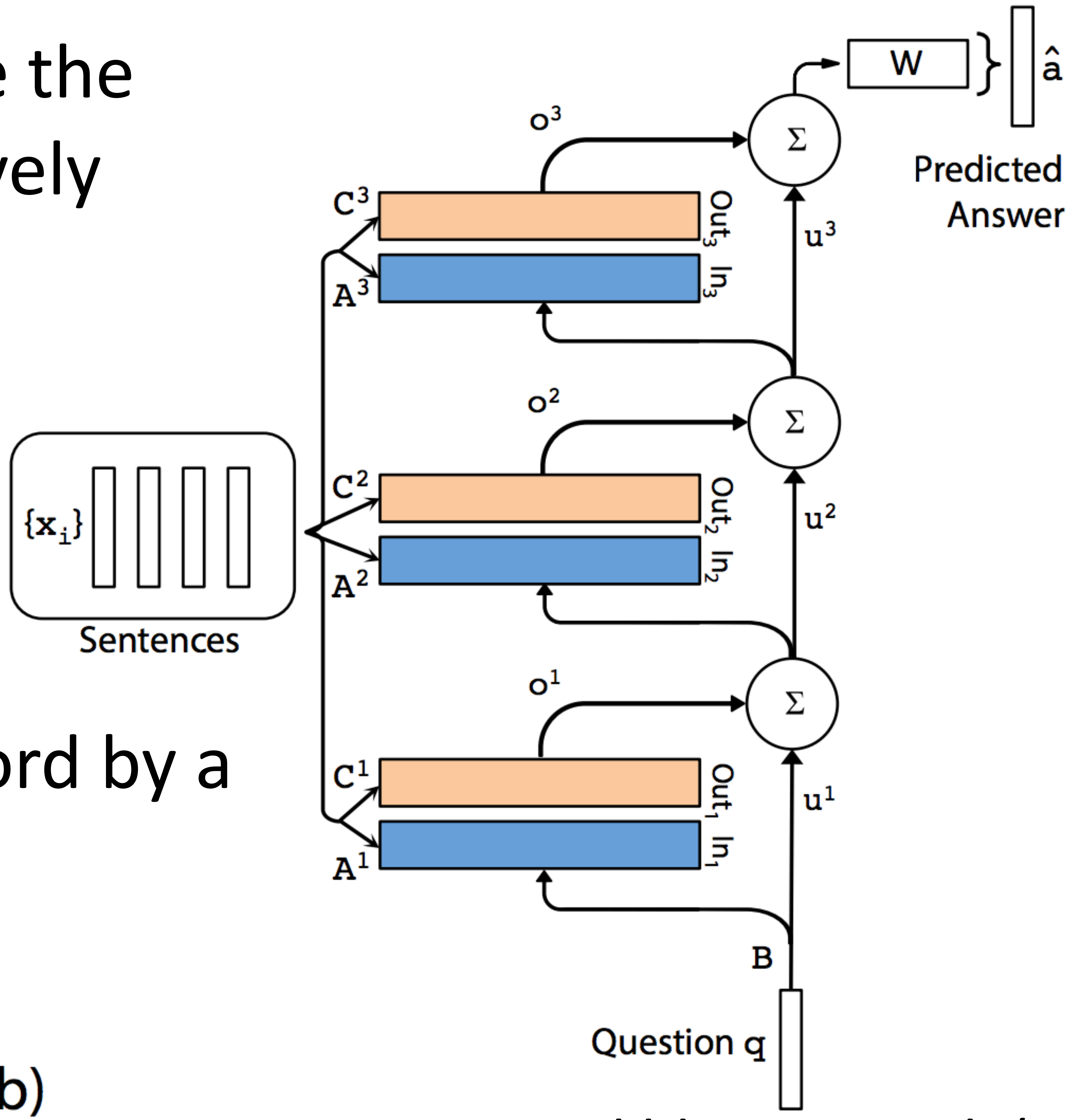

(b)

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

| 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 |

# 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

# 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**

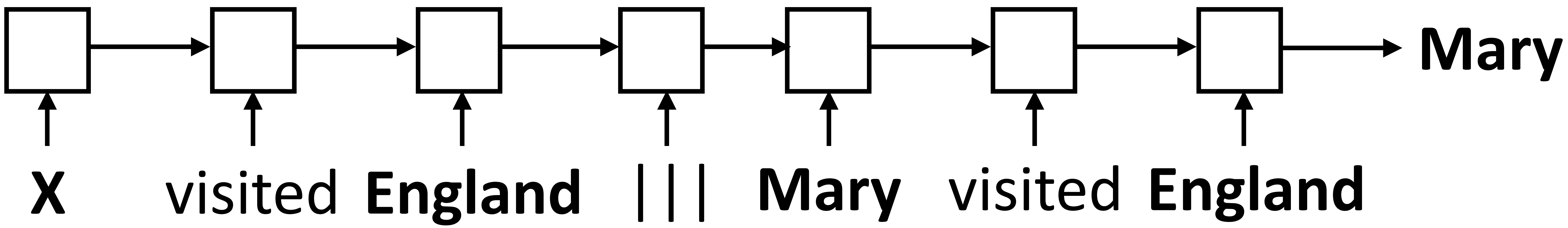
( @entity4 ) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens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

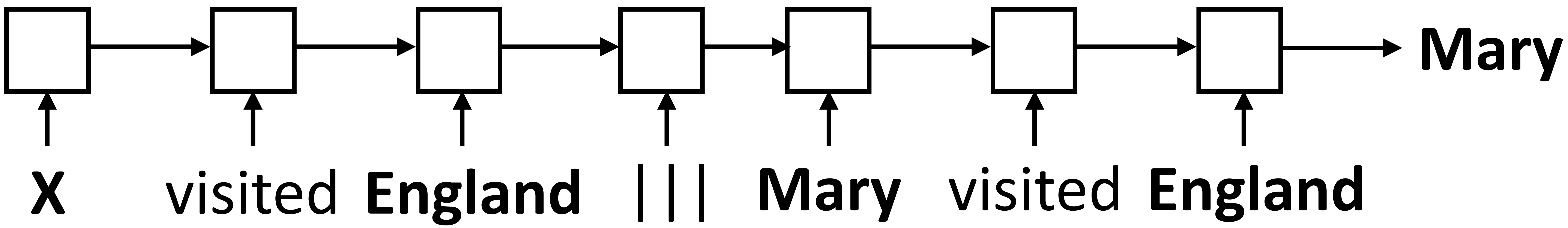
- LSTM reader: encode question, encode passage, predict entity



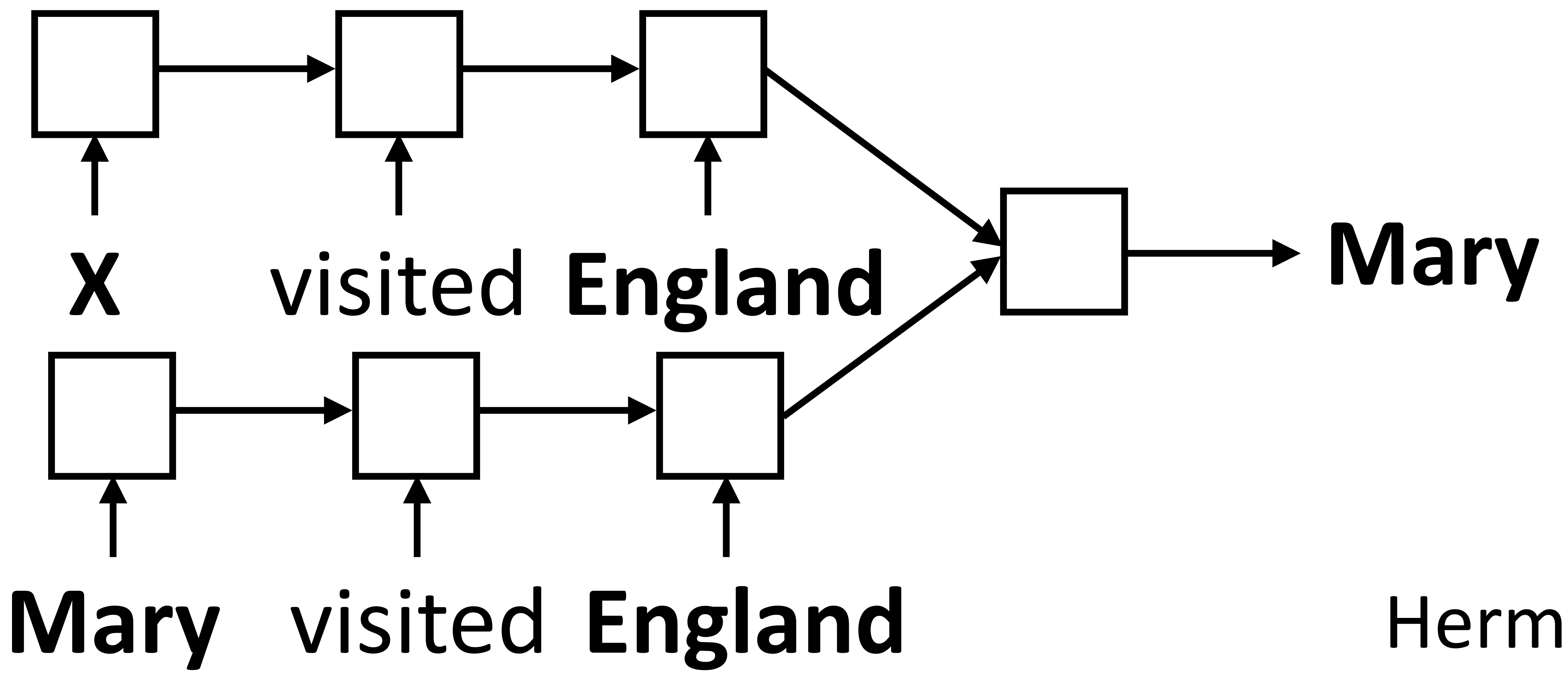

Hermann et al. (2015), Chen et al. (2016)

# CNN/Daily Mail

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




- Can also use textual entailment-like models

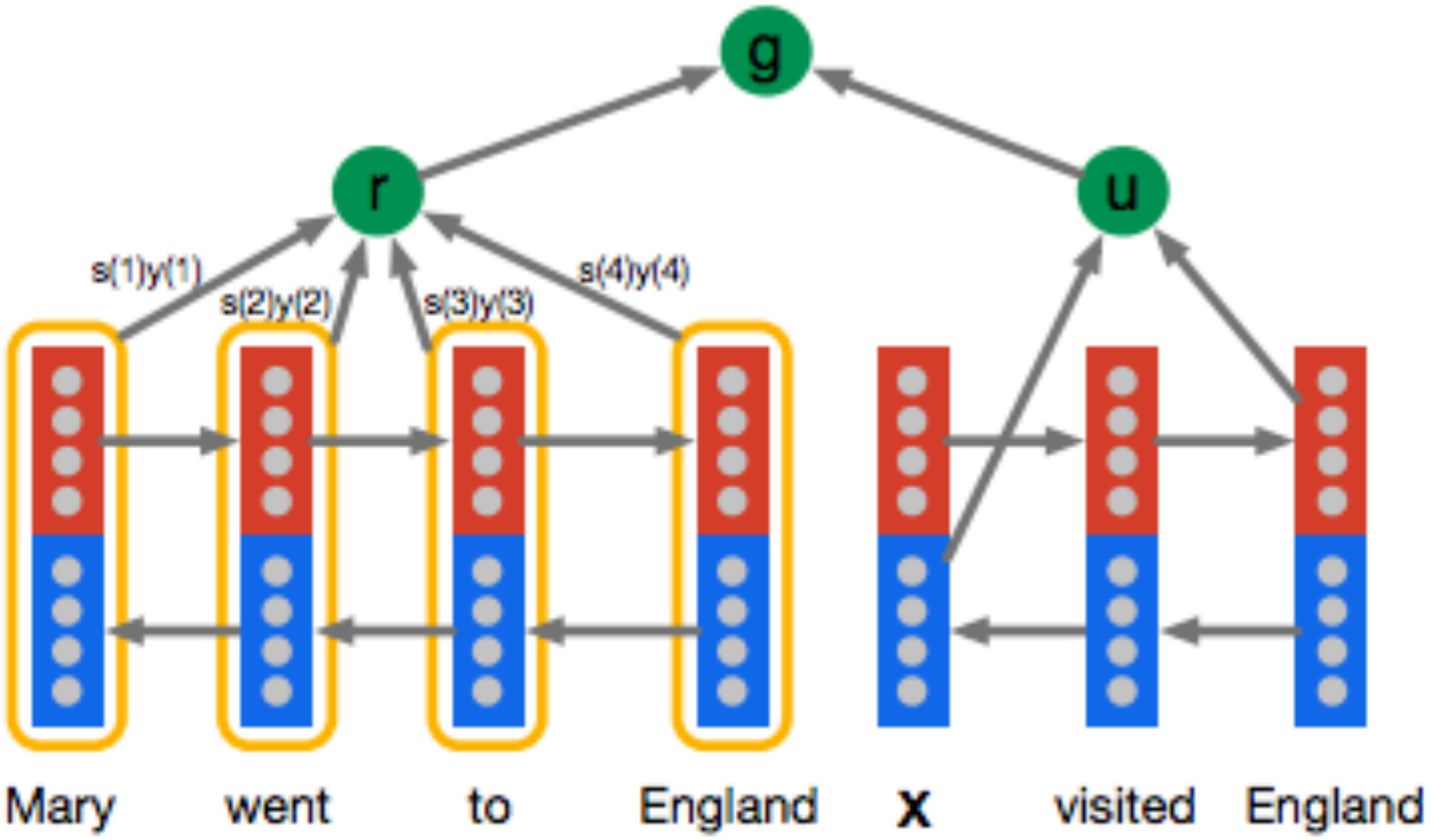



Multiclass classification problem over entities in the document

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)

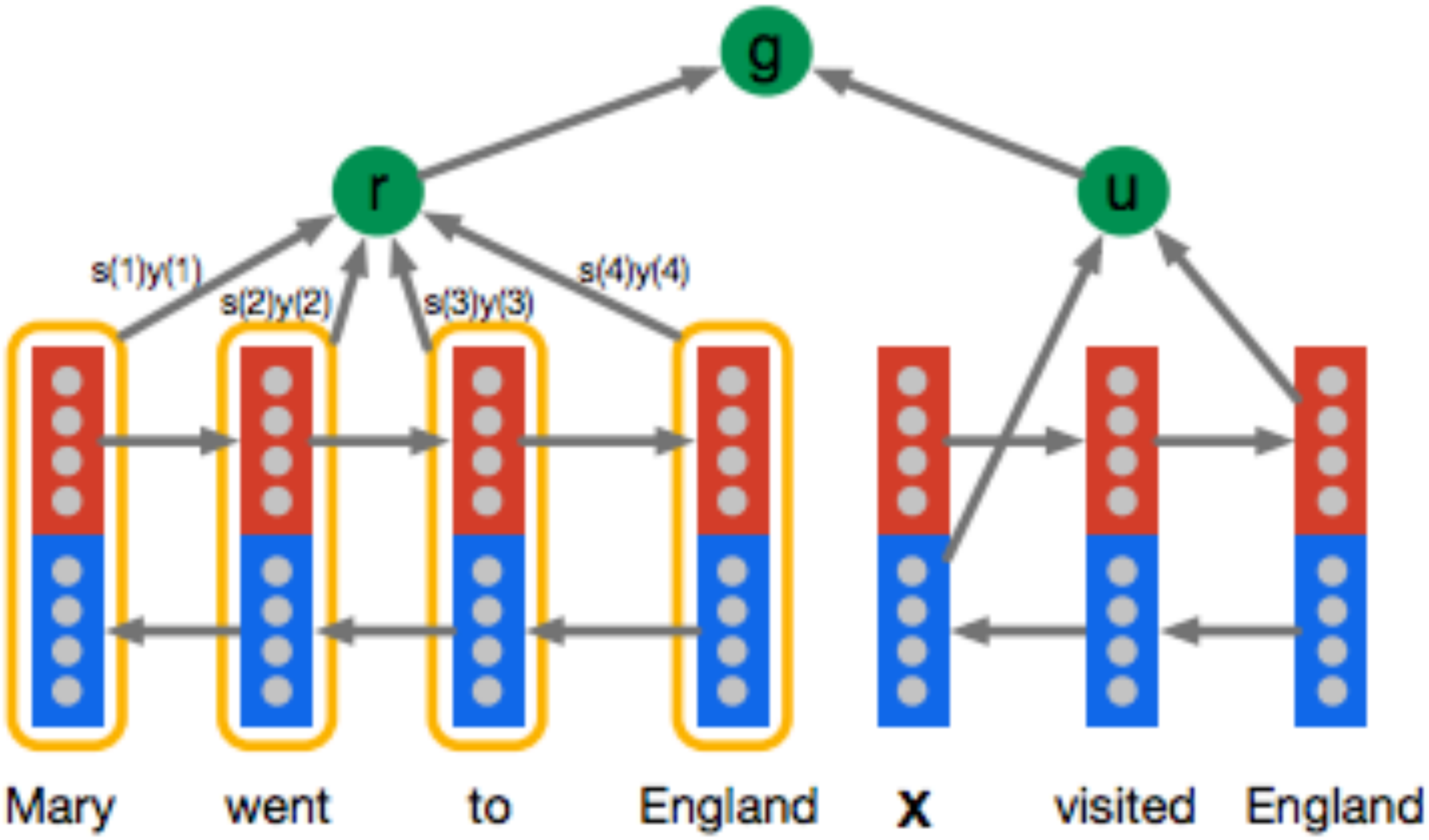



Hermann et al. (2015)

# 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

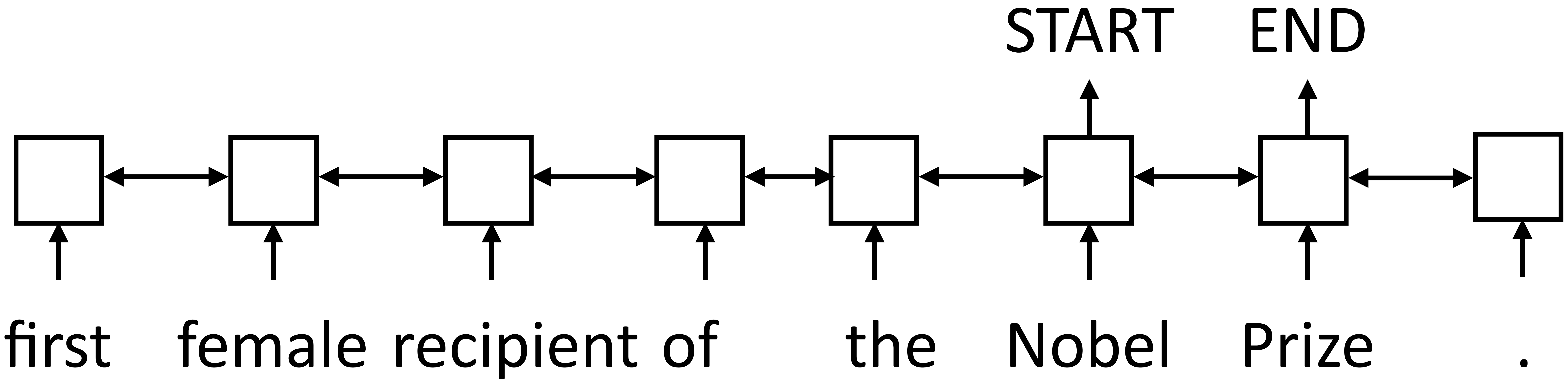
**What year was Casimir Pulaski born in Warsaw?**
*Ground Truth Answers:* 1745 | 1745 | 1745

**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

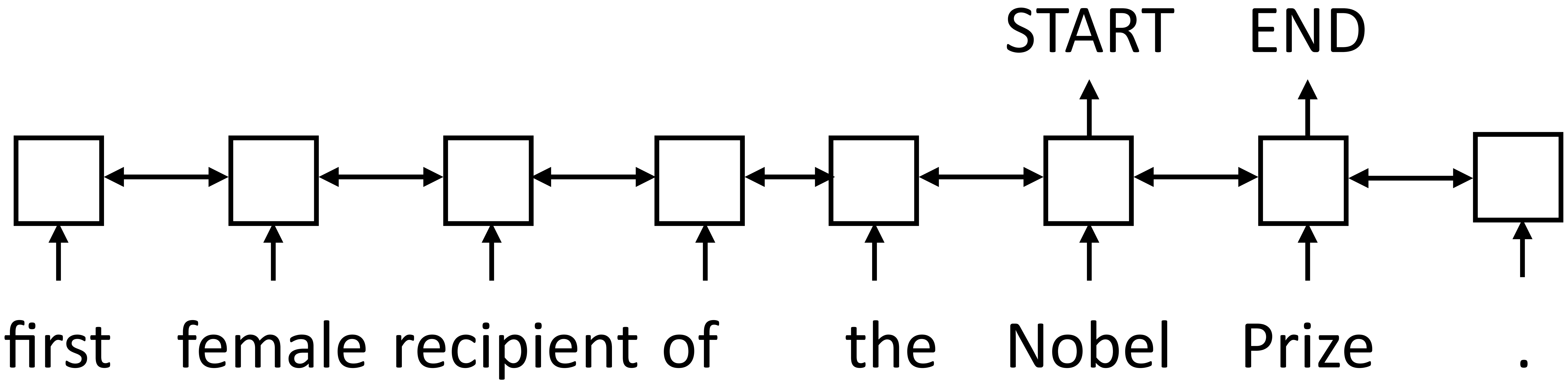
What was Marie Curie the first female recipient of?




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)

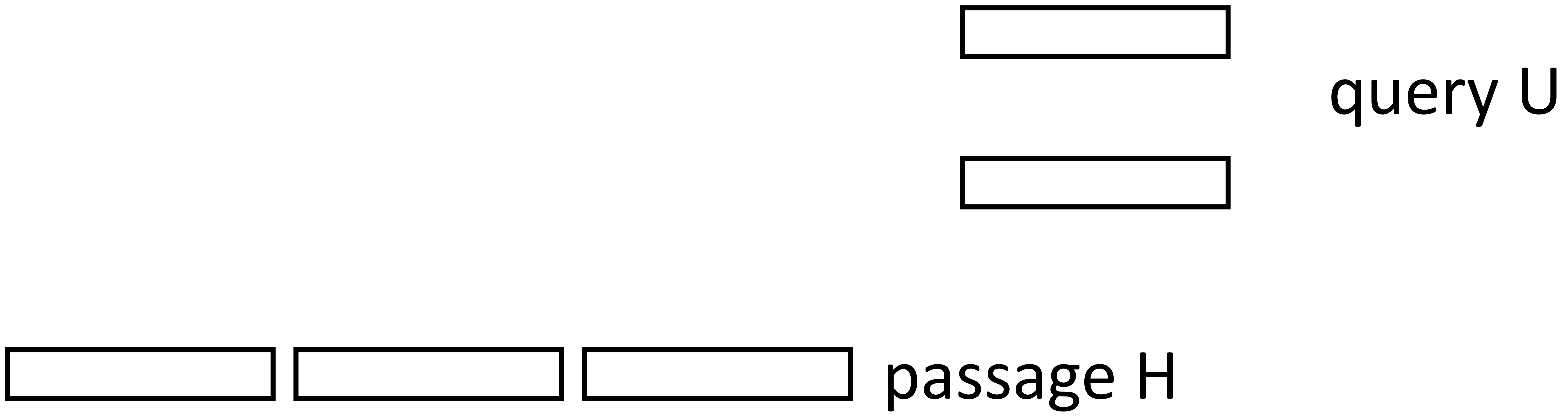
# Bidirectional Attention Flow

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

Seo et al. (2016)

# Bidirectional Attention Flow

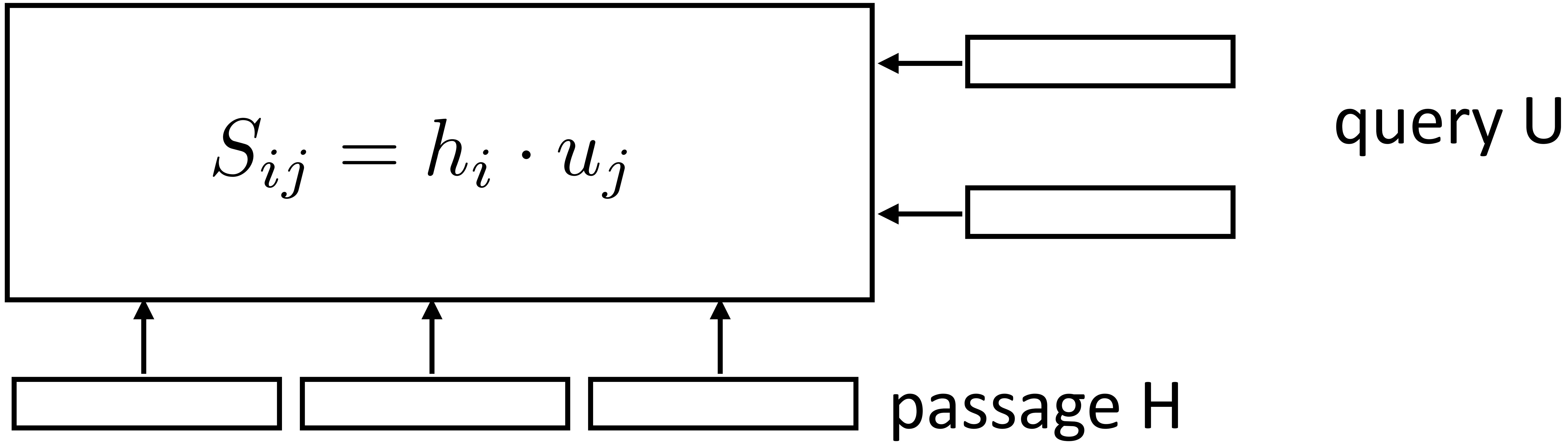
- Passage (context) and query are both encoded with BiLSTMs




Seo et al. (2016)

# Bidirectional Attention Flow

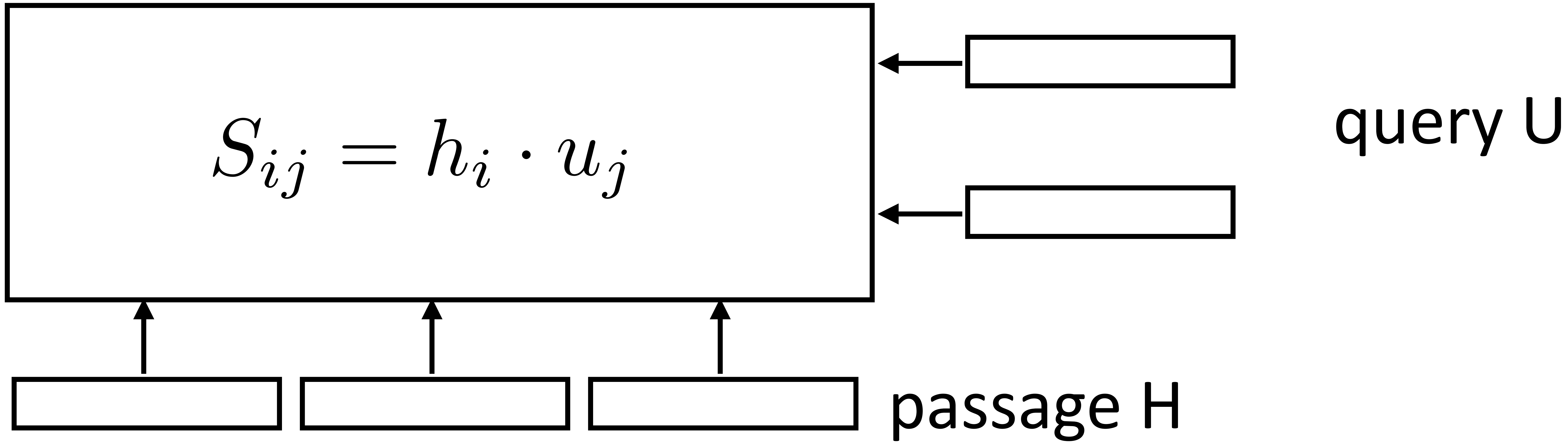
- Passage (context) and query are both encoded with BiLSTMs




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

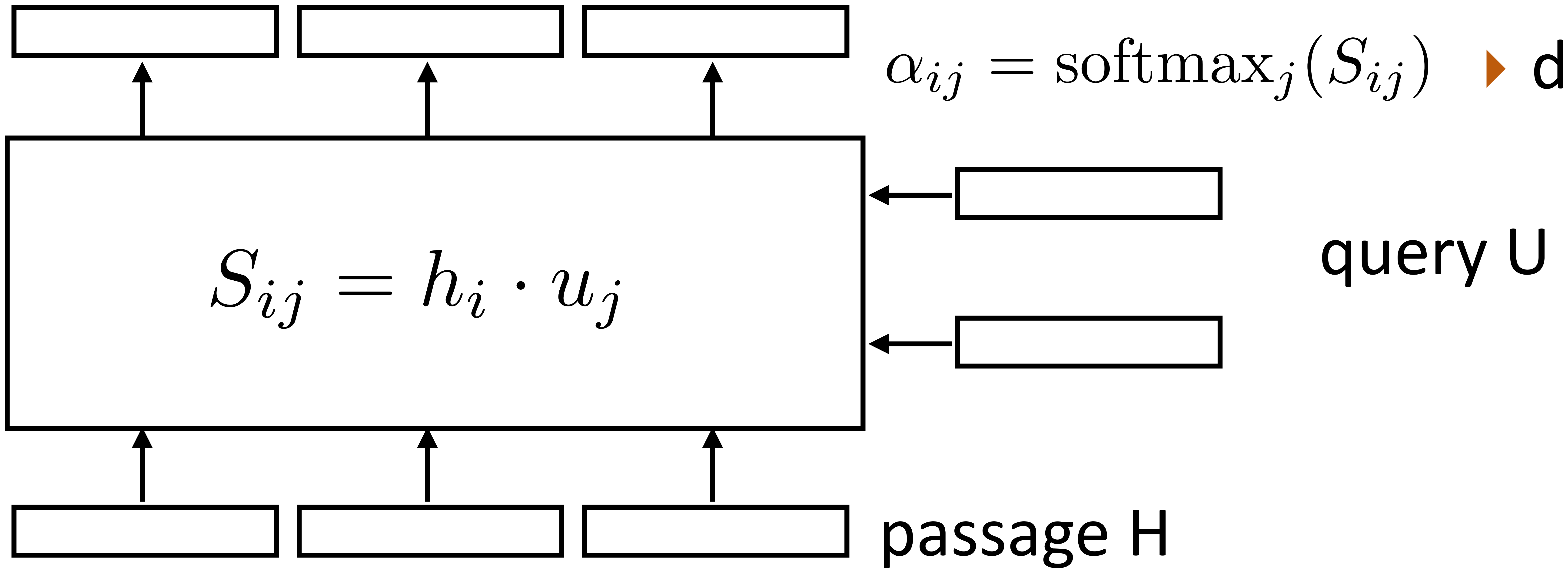
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

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

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

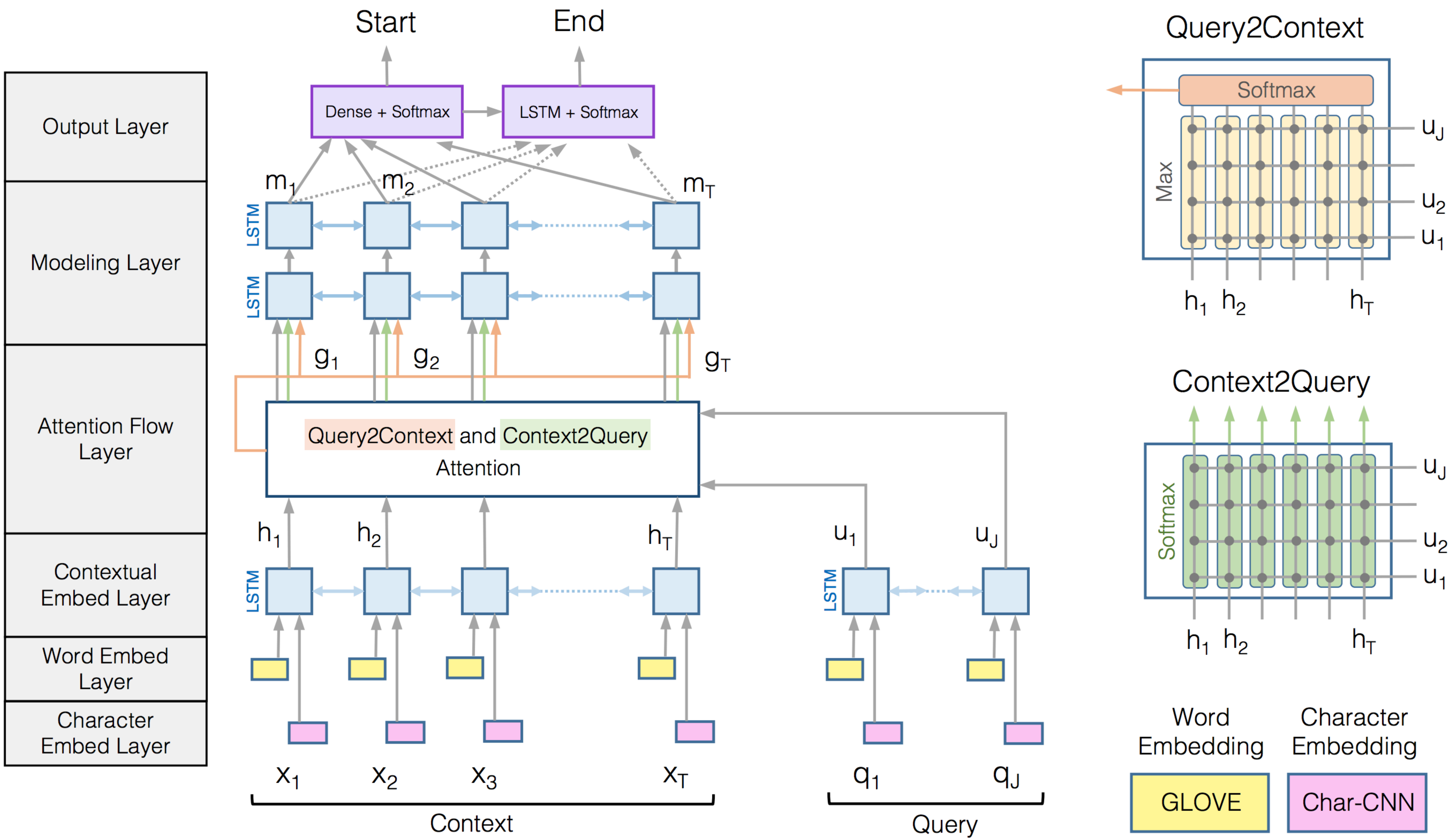
$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- dist over query



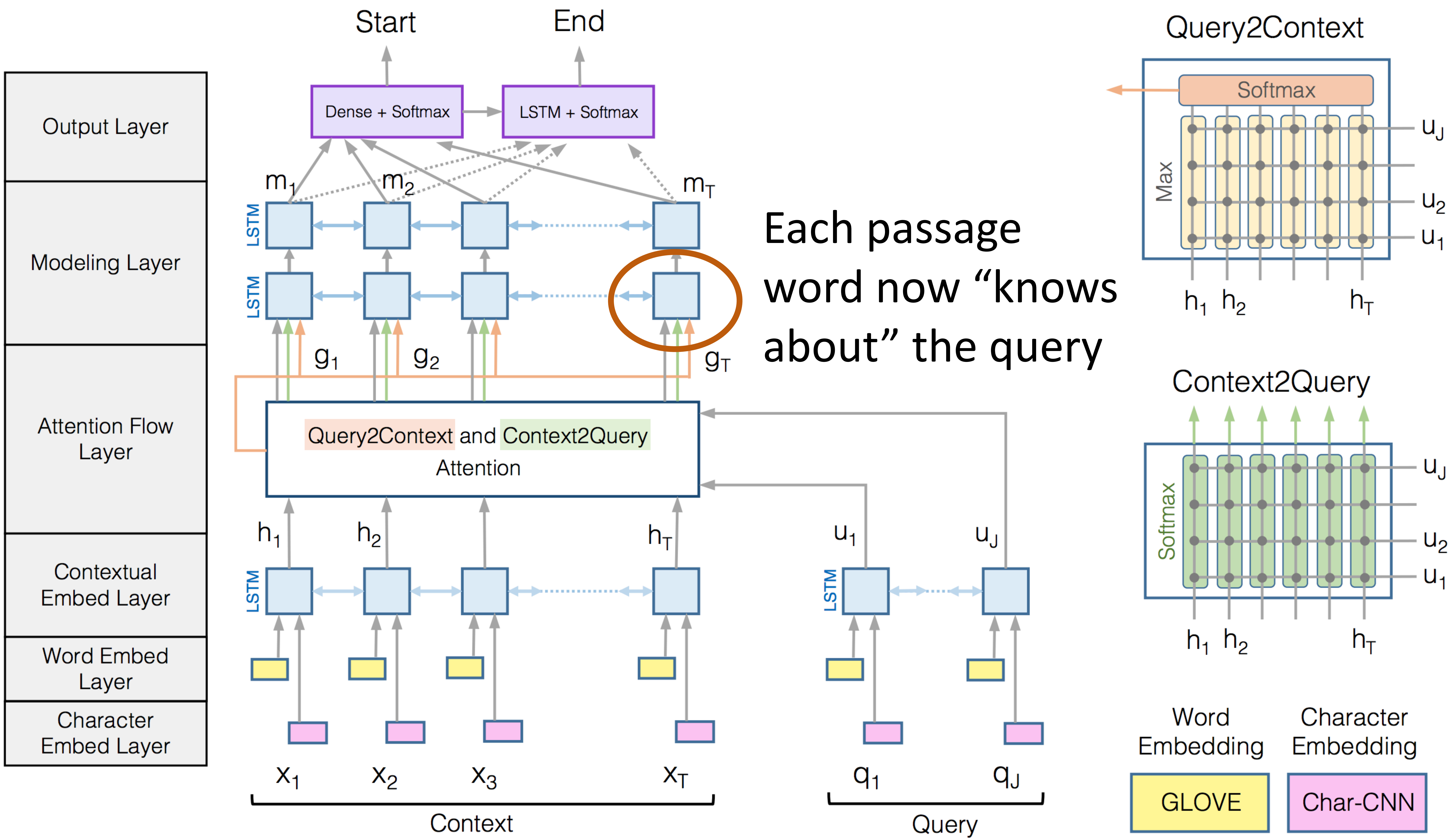

Seo et al. (2016)

# Bidirectional Attention Flow




Seo et al. (2016)

# Bidirectional Attention Flow




Each passage word now "knows about" the query

Seo et al. (2016)

# SQuAD SOTA

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

# SQuAD SOTA

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

- BiDAF: 73 EM / 81 F1

# SQuAD SOTA

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

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

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

# But how well are these doing?

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

**Article:** Super Bowl 50
**Paragraph:** "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
**Question:** "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
**Original Prediction:** John Elway
**Prediction under adversary:** Jeff Dean

Jia and Liang (2017)

# Takeaways

- Many flavors of reading comprehension tasks: cloze or actual questions, single or multi-sentence
- Memory networks let you reference input in an attention-like way, useful for generalizing language models to long-range reasoning
- Complex attention schemes can match queries against input texts and identify answers