CS11-747 Neural Networks for NLP

Machine Reading w/ Neural Networks

Graham Neubig

Carnegie Mellon University

Language Technologies Institute

Site

https://phontron.com/class/nn4nlp2017/

What is Machine Reading?

- Read a passage, try to answer questions about that passage
- Contrast to knowledge-base QA, need to synthesize the information in the passage as well
    - The passage is the KB!

Machine Reading Tasks

Machine Reading Tasks

- Multiple choice question
- Span selection
- Cloze (fill-in-the-blank) style

Multiple-choice Question Tasks

- **MCTest** (Richardson et al. 2013): 500 passages 2000 questions about simple stories

- **RACE** (Lai et al. 2017): 28,000 passages 100,000 questions from English comprehension tests

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

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.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

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

4) What did James do after he ordered the fries?
A) went to the grocery store
B) went home without paying
C) ate them
D) made up his mind to be a better turtle

Span Selection

- **SQuAD** (Rajpurkar et al. 2016): 500 passages 100,000 questions on Wikipedia text
- **TriviaQA** (Joshi et al. 2017): 95k questions, 650k evidence documents (distant supervision)

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

What causes precipitation to fall?
gravity

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

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

Cloze Questions

- **CNN/Daily Mail dataset:** Created from summaries of articles, have to guess the entity

	Original Version	Anonymised Version
Context		
	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " …
Query		
	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer		
	Oisin Tymon	*ent193*

- Entities anonymized to prevent co-occurance clues

What is Necessary for Machine Reading?

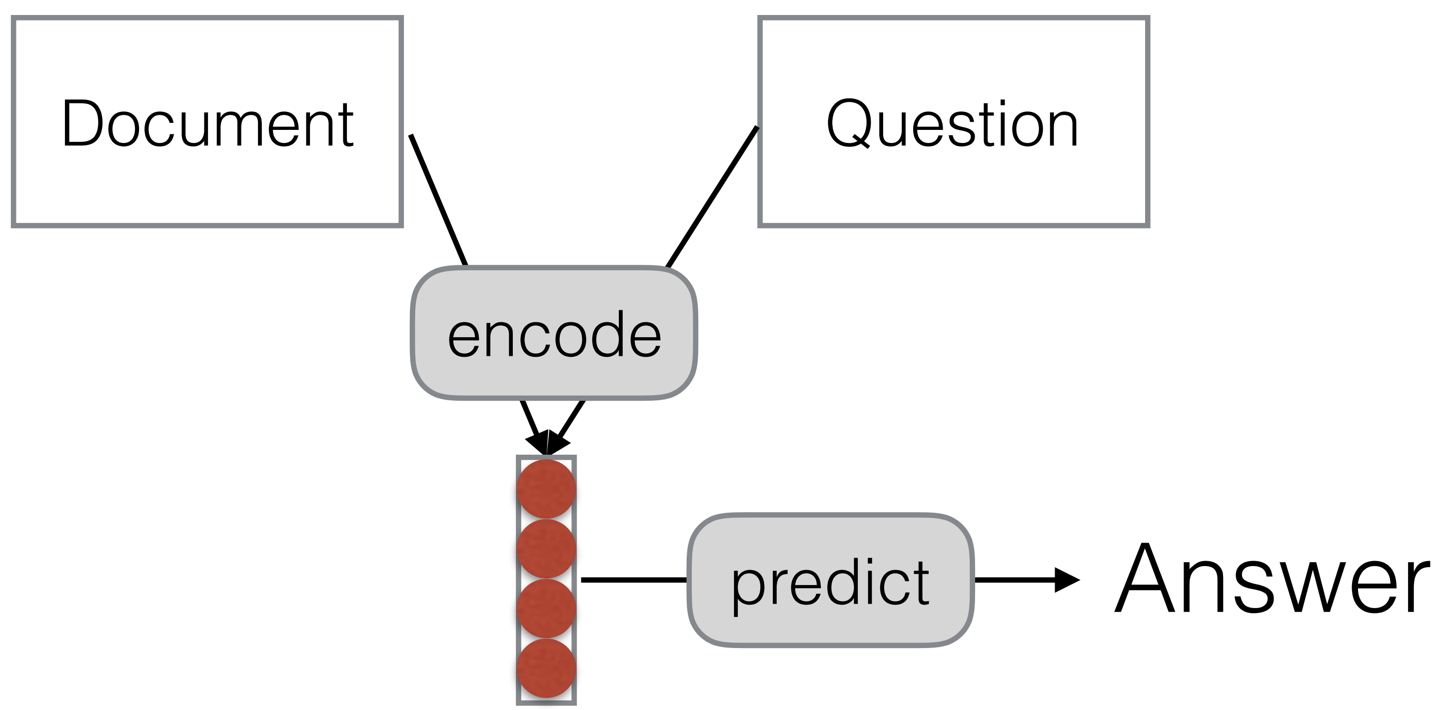
- We must take a large amount of information and extract only the salient parts
  → **Attention**

- We must perform some sort of reasoning about the information that we’ve extracted
  → **Multi-step Reasoning**

Attention Models for Machine Reading

A Basic Model for Document Attention

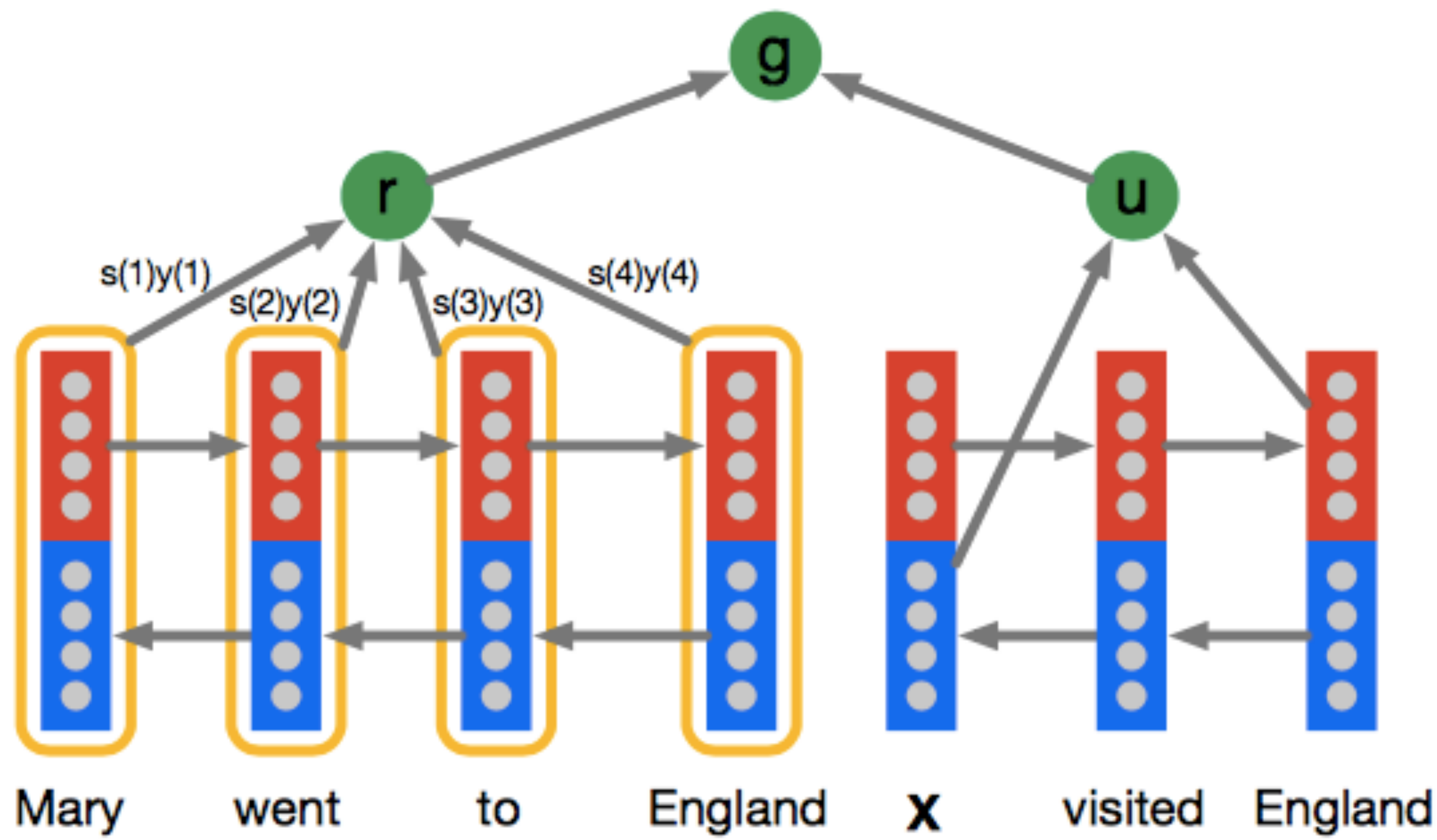
- Encode the document and the question, and generate an answer (e.g. a sentence or single word)



- Problem: encoding whole documents with high accuracy and coverage is hard!

A First Try: Attentive Reader
(Hermann et al. 2015)

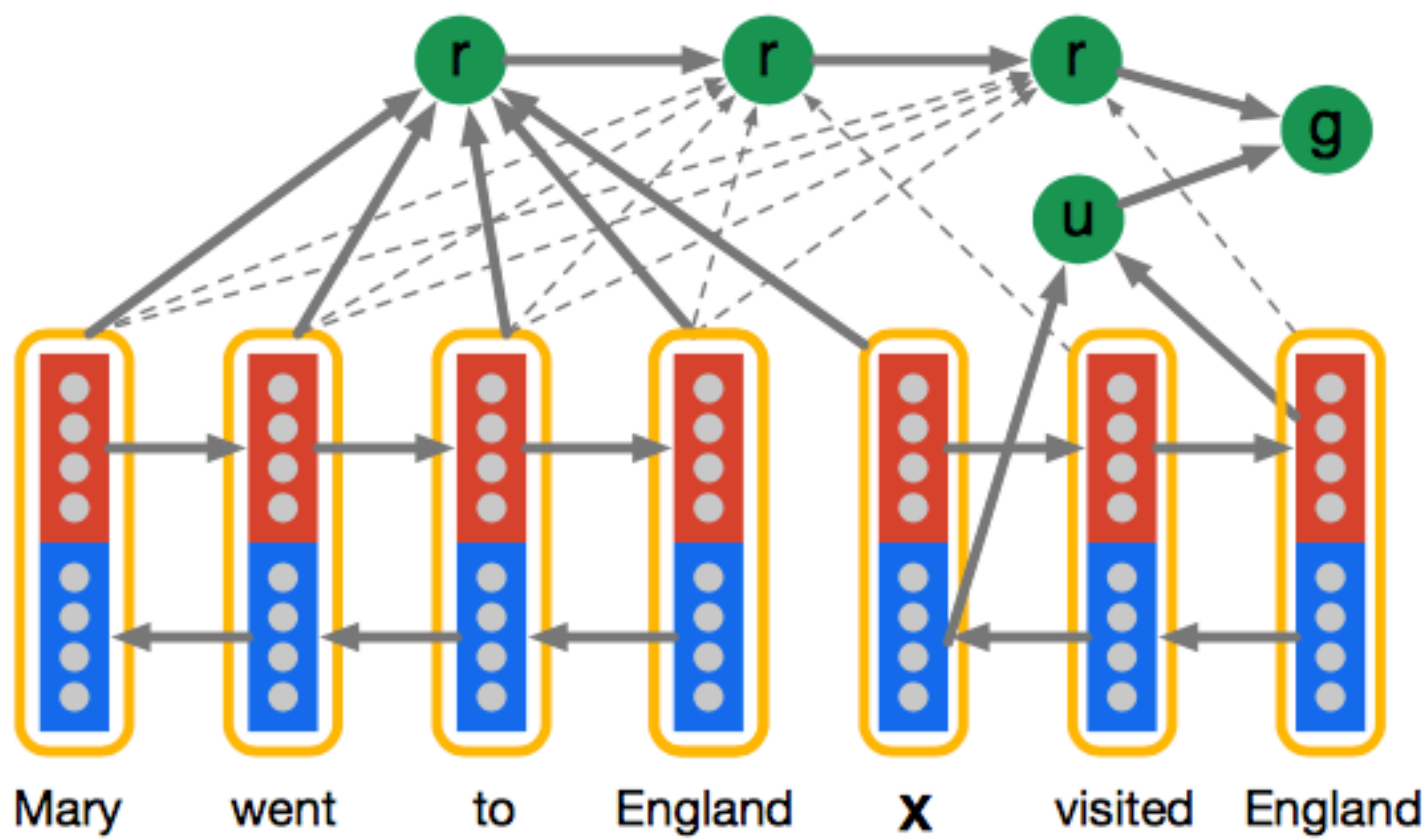
- Read the query (u) first, then attend to the values in the context vector



- Allows the model to focus on relevant information, but query is not considered during encoding

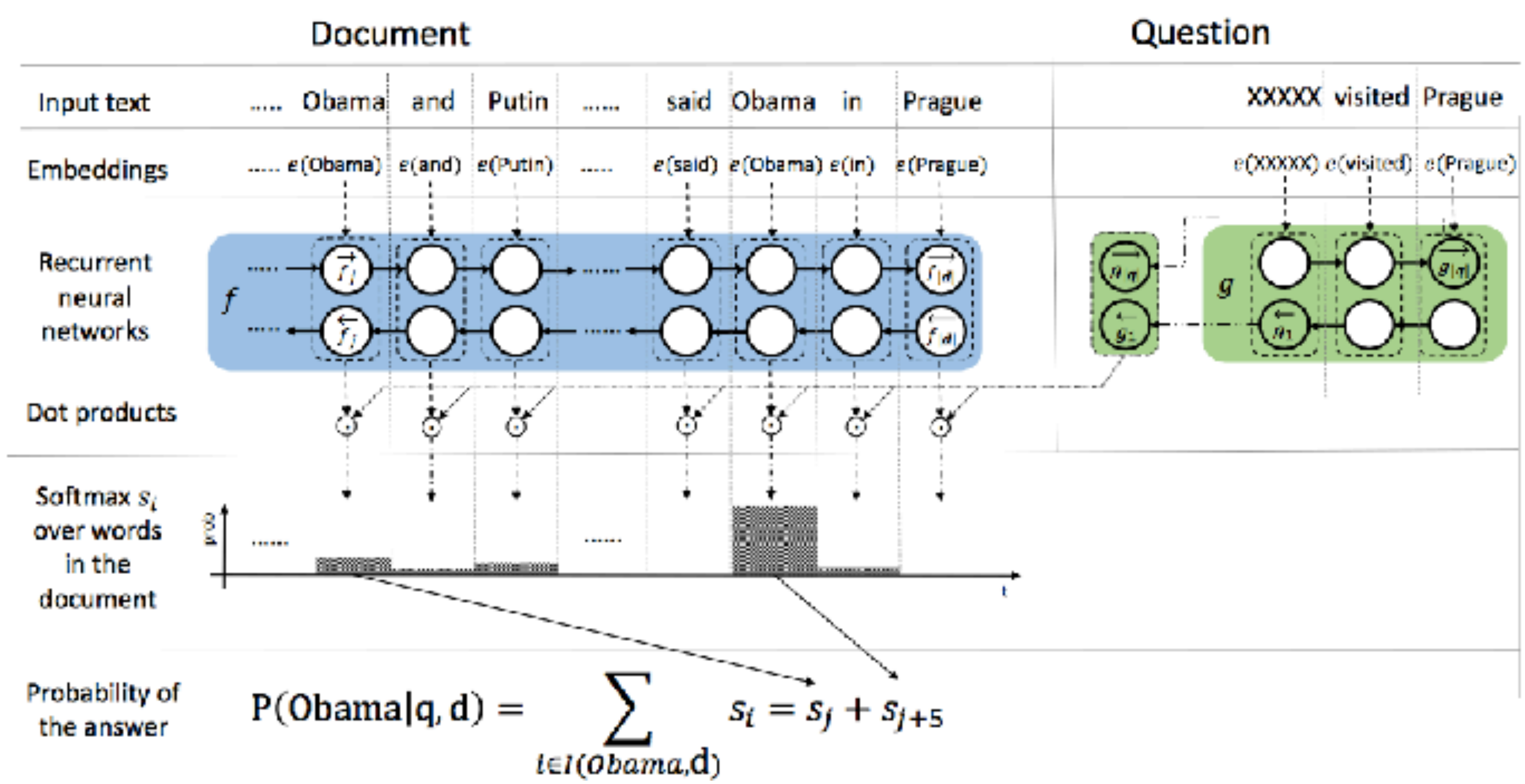
Impatient Reader
(Hermann et al. 2015)

- Re-read the document every time you get a new query token and update understanding



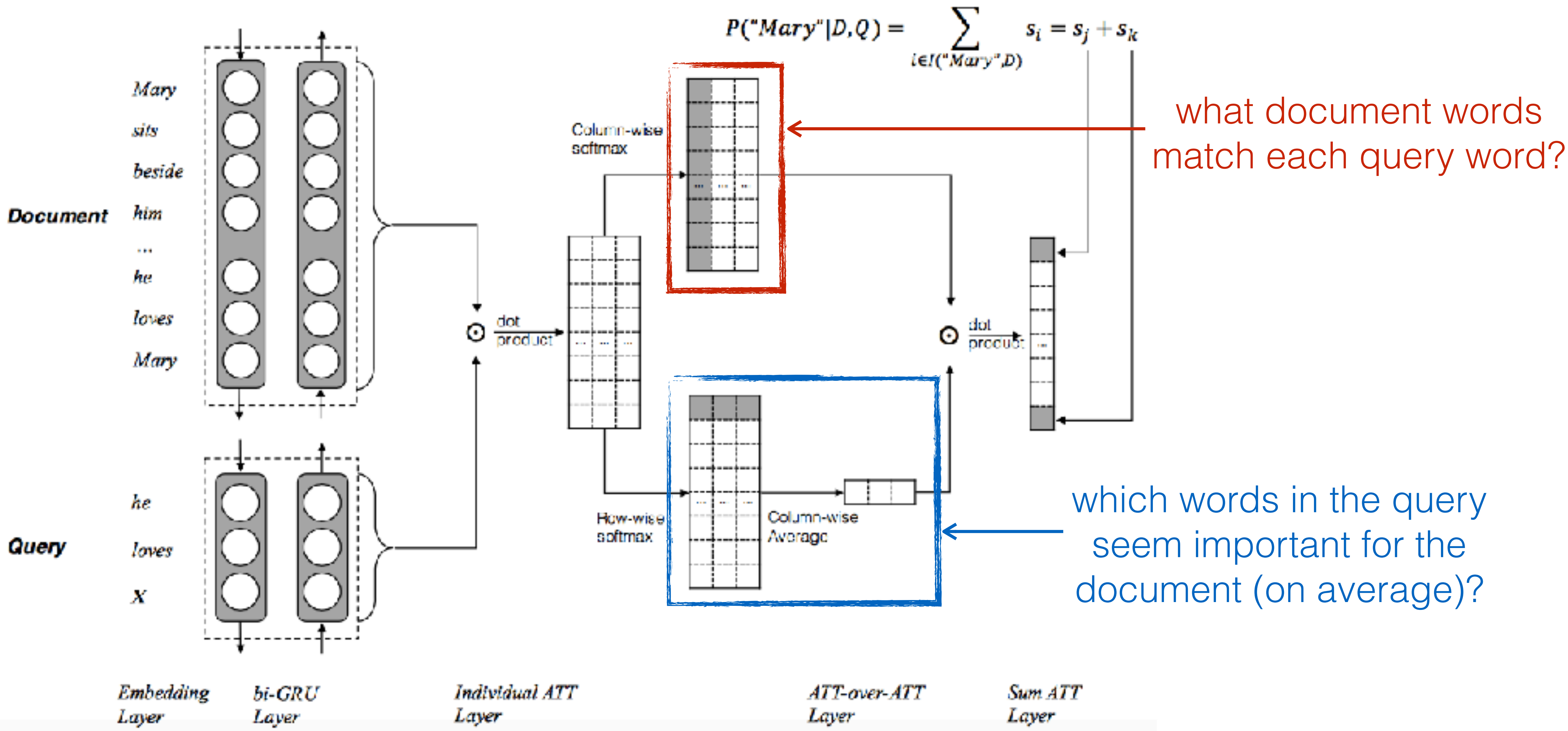
Attention Sum Reader
(Kadlec et al. 2016)

- Instead of attending to get representation, attend to each entity in the source document
- The score of the entity is the sum of the attention scores over all mentions of the entity



Attention-over-attention
(Cui et al. 2017)

- Idea: we want to know the document words that match best with the *most important words* in the query



Choosing Answer Spans

Word Classification vs. Span Classification

- In span-based models, we need to choose a multi-word span

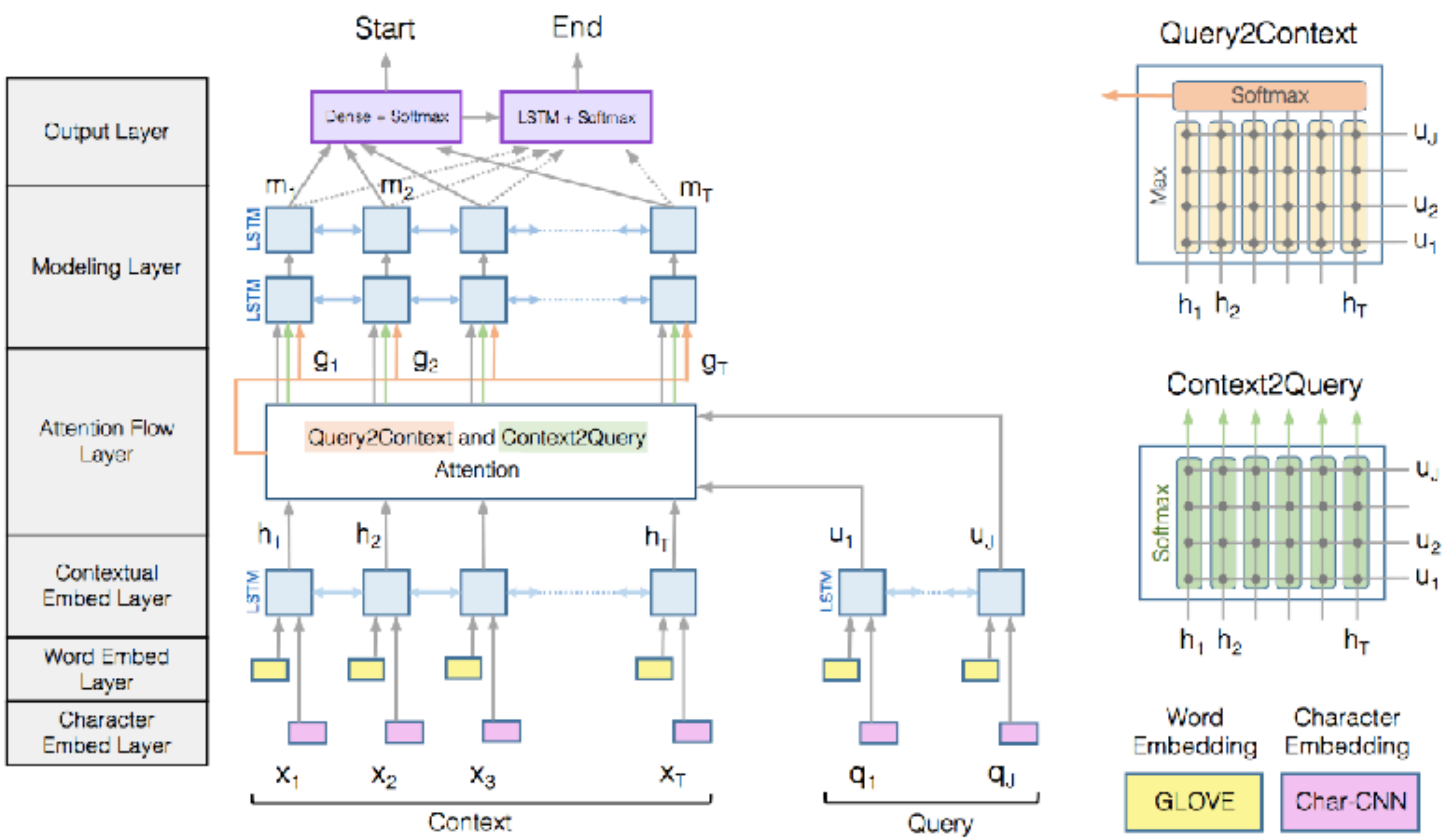
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or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

- In contrast:
    - Previous single-word machine reading models choose a single word or entity
    - Other models such as NER choose multiple spans

Bidirectional Attention Flow

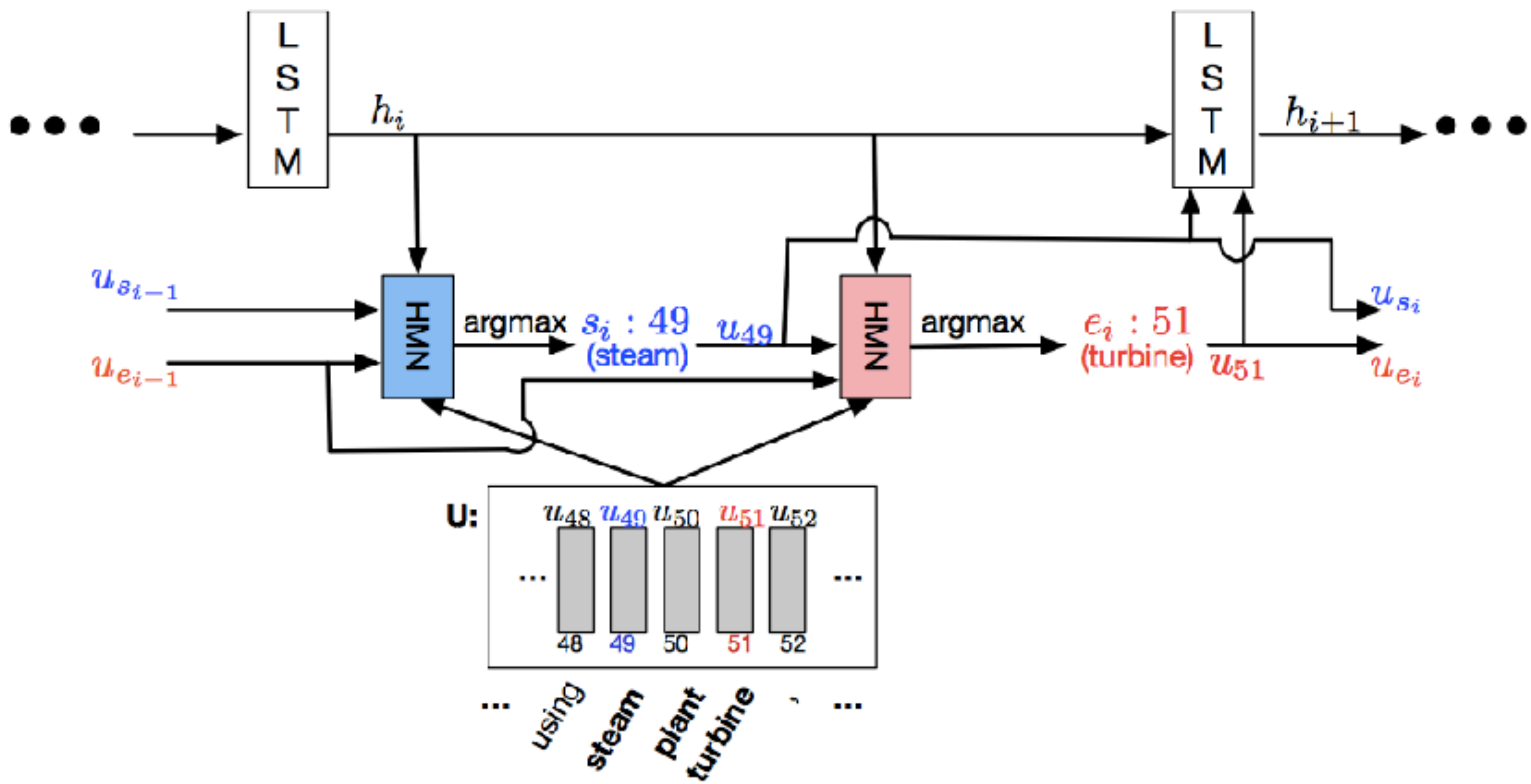
(Seo et al. 2017)

- Calculate doc2ctxt, ctxt2doc attention
- Both representations concatenated to word representations themselves in the document



Dynamic Span Decoder
(Xiong et al. 2017)

- Iteratively refine the left and right boundaries



Multi-step Reasoning

Multi-step Reasoning

- It might become clear that more information is necessary post-facto

John went to the hallway

John put down the football

Q: Where is the football?

Step 1: Attend to football   Step 2: Attend to John

Example: Kumar et al. 2016

Memory Networks
(Weston et al. 2014)

- A general formulation of models that access external memory through attention and specific instantiation for document-level QA
- In specific QA model, first do arg-max attention:

$$o_1 = O_1(x, \mathbf{m}) = \underset{i=1,\ldots,N}{\arg\max}\ s_O(x, \mathbf{m}_i)$$

- But with additional argmax step to get a second element from memory, conditioned on first

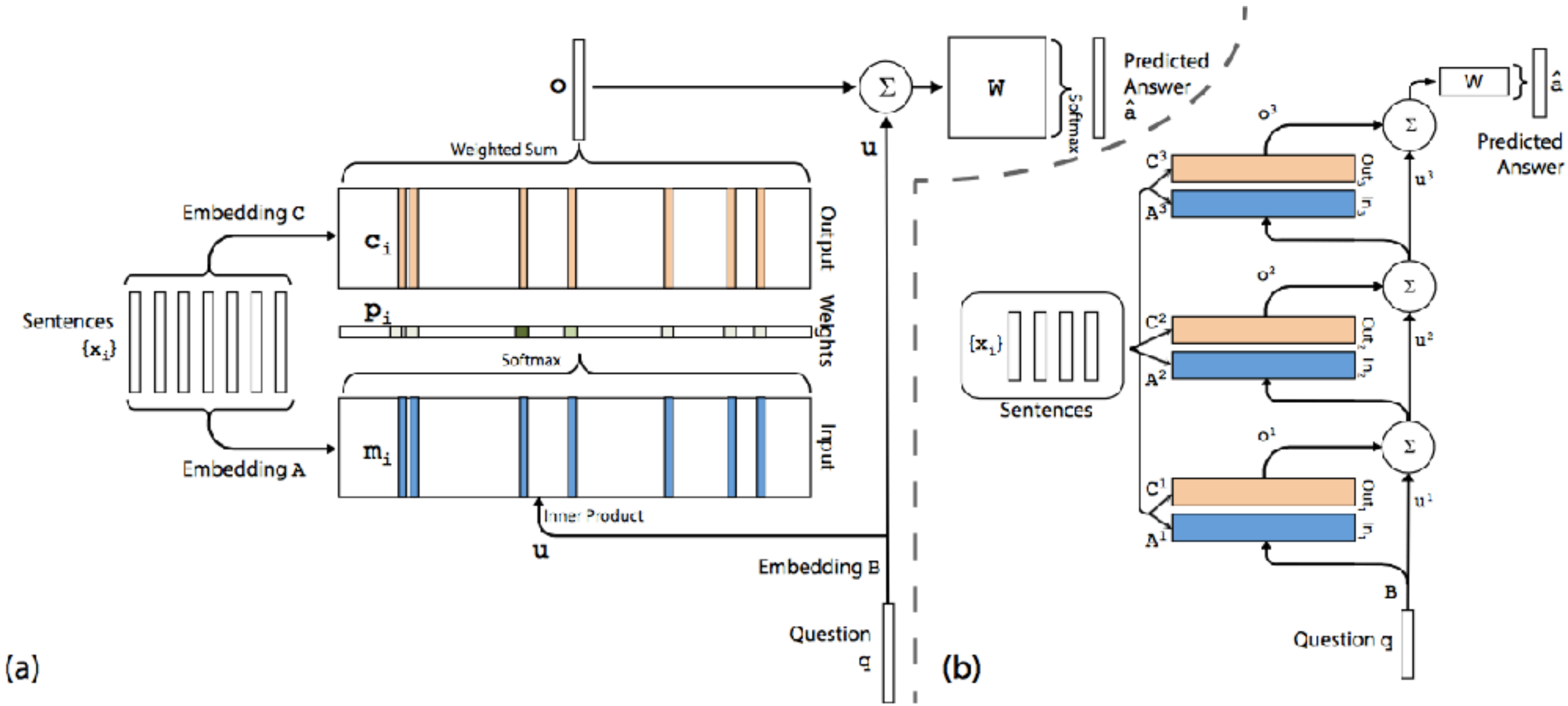
$$o_2 = O_2(x, \mathbf{m}) = \underset{i=1,\ldots,N}{\arg\max}\ s_O([x, \mathbf{m}_{o_1}], \mathbf{m}_i)$$

- Use both to get the answer

$$r = \mathrm{argmax}_{w \in W}\ s_R([x, \mathbf{m}_{o_1}, \mathbf{m}_{o_2}], w)$$

Softened, and Multi-layer Memory Networks (Sukhbaatar et al. 2015)

- Use standard softmax attention, and multiple layers

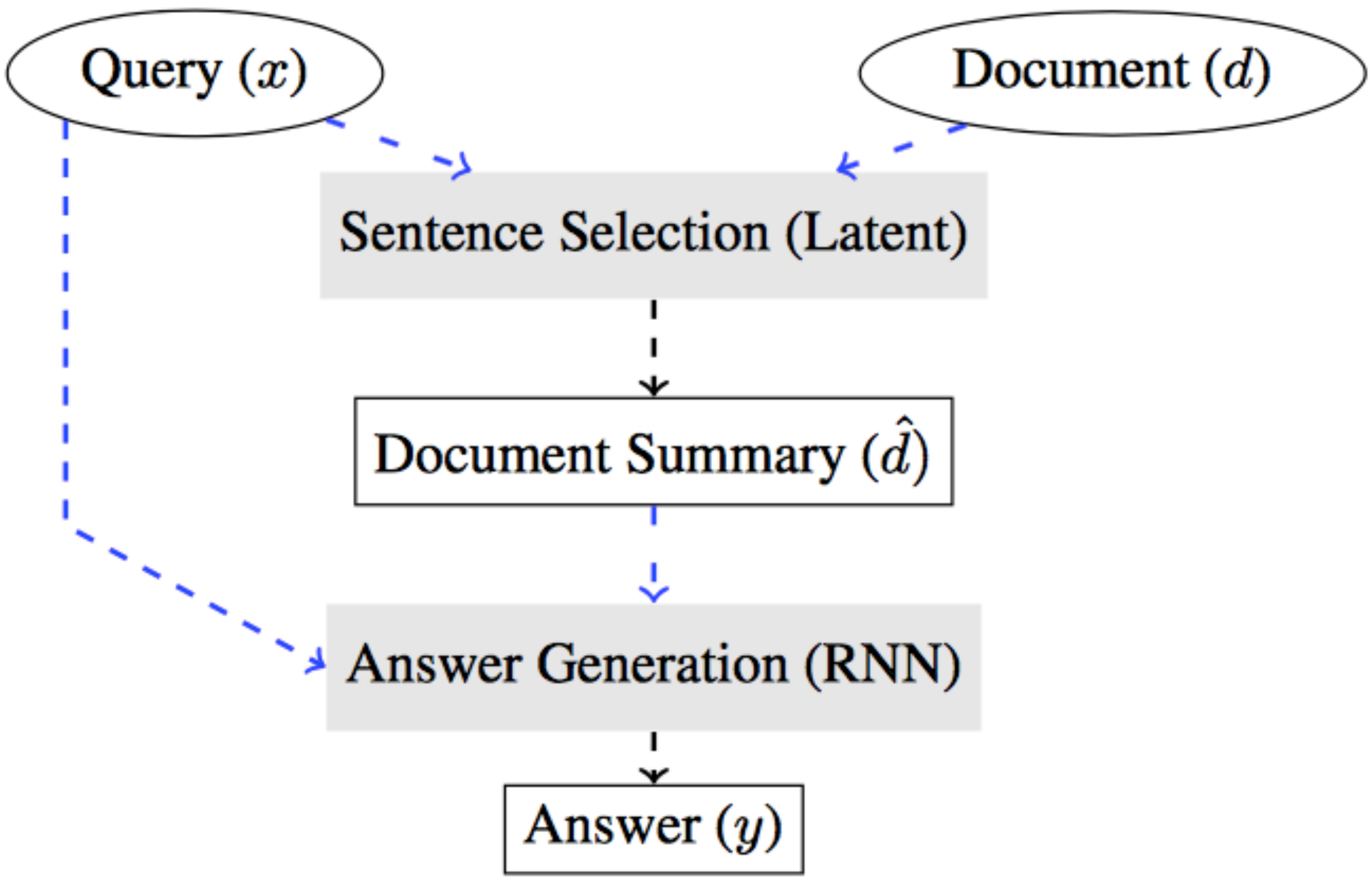


When to Stop Reasoning?

- A fixed number of sequences (e.g. Weston et al. 2014)
- When we attend to a “stop reasoning” symbol (e.g. Kumar et al. 2016)
- Have an explicit “stop reasoning” predictor (e.g. Shen et al. 2017)

Coarse-to-fine Question Answering (Choi et al. 2017)

- First, decide which sentence to cover, then reason



- This is also a variety of multi-hop reasoning

A Caveat about Data Sets

All Datasets Have Their Biases

- No matter the task, data bias matters
  - Domain bias
  - Simplifications
- In particular, for reading comprehension, real, large-scale (copyright-free) datasets are hard to come by
- Datasets created from weak supervision have not been vetted

A Case Study: bAbI
(Weston et al. 2014)

- Automatically generate synthetic text aimed at evaluating whether a model can learn certain characteristics of language

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 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

- Problem: papers evaluate *only* on this extremely simplified dataset, then claim about ability to learn language
- **Neubig's Law:** If you can write a rule-based system to solve a dataset w/ 100% accuracy in short order, it's probably less interesting
- **Extra credit:** reference Pereira and Shieber (2002) and write a Prolog program to solve bAbI

An Examination of CNN/ Daily Mail (Chen et al. 2015)

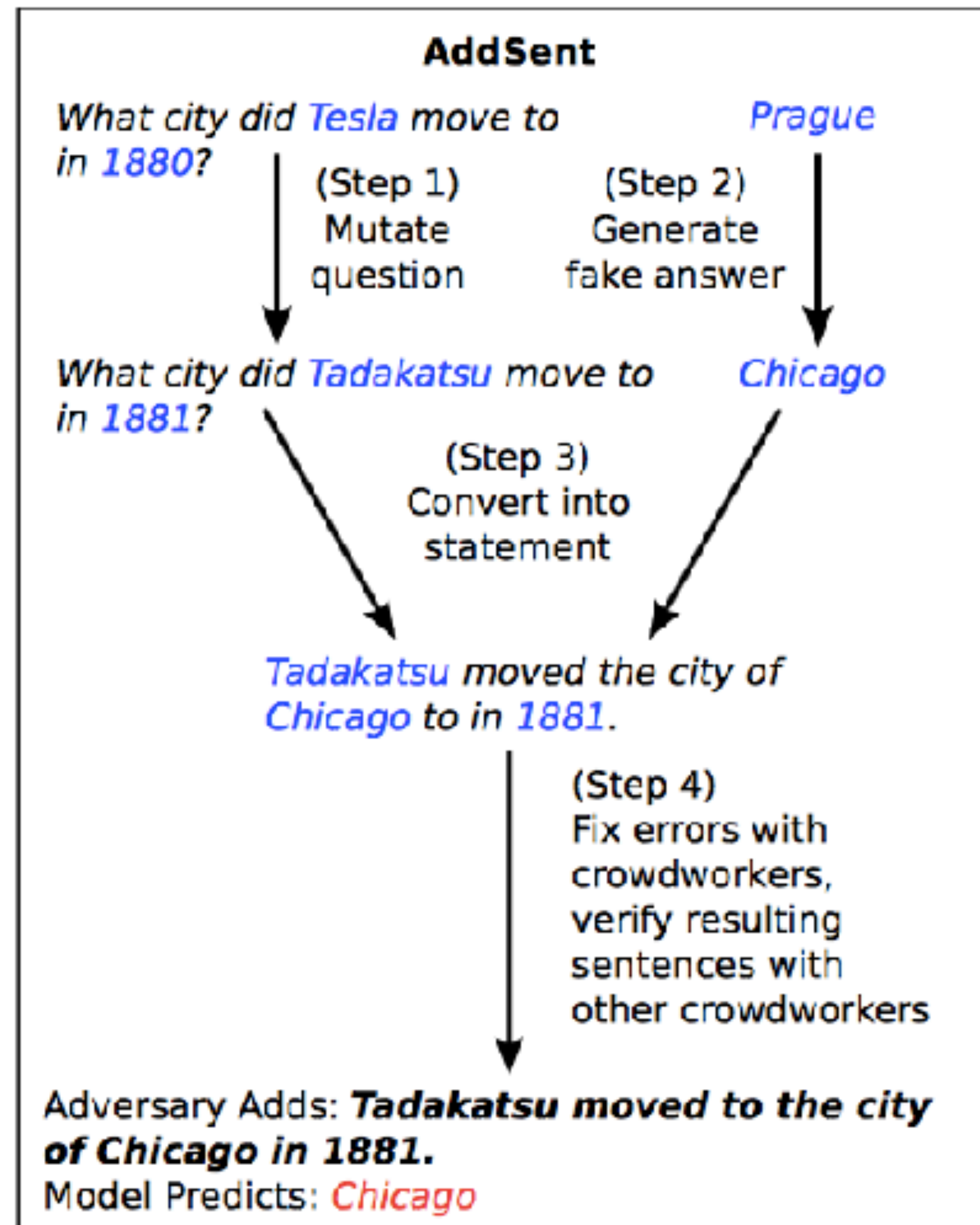
- Even synthetically created real datasets have problems!
- An analysis of CNN/Daily Mail revealed very few sentences required multi-sentence reasoning, and many were too difficult due to anonymization or wrong preprocessing

No.	Category	(%)
1	Exact match	13
2	Paraphrasing	41
3	Partial clue	19
4	Multiple sentences	2
5	Coreference errors	8
6	Ambiguous / hard	17

Adversarial Examples in Machine Reading (Jia and Liang 2017)

- Add a sentence or word string specifically designed to distract the model

- Drops accuracy of state-of-the-art models from 81 to 46



Questions?