CS769 Advanced NLP

Text-based Question Answering

Junjie Hu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Slides adapted from Graham
https://junjiehu.github.io/cs769-spring23/

Goal for Today

- Text-based QA Tasks
- Three types of QA Systems
    1. Attention Models for Reading Comprehension
    2. Multi-hop QA
    3. Retrieval-based (Open) QA
- Caveats about Dataset Curations

What is Text-based QA?

- Read a passage, try to answer questions about that passage (also called "machine reading")
- Contrast to knowledge-base QA (later class), need to match to unstructured data source.

Who was the oldest US president to take office?

Text-based QA

Biden and his running mate Kamala Harris defeated incumbent president Donald Trump and vice president Mike Pence in the 2020 presidential election. He is the oldest president and the first to have a female vice president. Biden proposed, lobbied for and signed into law the

Knowledge-base QA

#	President	Born	Age at start of presidency	Age at end of presidency	Post-presidency timespan
46	Joe Biden	Nov 20, 1942	78 years, 61 days Jan 20, 2021	*(incumbent)*	*(incumbent)*
45	Donald Trump	Jun 14, 1946	70 years, 220 days Jan 20, 2017	74 years, 220 days Jan 20, 2021	*68 days*
40	Ronald Reagan	Feb 6, 1911	69 years, 349 days Jan 20, 1981	77 years, 349 days Jan 20, 1989	15 years, 137 days

Text-based QA Tasks

Machine Reading Question Answering Formats

- Multiple choice question
- Span selection
- Cloze (fill-in-the-blank) style
- Generative QA
- *(Information extraction)*

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

Generative QA

- Generate an output, not constrained
- **NarrativeQA:** Generate an answer based on a story (Kočiský et al. 2018)
- Similarities to *query-based summarization*
- Evaluation difficult -- NLG metrics (BLEU/ROUGE) or retrieval metrics (MRR)

Title: Ghostbusters II

Question: How is Oscar related to Dana?

Answer: her son

Summary snippet: …Peter's former girlfriend Dana Barrett has had a son, Oscar…

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.

What is Necessary for Text-based QA?

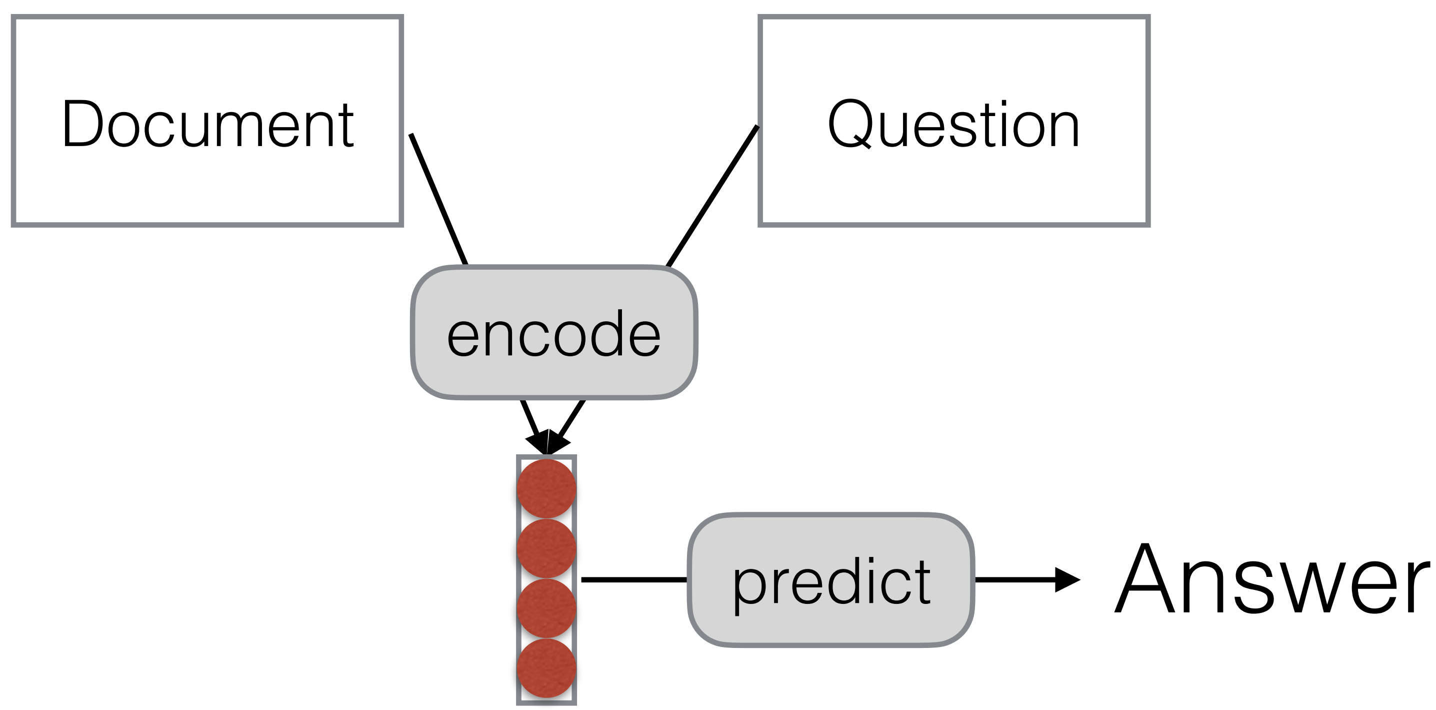
- We must take a large amount of information and extract only the salient parts
→ **Attention models**
→ **Retrieval models**

- We must perform some sort of reasoning about the information that we've extracted
→ **Multi-hop 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)

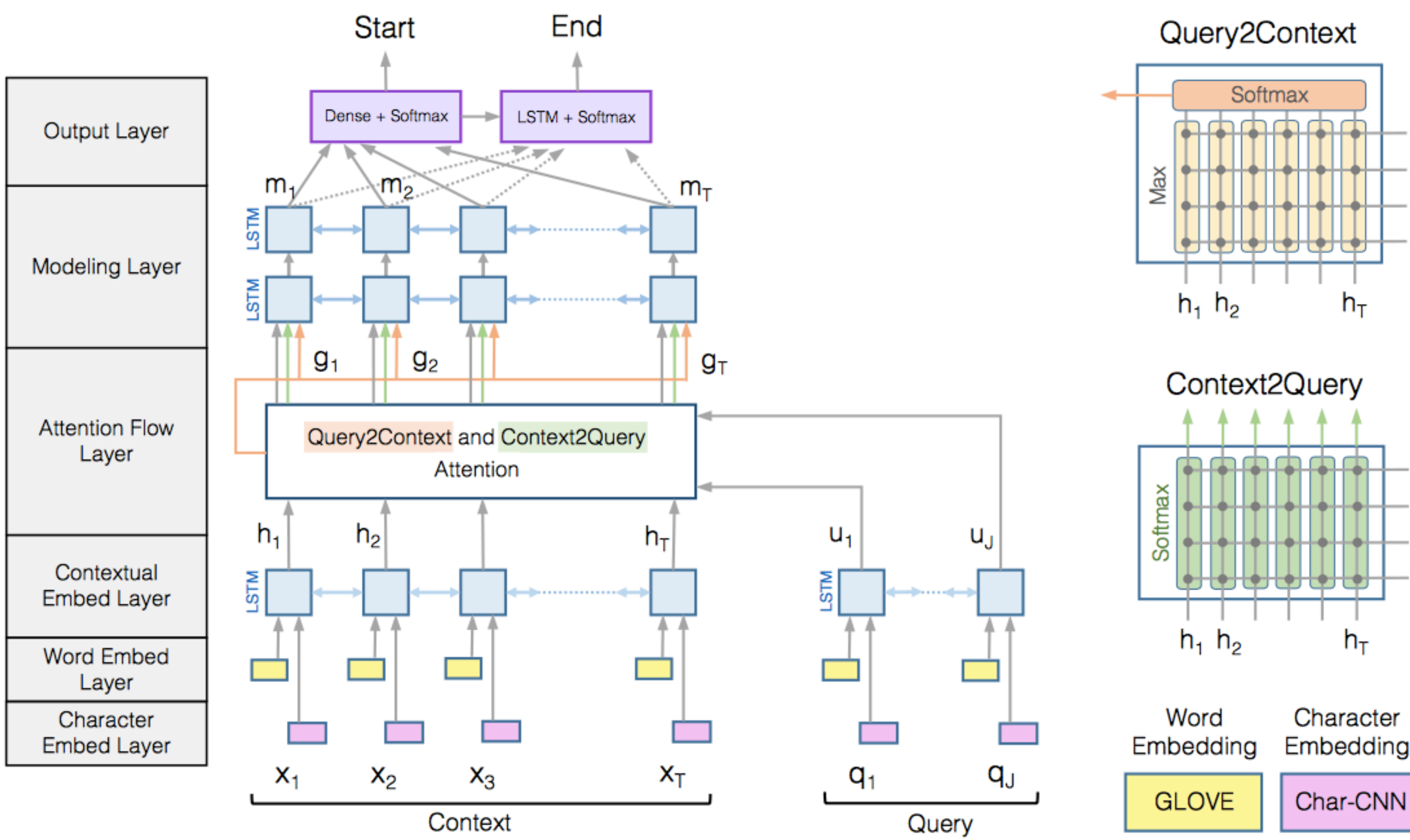


- Problem: encoding whole documents with high accuracy and coverage is hard!

A Popular Attempt: Bidirectional Attention Flow (BiDAF)

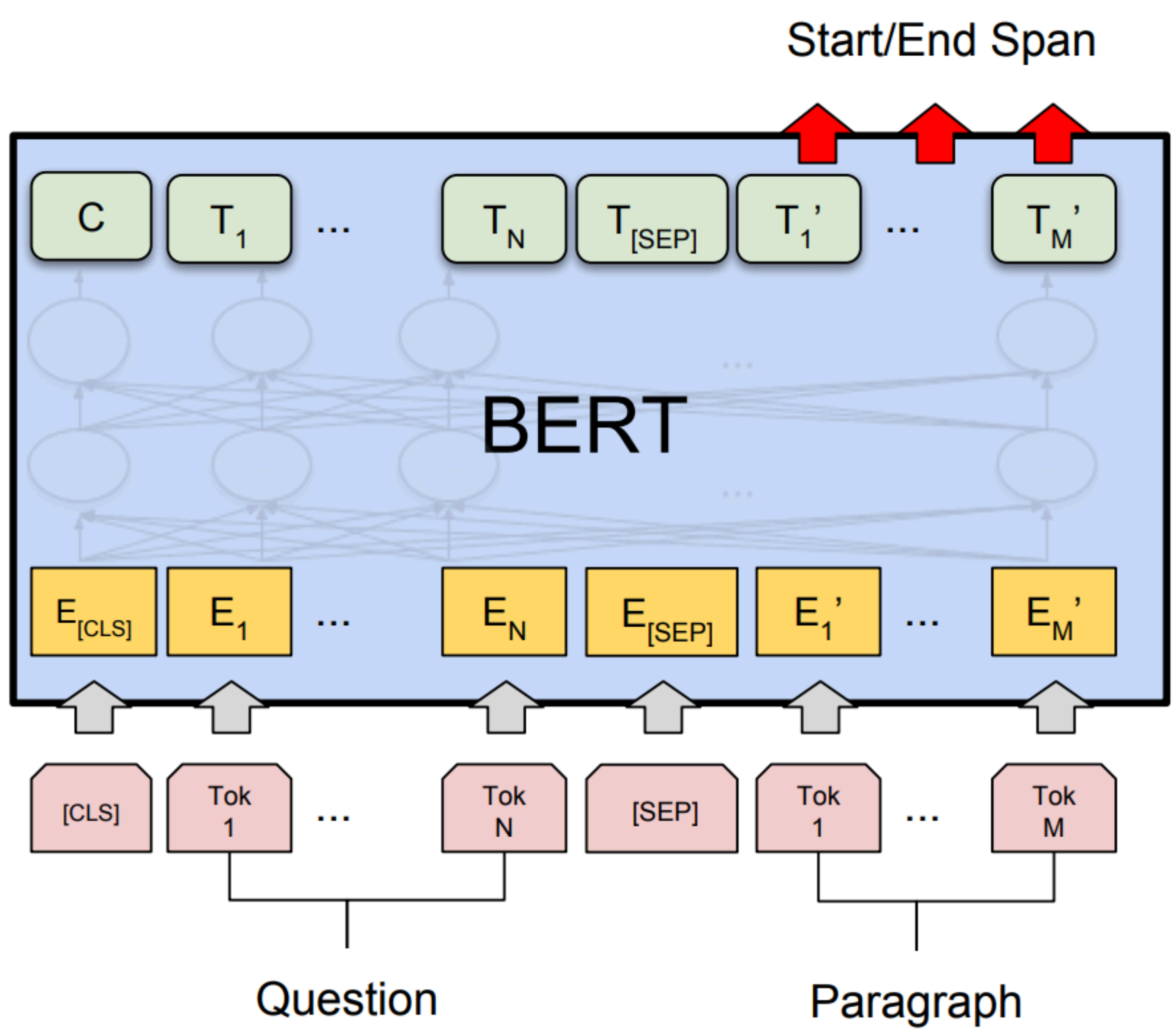
(Seo et al. 2017)

- Calculate query2ctxt, ctxt2query attention
- Both representations concatenated to word representations themselves in the document



Pre-trained Contextualized Representations

- Now standard to use BERT or other contextualized representations (Devlin et al. 2019)



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

Generative QA Models

- Feed in input passage and question, use decoder to output answer
- Example: UnifiedQA (Khashabi et al. 2020), trained on many different datasets
  - Format each dataset into input/output format
  - Base model: T5

EX	**Dataset**	SQuAD 1.1
	Input	At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
	Output	16,000 rpm
AB	**Dataset**	NarrativeQA
	Input	What does a drink from narcissus's spring cause the drinker to do? \n Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to ``Grow dotingly enamored of themselves.'' ...
	Output	fall in love with themselves
MC	**Dataset**	ARC-challenge
	Input	What does photosynthesis produce that helps plants grow? \n (A) water (B) oxygen (C) protein (D) sugar
	Output	sugar
	Dataset	MCTest
	Input	Who was Billy? \n (A) The skinny kid (B) A teacher (C) A little kid (D) The big kid \n Billy was like a king on the school yard. A king without a queen. He was the biggest kid in our grade, so he made all the rules during recess. ...
	Output	The big kid
YN	**Dataset**	BoolQ
	Input	Was America the first country to have a president? \n (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
	Output	no

Multi-hop Reasoning

Multi-hop 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

Multi-hop Reasoning Datasets

- Datasets explicitly created to require multiple steps through text
- Often labeled with "**supporting facts**" to demonstrate that multiple steps are necessary
- e.g. HotpotQA (Yang et al. 2018), WikiHop (Welbl et al. 2018)
- As always, be aware of dataset bias: we can **predict where the answer is** for more than half of questions in HotpotQA & WikiHop (Chen and Durrett 2019)

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

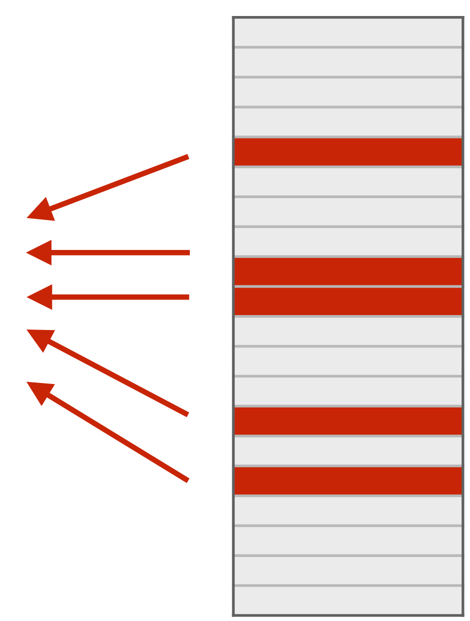
Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Memory Networks
(Weston et al. 2014)

- A general formulation of models that access **external memory** through attention and specific instantiation for document-level QA

Supporting sentences

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office? A: kitchen



Memory **m**

Memory Networks
(Weston et al. 2014)

- Given a question x, and an external memory **m**
- 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

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

Memory Networks

(Weston et al. 2014)

- During training, we are given a QA pair (x,r) as well as k ground-truth "supporting sentences" (i.e., m_{o1}, m_{o2}, ...) in the memory. When k=2, we minimize three margin-based hinge loss terms jointly:

$$\sum_{\bar{f} \neq \mathbf{m}_{o_1}} \max(0, \gamma - s_O(x, \mathbf{m}_{o_1}) + s_O(x, \underline{\bar{f}})) +$$

Random supporting sentences from the memory

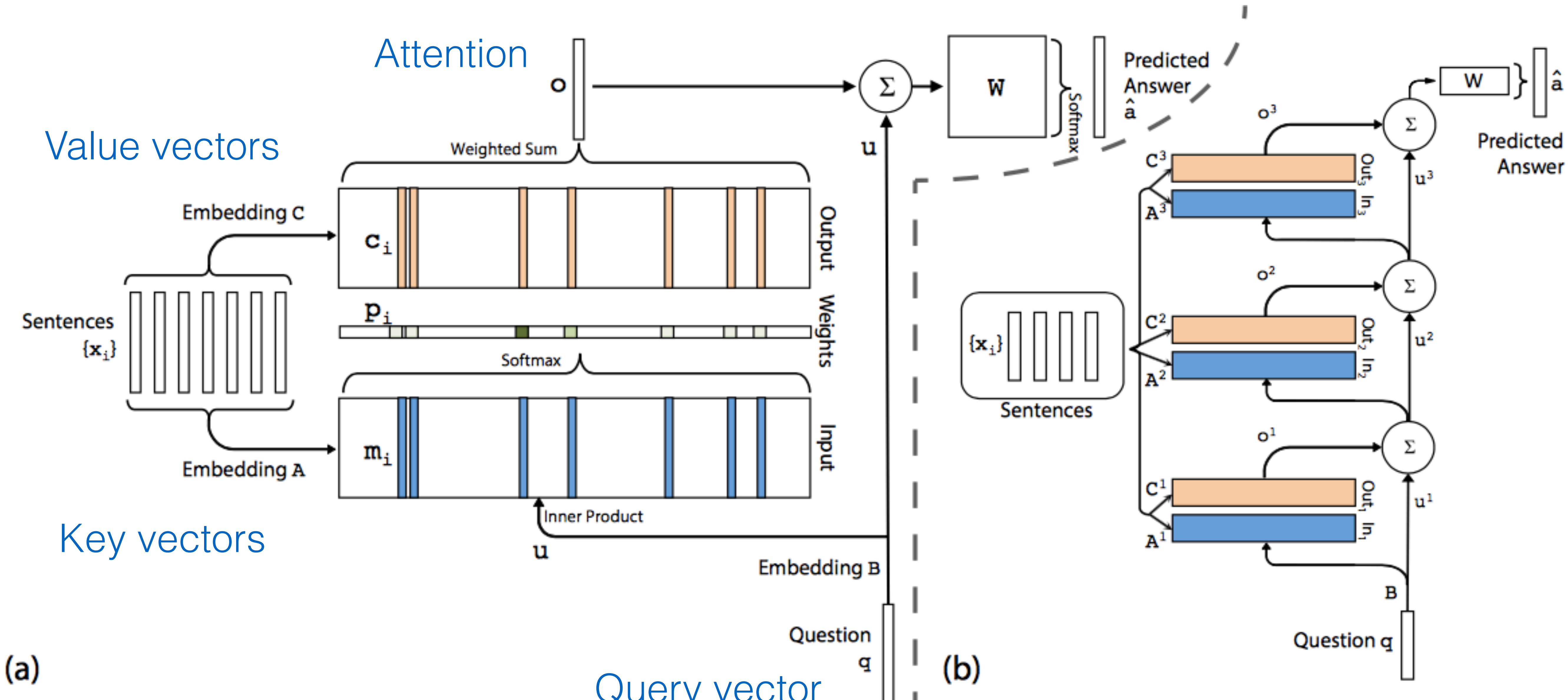
$$\sum_{\bar{f}' \neq \mathbf{m}_{o_2}} \max(0, \gamma - s_O([x, \mathbf{m}_{o_1}], \mathbf{m}_{o_2}]) + s_O([x, \mathbf{m}_{o_1}], \underline{\bar{f}'}])) +$$

$$\sum_{\bar{r} \neq r} \max(0, \gamma - s_R([x, \mathbf{m}_{o_1}, \mathbf{m}_{o_2}], r) + s_R([x, \mathbf{m}_{o_1}, \mathbf{m}_{o_2}], \underline{\bar{r}}]))$$

Random response

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

- Use standard softmax attention, and multiple layers



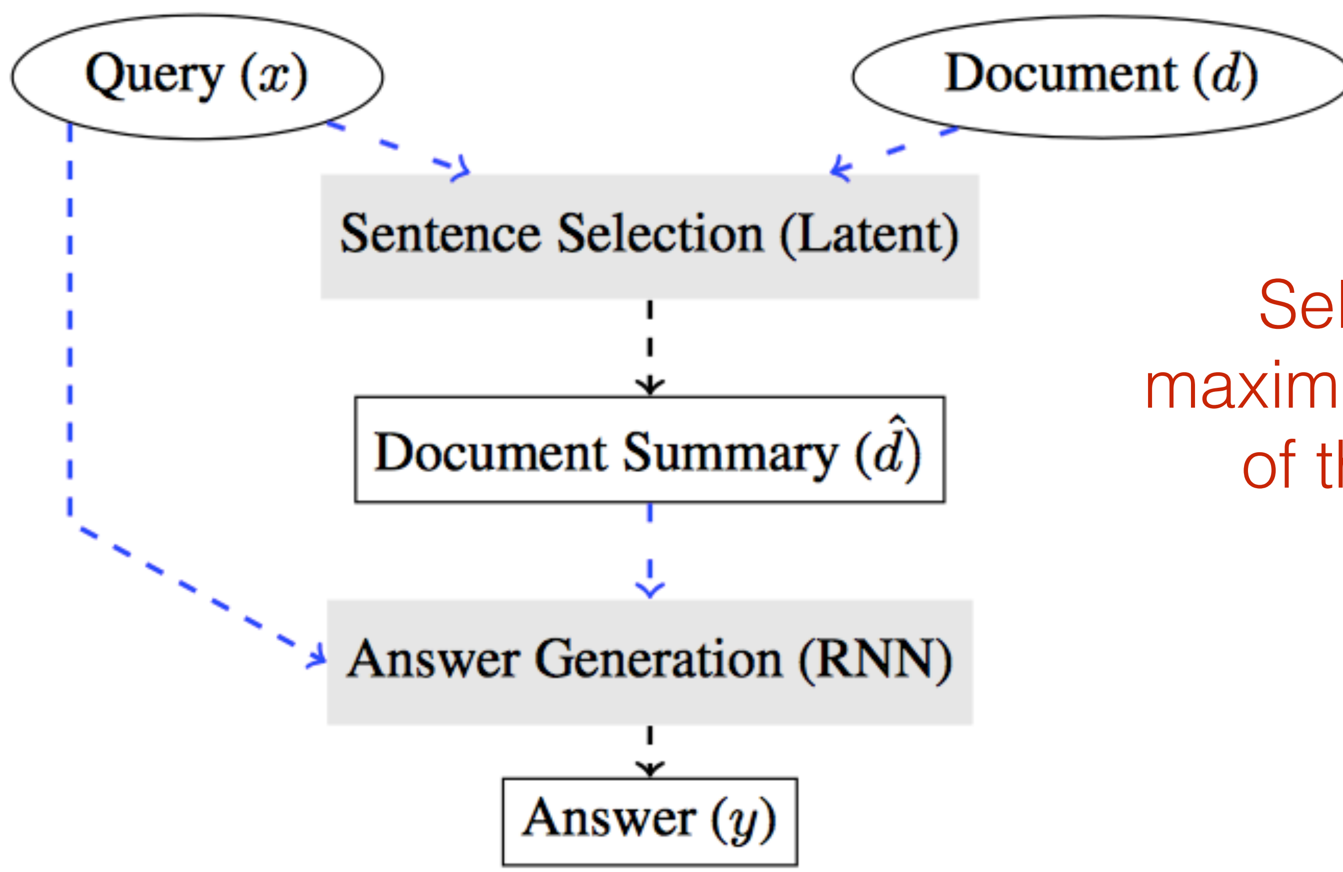
When to Stop Reasoning?

- 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)

Retrieval-based QA Models

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

- First, decide which sentence to cover, then reason



Select sentence that maximize the log probability of the correct answer.

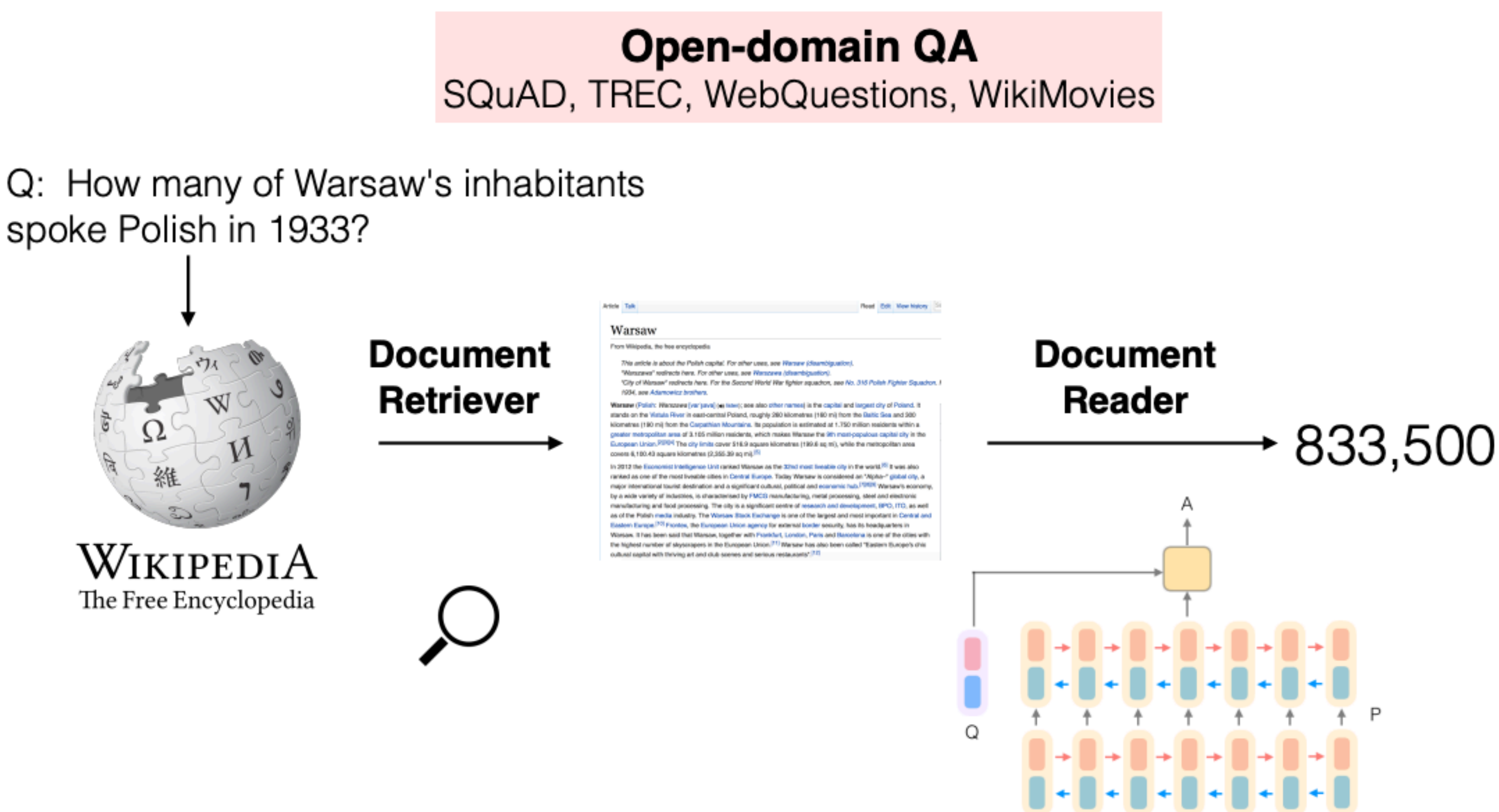
- Use reinforcement learning to selection sentences.
- This is also a variety of multi-hop reasoning

Choi et al. 2017. Coarse-to-Fine Question Answering for Long Documents

Retriever + Reader

(Chen et al. 2017)

- Retrieve a doc from Wikipedia, and apply a QA reader model to predict answer span.

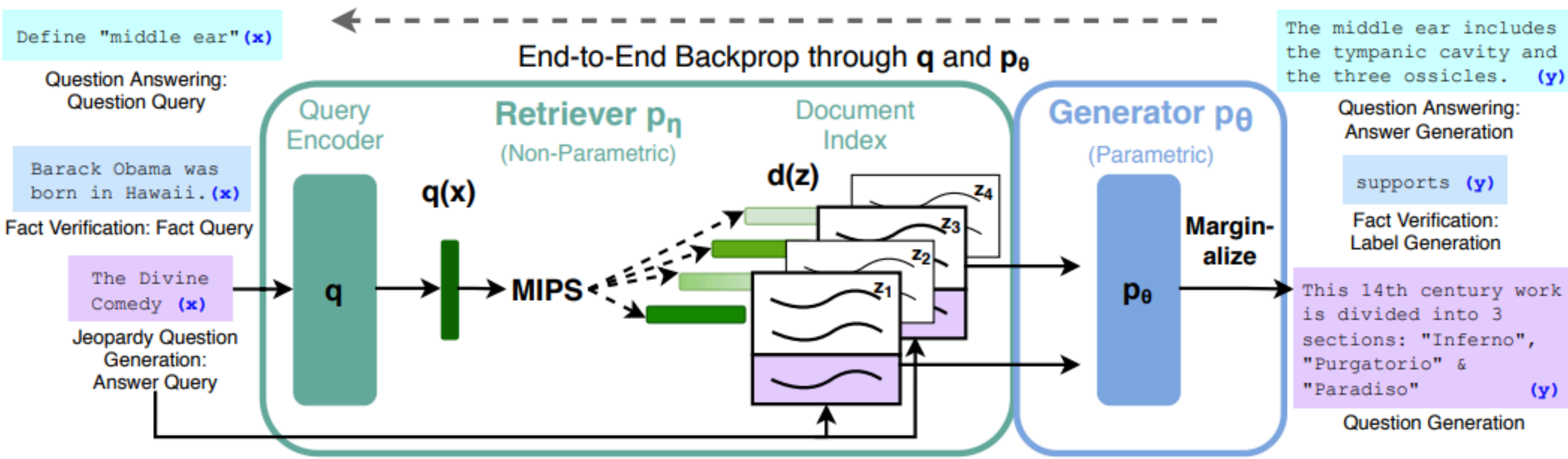


Chen et al. 2017. Reading Wikipedia to Answer Open-Domain Questions

Retrieval Augmented Generation (RAG)

(Lewis et al. 2020)

- Scale retrieval + generation based QA to large document collections (e.g. Wikipedia)



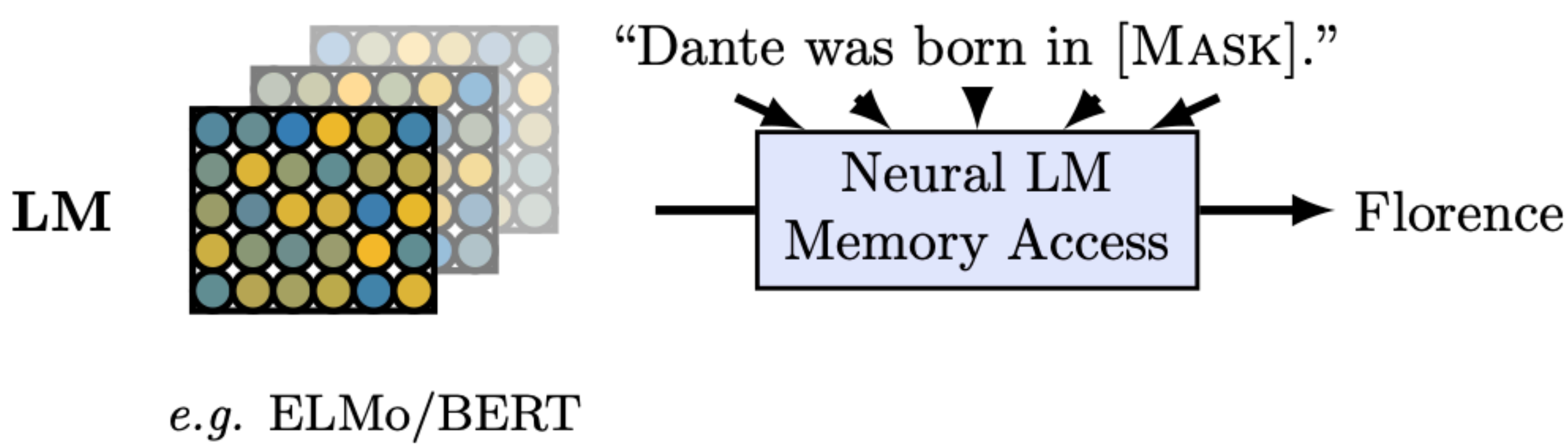
- Generative QA: given a question x, retrieve top-k docs {z}, feed x and z's to the decoder for generation

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right) \qquad \mathbf{d}(z) = \text{BERT}_d(z), \;\; \mathbf{q}(x) = \text{BERT}_q(x)$$

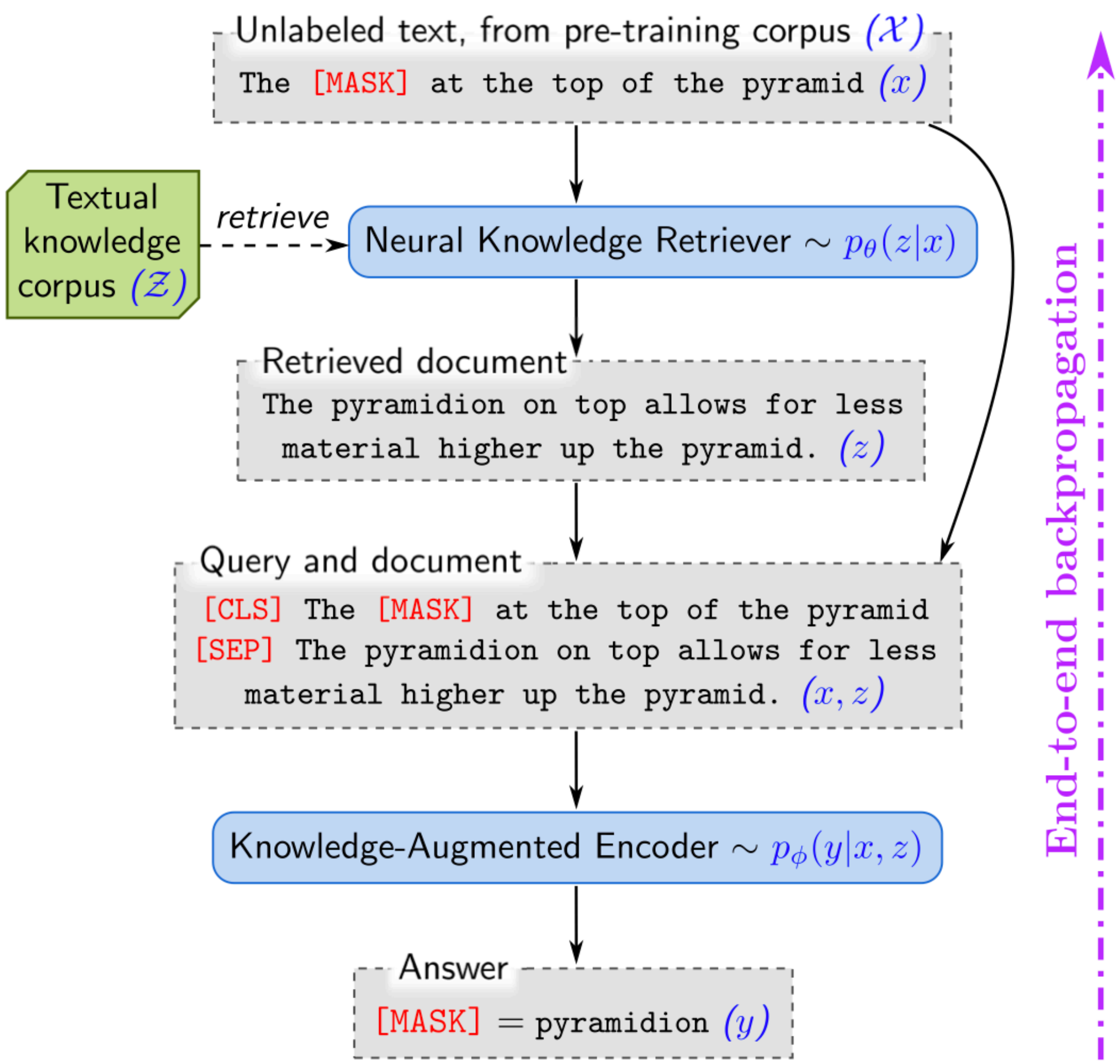
QA using LMs w/o Text Retrieval (Neural Knowledge Retrieval)

- A language model can be used to solve Cloze-style questions (e.g. Petroni et al. 2019)



Retrieval + Masked Language Model

- These two paradigms can be combined! Retrieve + masked language model (e.g. REALM, Guu et al. 2020)



Question Answering with Context (Choi et al. 2018, Reddy et al. 2018)

- Answer questions in sequence, so context from previous questions must be used in next answer

Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
...

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

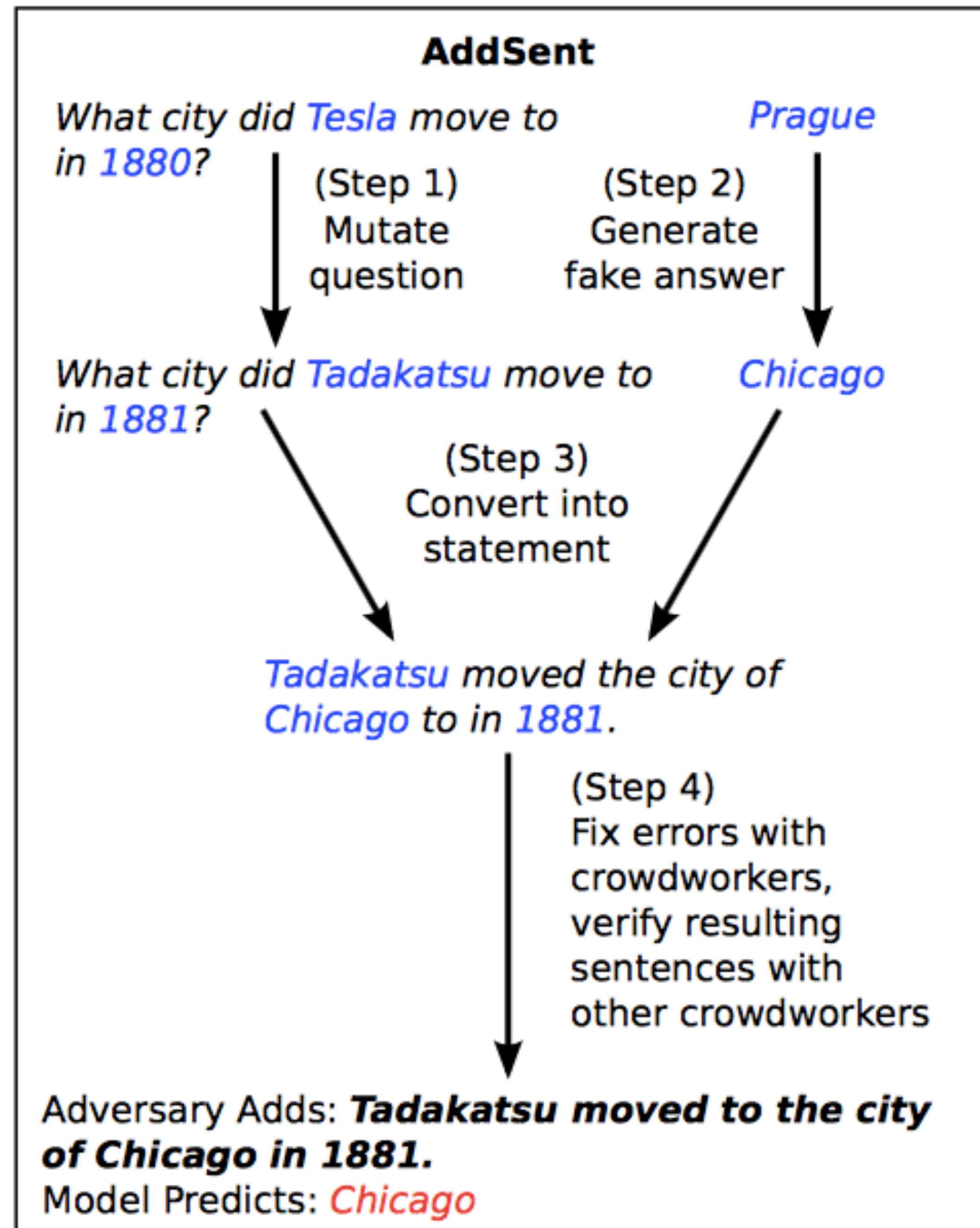
An Examination of CNN/ Daily Mail (Chen et al. 2015)

- 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



Adversarial Creation of New Datasets? (Zellers et al. 2018)

- **Idea:** create datasets that current models do poorly on, but humans do well
- **Process:**
  - Generate potential answers from LM
  - Find ones that QA model does poorly on
  - Have humans filter for naturalness
- **Problem:** Adversarial examples can be artificially hard/noisy, not representative

Natural Questions
(Kwiatkowski et al. 2019)

- Opposite approach:
  - create questions naturally from search logs
  - use crowdworkers to find corresponding evidence

Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

Example 3
Question: why does queen elizabeth sign her name elizabeth r
Wikipedia Page: Royal_sign-manual
Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).

Short answer: NULL

Questions?