Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering

Thomas Lukasiewicz

Hilary Term 2019

Questions

Question
When were the first pyramids built?
Jean-Claude Juncker
How old is Keir Starmer?
What is the current price for AAPL?
What's the weather like in London?
Whom did Juncker meet with?
When did you get to this lecture?
Why do we yawn?

Based on slides by Christopher Manning, Danqi Chen, Karl Moritz Hermann, and Edward Grefenstette.

Questions

Question	Answer
When were the first pyramids built?	*2630 BC*
Jean-Claude Juncker	*Jean-Claude Juncker is a Luxembourgish politician. Since 2014, Juncker has been President of the European Commission.*
How old is Keir Starmer?	*54 years*
What is the current price for AAPL?	*136.50 USD*
What's the weather like in London?	*7 degrees Celsius. Clear with some clouds.*
Whom did Juncker meet with?	*The European Commission president was speaking after meeting with Irish Taoiseach Enda Kenny in Brussels.*
When did you get to this lecture?	*Five minutes after it started.*
Why do we yawn?	*When we're bored or tired we don't breathe as deeply as we normally do. This causes a drop in our blood-oxygen levels and yawning helps us counter-balance that.*

Why do we care about question answering (QA)?

Because QA is awesome

1. QA is an AI-complete problem.
   If we solve QA, we have solved every other problem, too.
2. Many immediate and obvious applications
   Search, dialogue, information extraction, summarisation, ...
3. Some pretty nice results already
   IBM Watson and Jeopardy!, Siri, Google Search ...
4. Lots left to do!
   Plenty of interesting research and hard problems as well as low-hanging fruit.

Questions (again)

Question
When were the first pyramids built?
Jean-Claude Juncker
How old is Keir Starmer?
What is the current price for AAPL?
What's the weather like in London?
Whom did Juncker meet with?
When did you get to this lecture?
Why do we yawn?

Questions (again)

Question	Answer Source
When were the first pyramids built?	*Encyclopedia*
Jean-Claude Juncker	*Recent encyclopedia / Wikipedia*
How old is Keir Starmer?	*Very recent encyclopedia (i.e. Wikipedia). Extrapolate from date of birth.*
What is the current price for AAPL?	*NASDAQ Ticker*
What's the weather like in London?	*MET Office*
Whom did Juncker meet with?	*The Independent article "Jean-Claude Juncker doesn't want Northern Ireland and Republic to have post-Brexit hard border"*
When did you get to this lecture?	*Personal observation*
Why do we yawn?	*Various studies on the matter.*

Question answering depends on three kinds of data

And this gives us a good system for thinking about various QA tasks.

Question	**Context/Source**	**Answer**
Factual questions	Sets of documents (corpus)	A single fact
Complex/narrative questions	A single document	An explanation
Information Retrieval	Knowledge Base	A document
	Non-linguistic types of data (GPS, images, sensors, ...)	A sentence or paragraph extracted from somewhere
		An image or other type of object
		Another question

Question Taxonomy

Many possible taxonomies for questions

- Wh- words
- Subject of question
- The form of expected answers
- Types of sources from which answers may be drawn

For the purposes of building QA systems it is useful to start by considering the sources an answer may be drawn from.

Focus on the answer rather than the question.

QA Taxonomy Discovery

Three Questions for building a QA System

- What do the answers look like?
- Where can I get the answers from?
- What does my training data look like?

Areas in Question Answering

Reading Comprehension	• Answer based on a document • Context is a specific document
Semantic Parsing	• Answer is a logical form, possible executed against a KB • Context is a Knowledge Base
Visual QA	• Answer is simple and factual • Context is one/multiple image(s)
Information Retrieval	• Answer is a document/paragraph/sentence • Context is a corpus of documents

Question Answering: Example

Technical note: This is a "featured snippet" answer extracted from a web page, not a question answered using the (structured) Google Knowledge Graph (formerly known as Freebase).

Question Answering and Reading Comprehension

- With massive collections of full-text documents, i.e., the web simply returning relevant documents is of limited use
- Rather, we often want **answers** to our **questions**
- Especially on mobile
- Or using a digital assistant device, like Alexa, Google Assistant, ...

- We can factor this into two parts:
  1. Finding documents that (might) contain an answer
     - Which can be handled by traditional information retrieval/web search
  2. Finding an answer in a paragraph or a document
     - This problem is often termed **Reading Comprehension**

A Brief History of Reading Comprehension

- Much early NLP work attempted reading comprehension
  - Schank, Abelson, Lehnert et al. c. 1977 – "Yale A.I. Project"
- Revived by Lynette Hirschman in 1999:
  - Could NLP systems answer human reading comprehension questions for 3rd to 6th graders? Simple methods attempted.
- Revived again by Chris Burges in 2013 with MCTest
  - Again answering questions over simple story texts
- Floodgates opened in 2015/16 with the production of large datasets which permit supervised neural systems to be built
  - Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
  - Rajpurkar et al. (EMNLP 2016) SQuAD
  - MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, ...

Datasets and Models for QA/RC

Before 2015:

- MCTest (Richardson et al, 2013): 2600 questions
- ProcessBank (Berant et al, 2014): 500 questions

After 2015:

- **CNN/Daily Mail**
- Children Book Test
- WikiReading
- LAMBADA
- **SQuAD**
- Who did What
- Maluuba NewsQA
- MS MARCO

Datasets and Models for QA/RC

Before 2015:

- MCTest (Richardson et al, 2013): 2600 questions
- ProcessBank (Berant et al, 2014): 500 questions

After 2015:

- **CNN/Daily Mail**
- Children Book Test
- WikiReading
- LAMBADA
- **SQuAD**
- Who did What
- Maluuba NewsQA
- MS MARCO

More than 100k questions!

Datasets and Models for QA/RC

Before 2015:

- Lexical matching
- Logistic regression

After 2015:

- Neural networks

Datasets and Models for QA/RC

Before 2015:

- Lexical matching
- Logistic regression

After 2015:

Attentive Reader

Memory Networks

Gated-**attention** Reader

ReasoNet

Match-LSTM

Attention Sum Reader

Attention-over-**Attention** Reader

Iterative **Attentive** Reader

Dynamic **coattention** networks

Bi-directional **Attention** Flow Network

Multi-Perspective Context **Matching**

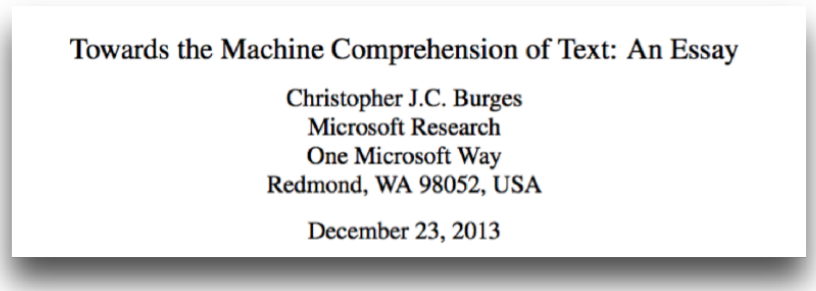
Reading Comprehension

- Candidate **reads** a document
- Candidate sees a **question about the document**
- Candidate must **select/generate an answer**
  - Span selection
  - Multiple Choice
  - Free form answers

Machine Comprehension

(Burges 2013)

"A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."


Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

MCTest Reading Comprehension

Passage (P) + Question (Q) $\longrightarrow$ Answer (A)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.....

Q What city is Alyssa in?

A Miami

"MCTest: A challenge dataset for the open-domain machine comprehension of text", Richardson et al., 2013. https://aclweb.org/anthology/D/D13/D13-1020.pdf

MCTest Reading Comprehension

Passage (P) + Question (Q) $\longrightarrow$ Answer (A)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.....

Q Why did Alyssa go to Miami?

A To visit some friends

MCTest Reading Comprehension

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.

Where did James go after he went to the grocery store?

1. his deck
2. his freezer
3. a fast food restaurant
4. his room

MCTest Reading Comprehension

What's great about it?

- Real data
- Relatively hard questions (many distractors)
- Some need for paraphrase, coreference resolution

What could be better?

- Little data (<3k questions from <1k articles)
- Multiple choice vs. free form.

A Brief History of Open-domain Question Answering

- Simmons et al. (1964) did first exploration of answering questions from an expository text based on matching dependency parses of a question and answer
- Murax (Kupiec 1993) aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing
- The NIST TREC QA track begun in 1999 first rigorously investigated answering fact questions over a large collection of documents
- IBM's Jeopardy! System (DeepQA, 2011) brought attention to a version of the problem; it used an ensemble of many methods
- DrQA (Chen et al. 2016) uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

Stanford Question Answering Dataset (SQuAD)

Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

"SQuAD: 100,000+ questions for machine comprehension of text", Rajpurkar et al., 2016. https://arxiv.org/pdf/1606.05250.pdf

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

SQuAD Evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate
    Precision = tp/(tp+fp), Recall = tp/(tp + fn), harmonic mean F1 = 2PR/(P+R)
    Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a**, **an**, **the** only)

SQuAD v1.1 Leaderboard, 2019-02-07

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
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or "Answer validation"

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

SQuAD 2.0 Example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234 [from Microsoft nlnet]

SQuAD 2.0 leaderboard, 2019-02-07

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035

Example

Good systems are great, but still basic NLU errors:

The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

Prediction: Ming dynasty [BERT (single model) (Google AI)]

SQuAD Limitations

- SQuAD has a number of other key limitations too:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at the passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
  - Barely any multi-fact/sentence inference beyond coreference
- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - It has been the most used and competed on QA dataset
  - It has also been a useful starting point for building systems in industry (though in-domain data always really helps!)

bAbI Dataset (Facebook)

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.

Q: Where was the apple before the kitchen?

A: office

https://research.fb.com/downloads/babi/

bAbI Dataset (Facebook)

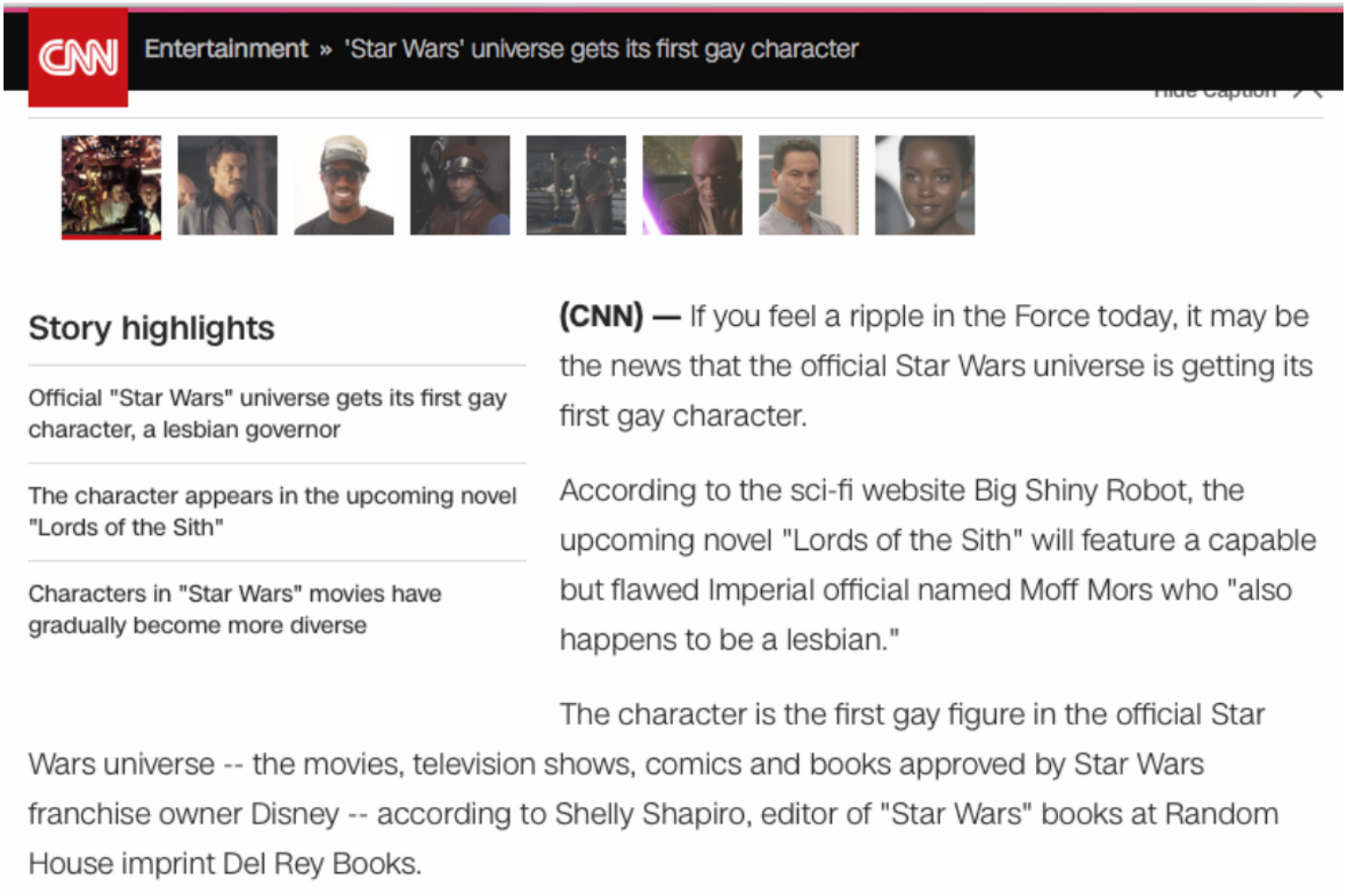
What's so great about it?

- Lots of data (can generate more)
- Many sub-tasks based on modes of reasoning

What could be better?

- Synthetic data (better as a unit test)
- Very predictable structure
- Lack of grammatical or lexical diversity

CNN / Daily Mail Dataset (DeepMind)



CNN Entertainment » 'Star Wars' universe gets its first gay character

Hide Caption

Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in "Star Wars" movies have gradually become more diverse

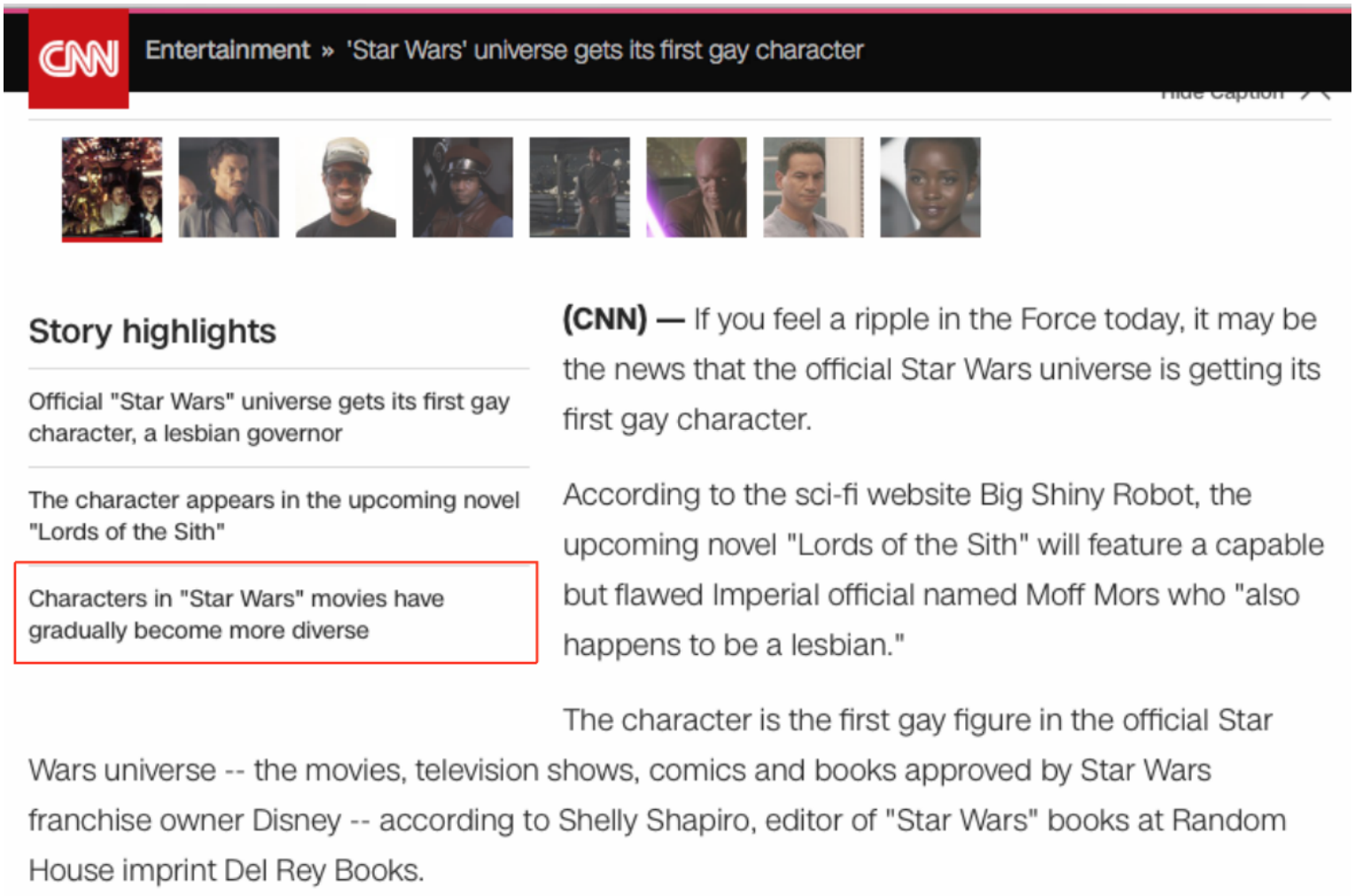
(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

https://github.com/deepmind/rc-data/

CNN / Daily Mail Dataset (DeepMind)



CNN Entertainment » 'Star Wars' universe gets its first gay character

Hide Caption

Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

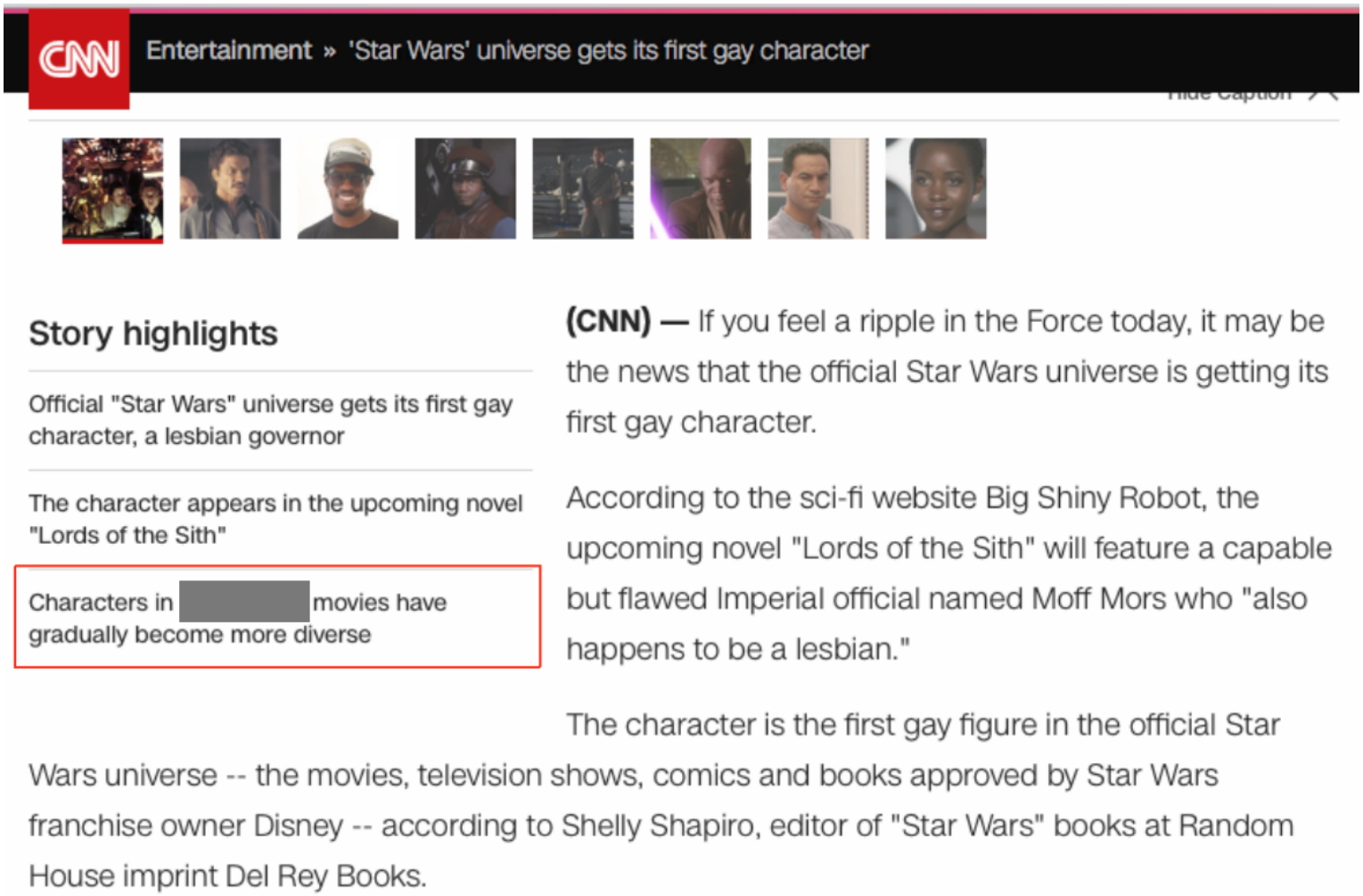
Characters in "Star Wars" movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

CNN / Daily Mail Dataset (DeepMind)



CNN / Daily Mail Dataset (DeepMind)

P

(@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 "

Q

characters in " @placeholder " movies have gradually become more diverse

A

@entity6

CNN / Daily Mail Dataset (DeepMind)

P

(@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 "

Q

characters in " @placeholder " movies have gradually become more diverse

A

@entity6

CNN: 380k, Daily Mail: 879k training - free!

CNN / Daily Mail Dataset (DeepMind)

What's so great about it?

- Lots of data (easy to generate more)
- Natural text (Cloze-questions from abstractive summaries)

What could be better?

- Semi-synthetic (lack of variety, unanswerable questions)
- Answers basically are just pointers to entities

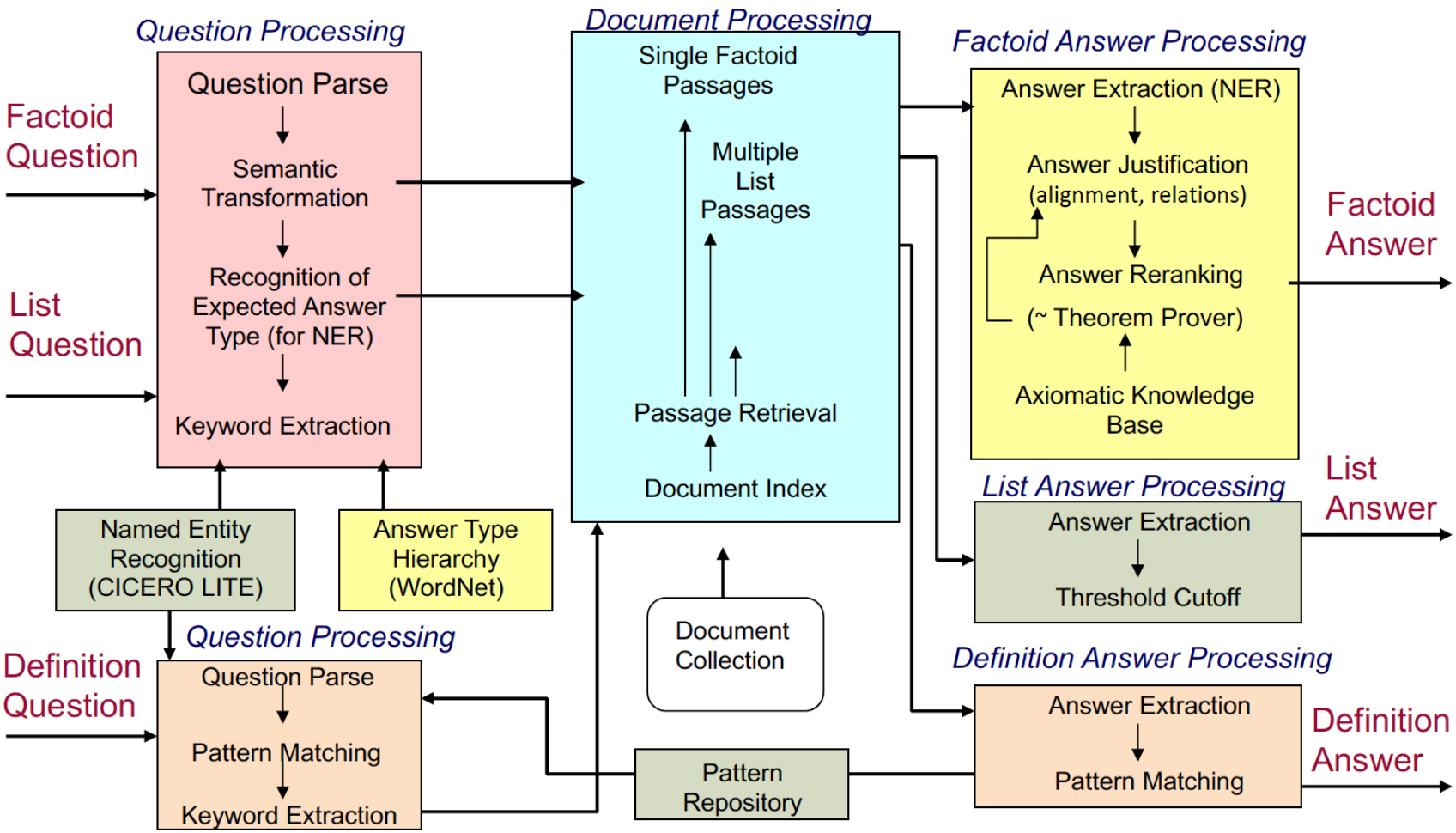
What's missing in datasets?

- Lots of natural, human-written, question/answer pairs.
- Questions that require linking information across different parts of the text (i.e. beyond simple anaphora resolution)
- More diverse questions. Not just who/what/where, but:
  - "how?", "in what manner …?".
  - Temporal questions.
  - Questions about abstract relations, narrative structure.

Turn-of-the-Millennium Full NLP QA

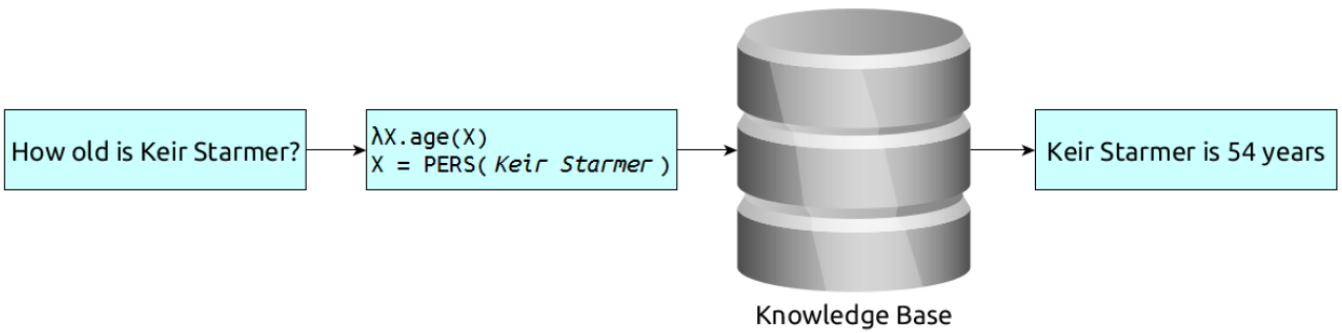
Architecture of LCC (Harabagiu/Moldovan) QA system, 2003.

Complex systems but they did work fairly well on factoid questions.



Semantic Parsing

Semantic Parsing is the process of mapping natural language into a formal representation of its meaning. Depending on the chosen formalism this logical representation can be used to query a structured knowledge base.



Question → Logical Form → KB Query → Answer

Semantic Parsing is Question→Logical Form.
We (often mistakenly) then assume that LF→Answer is trivial.

Knowledge Bases for QA with Semantic Parsing

Knowledge bases typically represent their data as triples
(married-to, Michelle Obama, Barack Obama)
(member-of, United Kingdom, European Union)
Generally: *(relation, entity1, entity2)*

There are several (large) databases freely available to use, e.g.:

Freebase 1.9 billion triples on general knowledge. Defunct as of 2016 and replaced by Google Knowledge Graph

WikiData Information on 25 million entities

OpenStreetMap 3 billion triples on geography

GeoQuery 700 facts about US geography. Tiny dataset, but frequently used in semantic parsing work.

KBs are cheap — Supervised Data is expensive!

Free917	917 freebase annotated questions
GeoQuery	880 questions on US geography
NLMaps	2,380 natural language queries on the OSM data

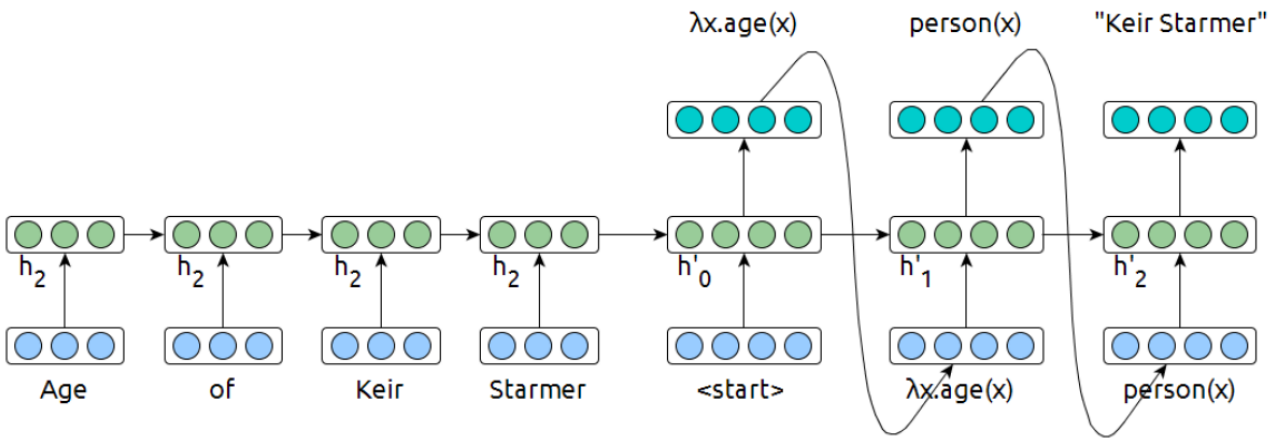
These kinds of datasets are incredibly expensive to create as they require experts for the manual annotation process, who are trained in using a given database schema:

"Where are kindergartens in Hamburg?"

```
query(area(keyval(name,Hamburg)),
      nwr(keyval(amenity,kindergarten)),
      qtype(latlong))
```

A Deep Learning Approach to Semantic Parsing

Semantic parsing can be viewed as a sequence to sequence model, not unlike machine translation.



Details

✔ Encode sentence with sequence models

✔ Decode with standard mechanisms from MT

✘ Supervised training data hard to come by

✘ Depending on formalism used, highly complex target side

✘ How to deal with proper nouns and numbers?

Semantic Parsing Summary

✔ LF instead of answer makes system robust
✔ Answer independent of question and parsing mechanism
✔ Can deal with rapidly changing information

✘ Constrained to queriable questions in DB schema
✘ No database is large enough
✘ Training data hard to find

✔ When were the pyramids built?
? Jean-Claude Juncker
✔ How old is Keir Starmer?
✔ What is the price for AAPL?
✔ What's the weather in London?
✘ Whom did Juncker meet with?
✘ When did you get here?
✘ Why do we yawn?

Caveat: Each of these examples requires a different underlying KB!

Stanford Attentive Reader

"A thorough examination of the CNN/Daily Mail reading comprehension task", Chen et al., 2016.
https://arxiv.org/pdf/1606.02858.pdf

Improved version of Attentive Reader in "Teaching machines to read and comprehend", Hermann et al., 2015.
https://arxiv.org/pdf/1506.03340.pdf

CNN / Daily Mail Dataset (DeepMind)

P

(@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 "

Q

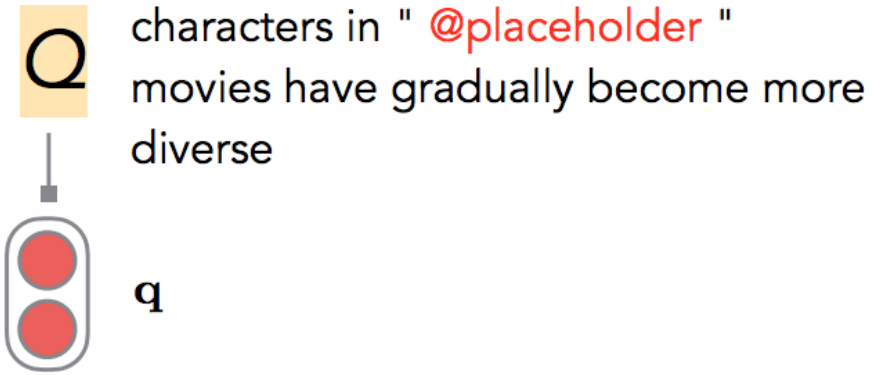
characters in " @placeholder " movies have gradually become more diverse

A

@entity6

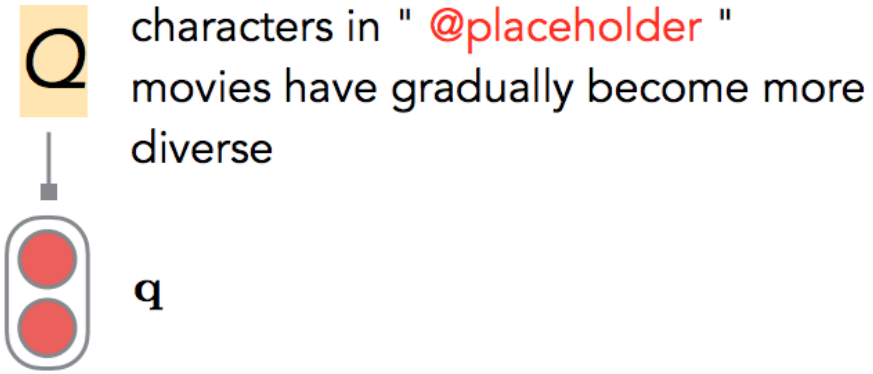
Stanford Attentive Reader

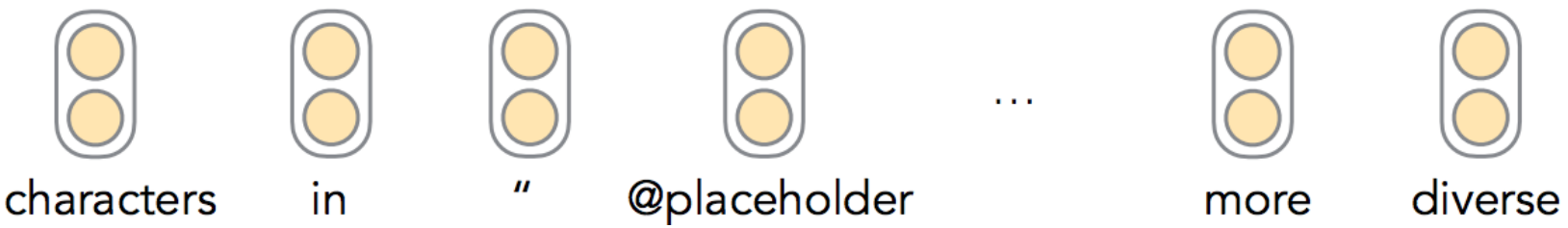
Bidirectional LSTMs



Stanford Attentive Reader

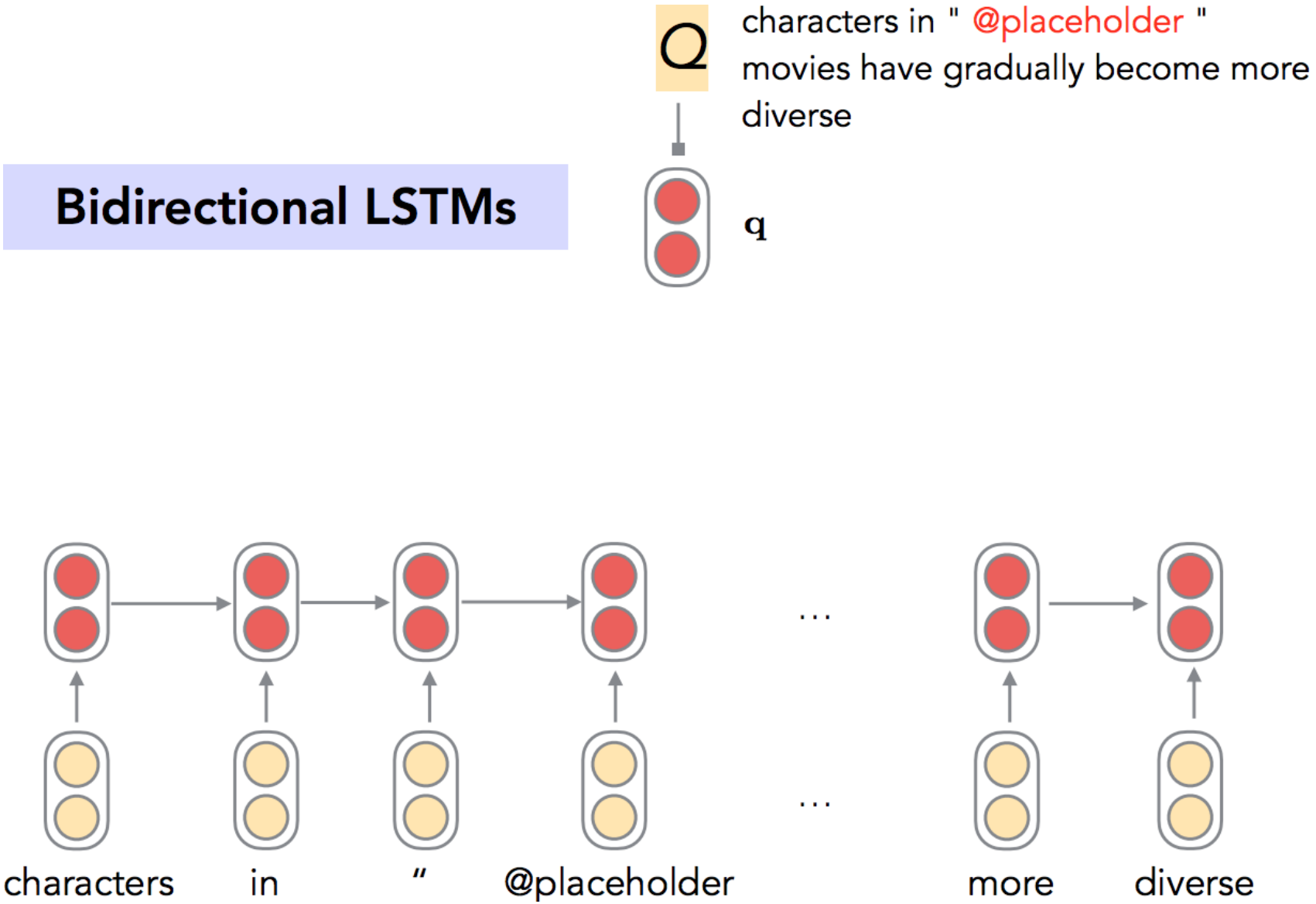
Bidirectional LSTMs



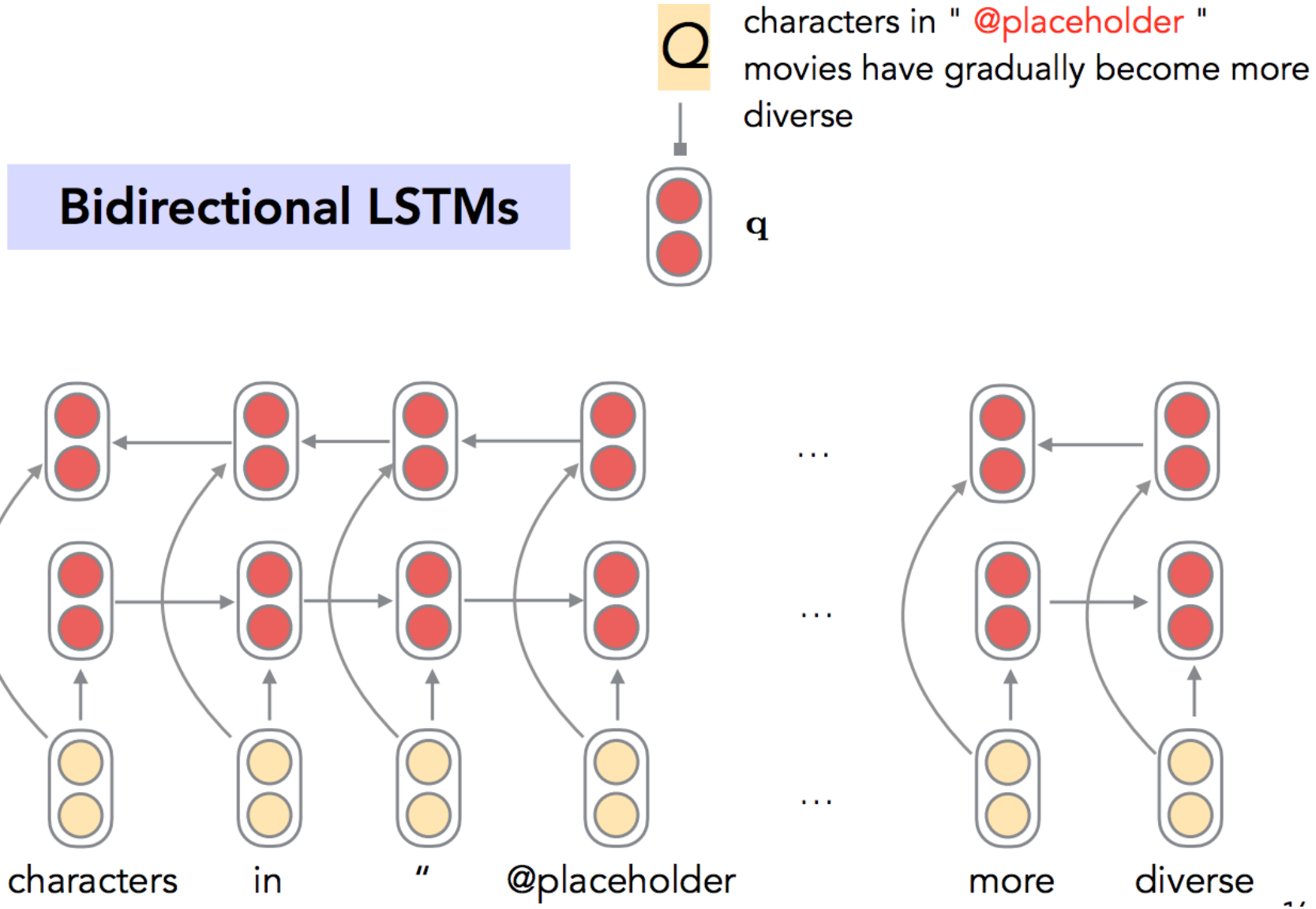


Stanford Attentive Reader

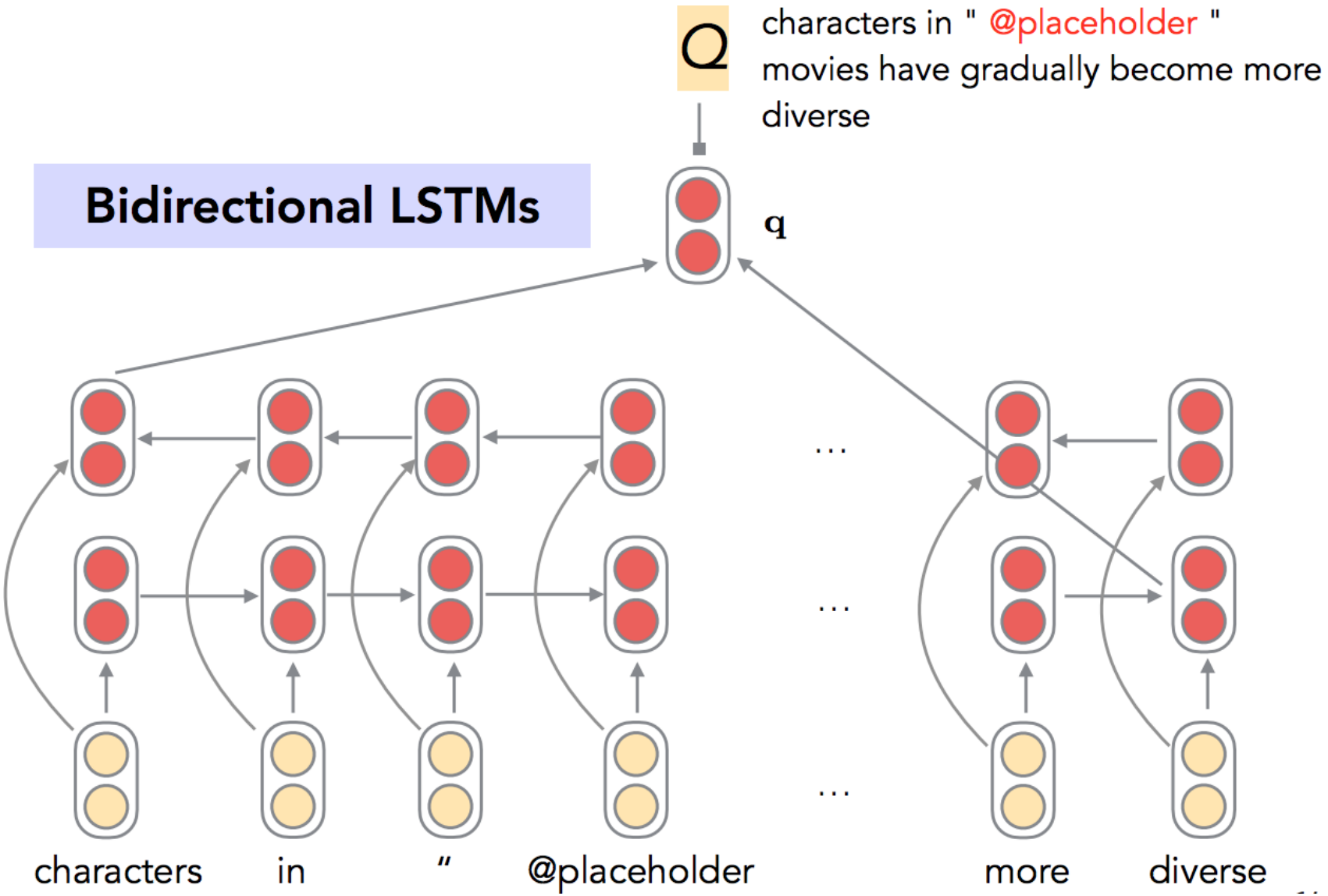
Bidirectional LSTMs



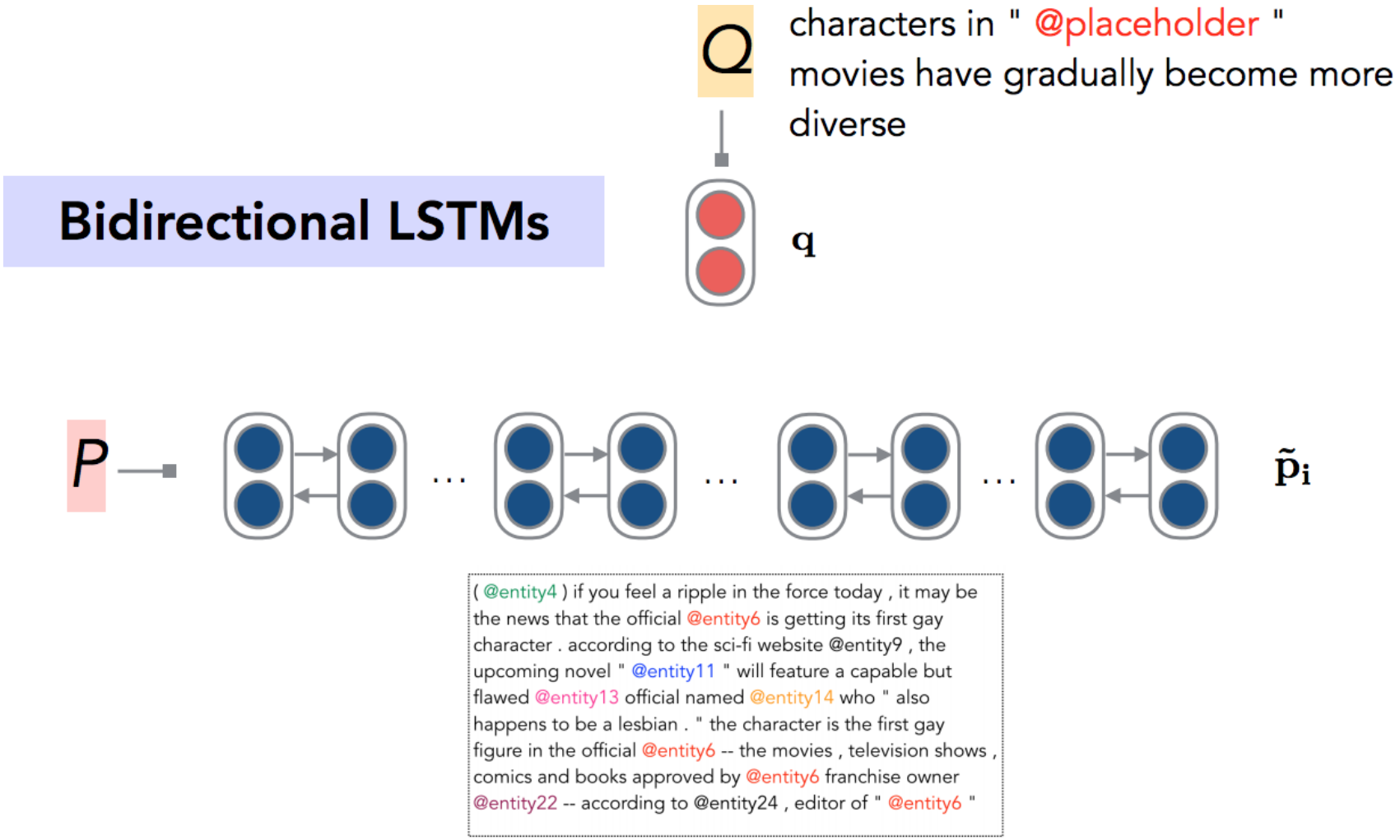
Stanford Attentive Reader



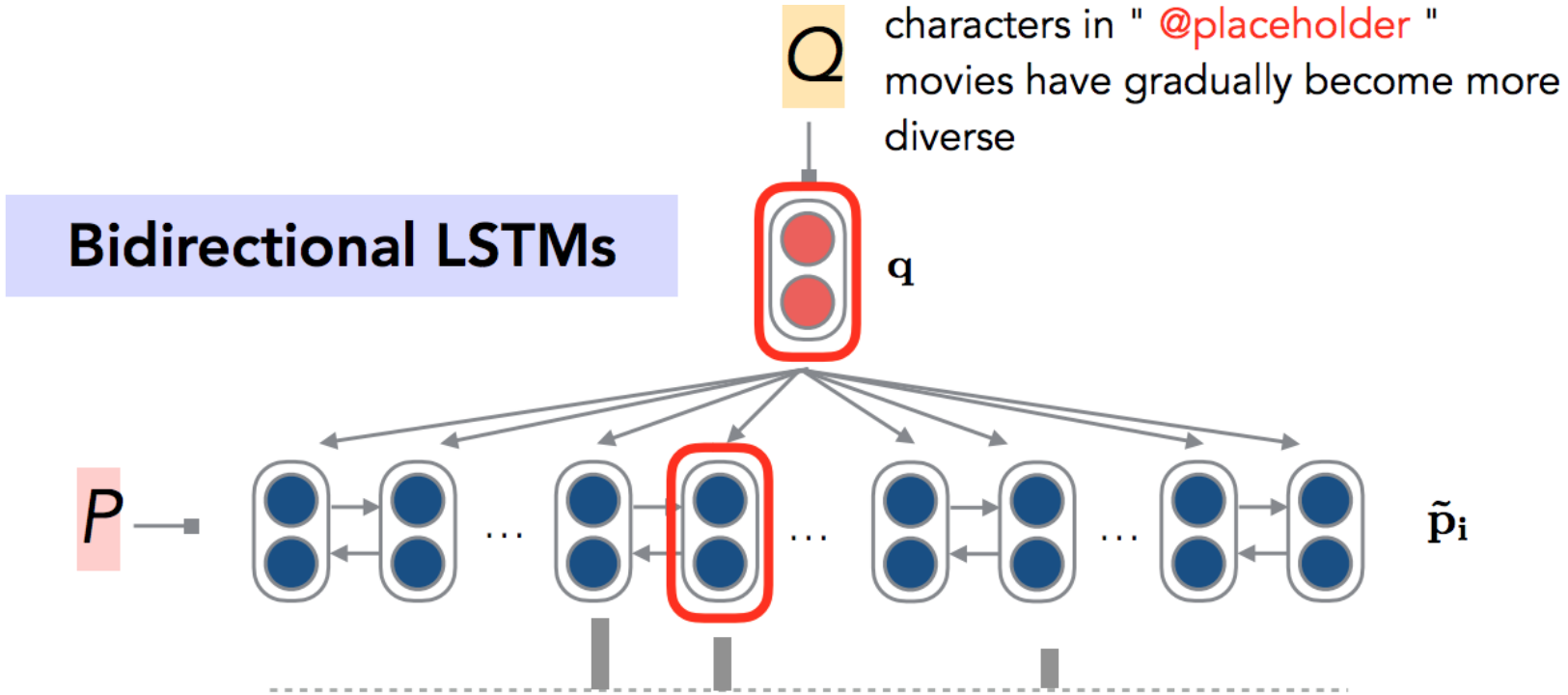
Stanford Attentive Reader



Stanford Attentive Reader



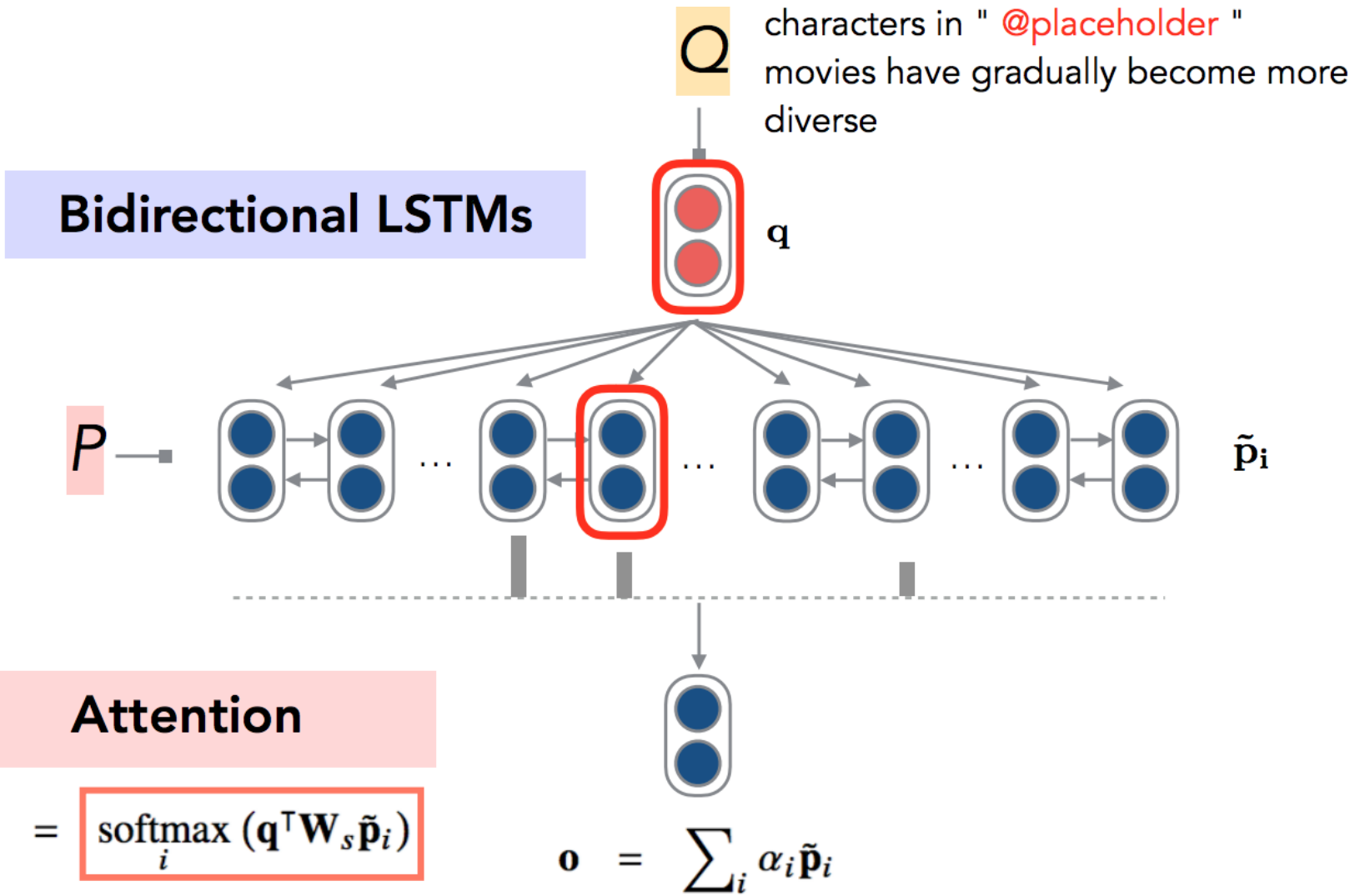
Stanford Attentive Reader



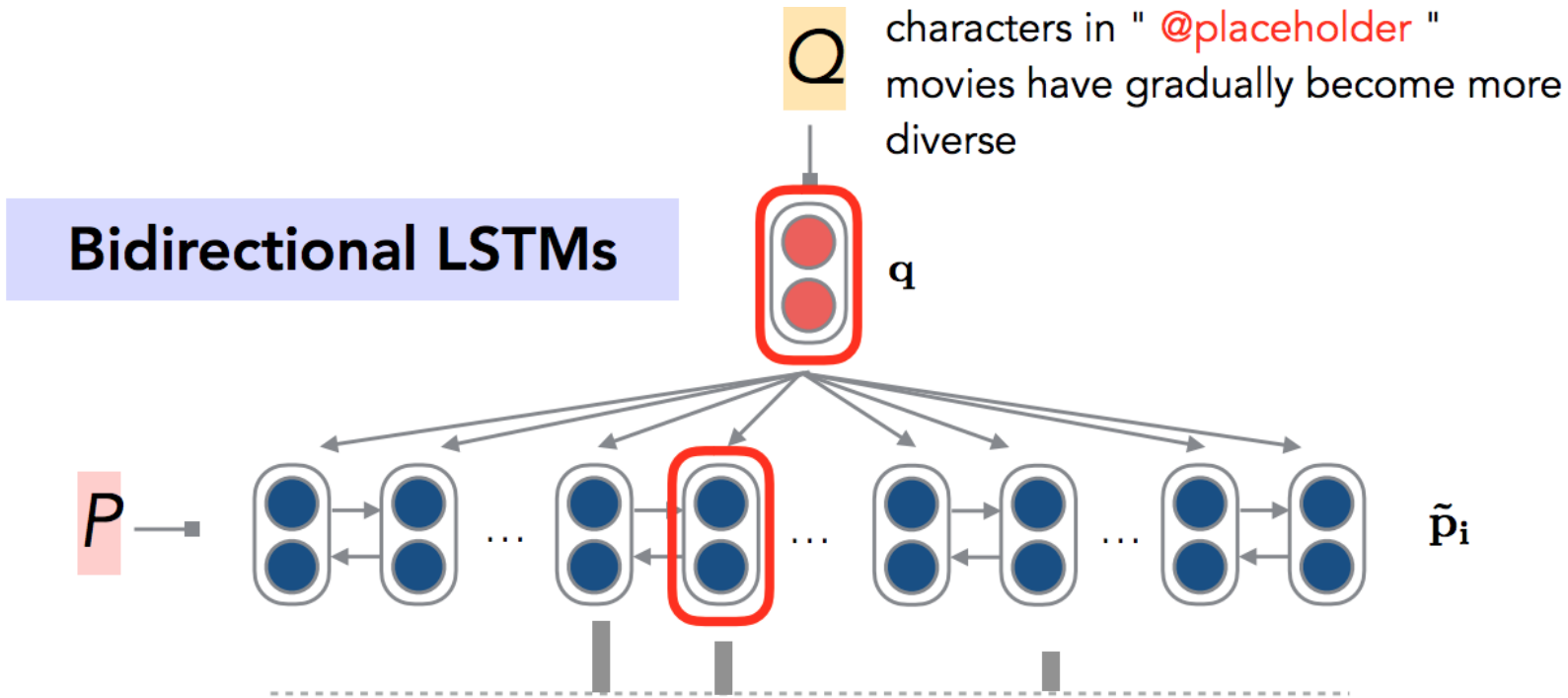
Attention

$$\alpha_i = \underset{i}{\text{softmax}}\left(\mathbf{q}^\top \mathbf{W}_s \tilde{\mathbf{p}}_i\right)$$

Stanford Attentive Reader



Stanford Attentive Reader

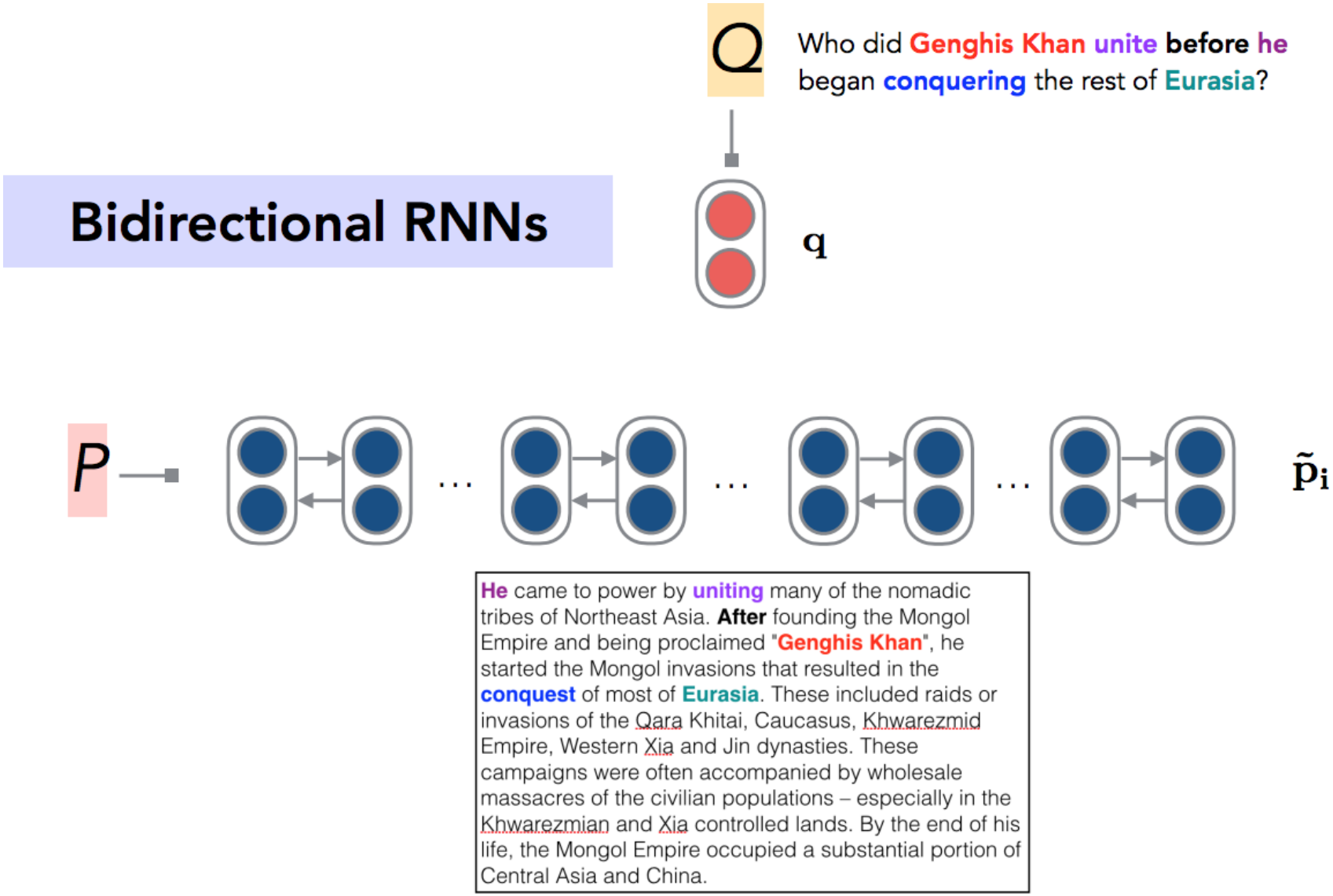


$$\alpha_i = \underset{i}{\text{softmax}}\left(\mathbf{q}^\intercal \mathbf{W}_s \tilde{\mathbf{p}}_i\right)$$

$$\mathbf{o} = \sum_i \alpha_i \tilde{\mathbf{p}}_i$$

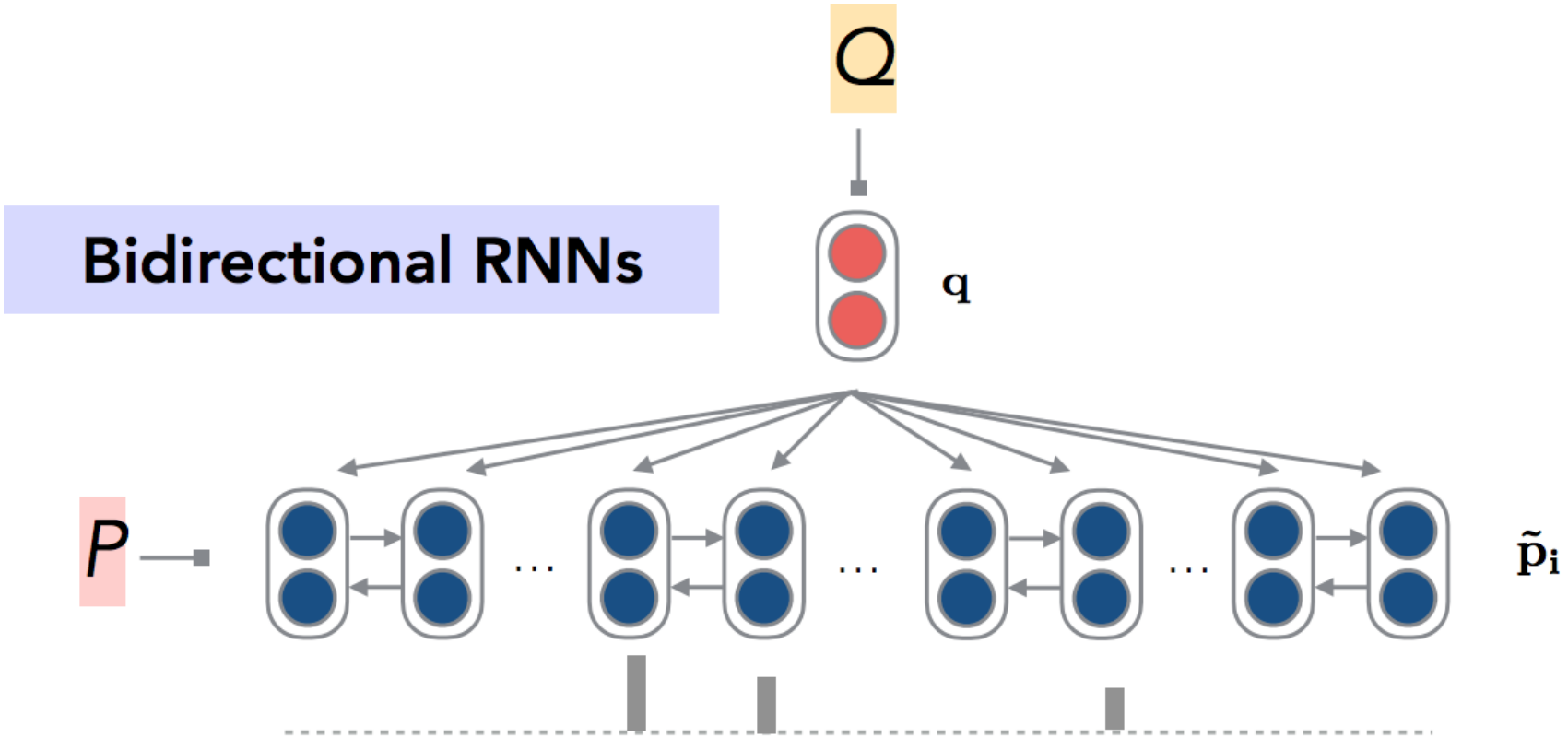
$$a = \arg\max_{a \in p \cap E} W_a^\intercal \mathbf{o}$$

Stanford Attentive Reader++



"Reading Wikipedia to answer open-domain questions", Chen et al., 2017.
https://arxiv.org/pdf/1704.00051.pdf

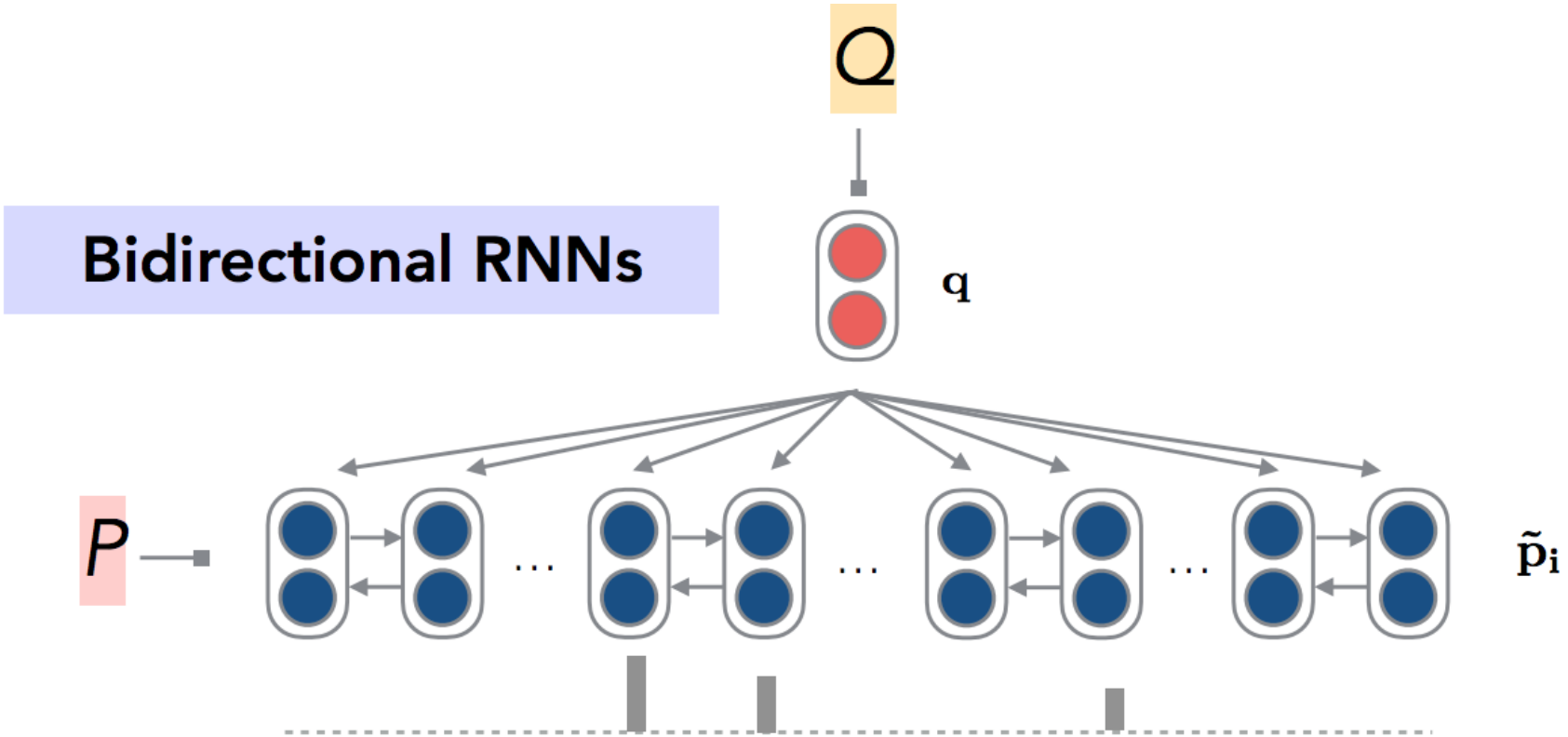
Stanford Attentive Reader++



Attention

$$\alpha_i = \underset{i}{\text{softmax}}\left(\mathbf{q}^\intercal \mathbf{W}_s \tilde{\mathbf{p}}_i\right)$$

Stanford Attentive Reader++

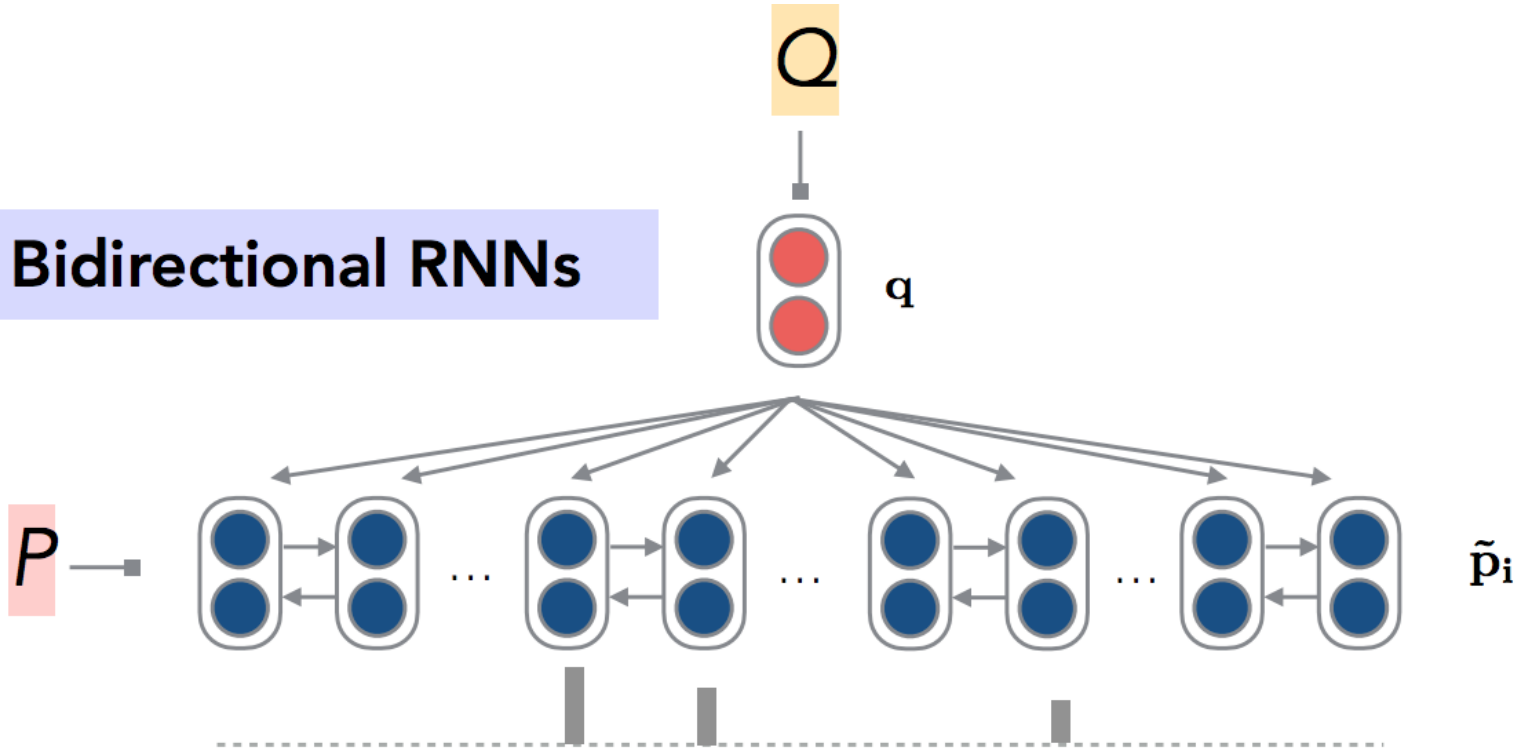


Attention

$$\alpha_i = \underset{i}{\text{softmax}}\left(\mathbf{q}^\top \mathbf{W}_s \tilde{\mathbf{p}}_i\right)$$

⟶ predict **start** token

Stanford Attentive Reader++



Attention

$$\alpha_i = \underset{i}{\text{softmax}}\left(\mathbf{q}^\intercal \mathbf{W}_s \tilde{\mathbf{p}}_i\right)$$

⟶ predict **start** token

Attention

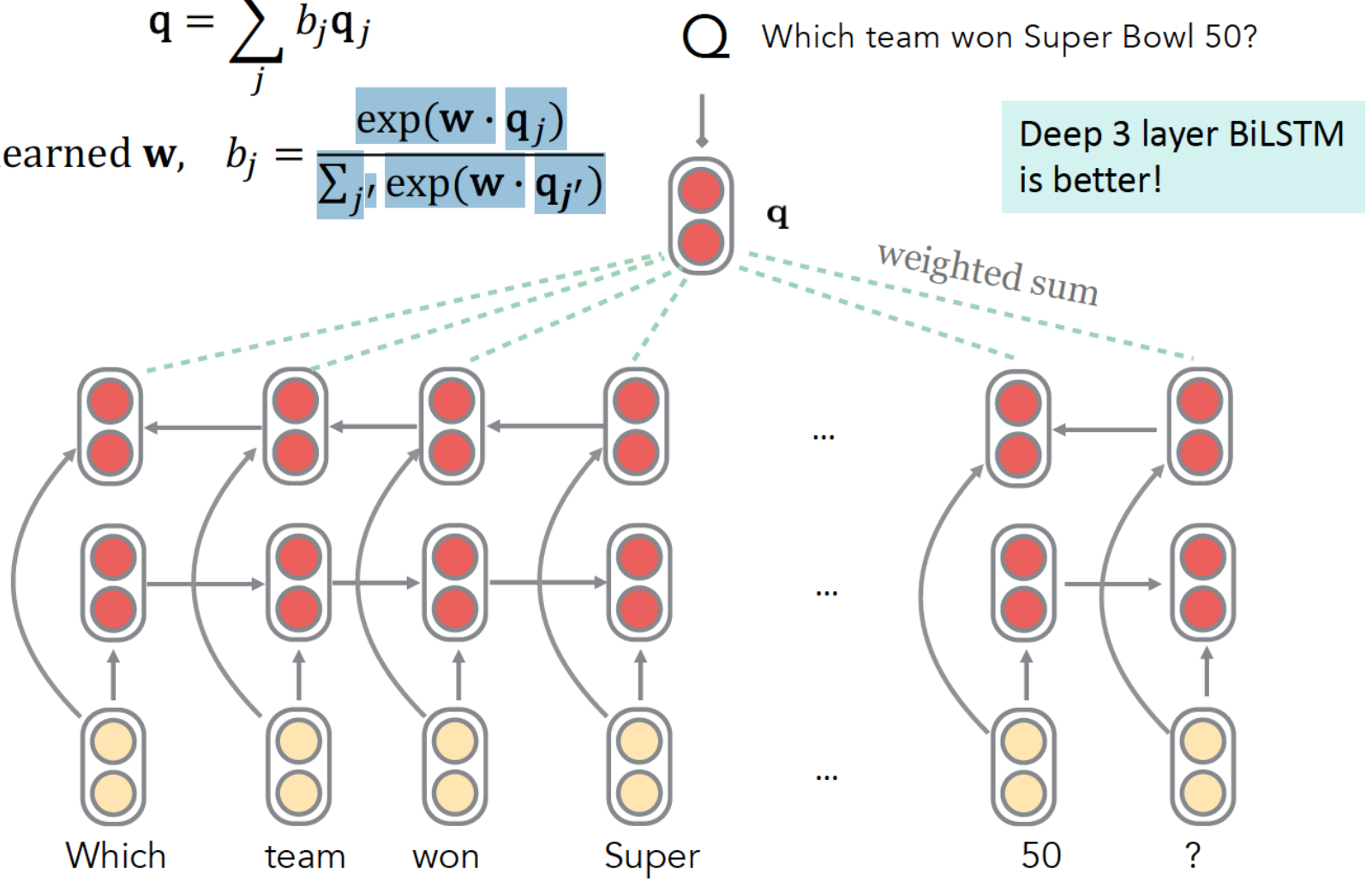
$$\alpha'_i = \text{softmax}_i\left(\mathbf{q}^\intercal \mathbf{W}'_s \tilde{\mathbf{p}}_i\right)$$

⟶ predict **end** token

Stanford Attentive Reader++

$$\mathbf{q} = \sum_j b_j \mathbf{q}_j$$

For learned $\mathbf{w}$, $b_j = \frac{\exp(\mathbf{w} \cdot \mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w} \cdot \mathbf{q}_{j'})}$



Stanford Attentive Reader++

- $\mathbf{p}_i$: Vector representation of each token in passage

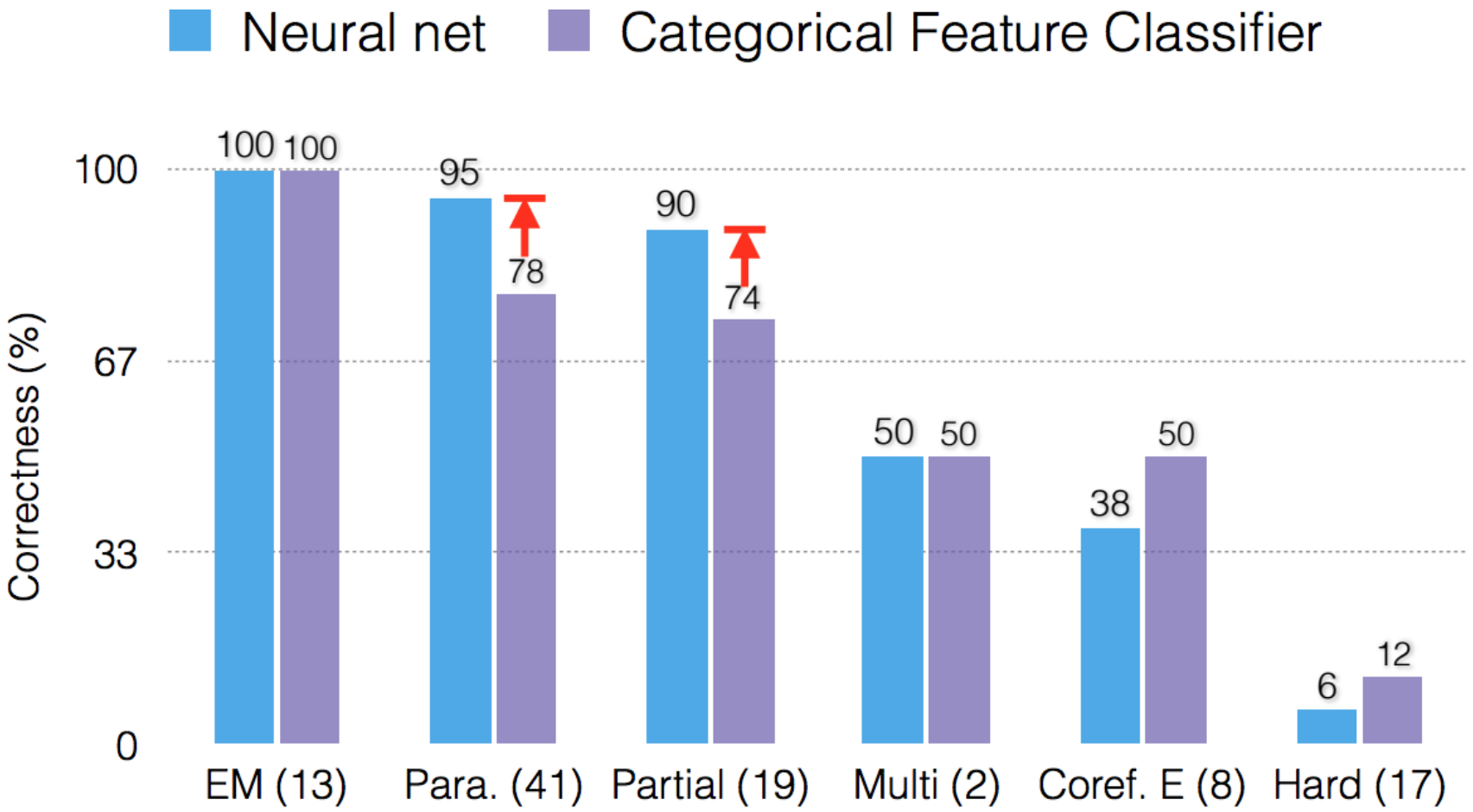
Made from concatenation of

- Word embedding (GloVe 300d)
- Linguistic features: POS & NER tags, one-hot encoded
- Term frequency (unigram probability)
- Exact match: whether the word appears in the question
  - 3 binary features: exact, uncased, lemma
- Aligned question embedding ("car" vs "vehicle")

$$f_{align}(p_i) = \sum_j a_{i,j}\mathbf{E}(q_j) \qquad q_{i,j} = \frac{\exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j)))}{\sum_{j'} \exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q'_j)))}$$

Where α is a simple one layer FFNN

What do these neural models do?



Exact match (EM), Paraphrasing (Para.), Partial clue (Partial), Multiple sentences (Multi), Coreference errors (Coref E), Ambiguous / hard (Hard)

Multiple sentences

P … " we got some groundbreaking performances , here too , tonight , " @entity6 said . " we got @entity17 , who will be doing some musical performances . he 's doing a his - and - her duet all by himself . "…

Q " he 's doing a his - and - her duet all by himself , " @entity6 said of @placeholder

A @entity17

Coreference errors

P … hip - hop star @entity246 saying on @entity247 that he was canceling an upcoming show for the @entity249 …

Q rapper @placeholder " disgusted , " cancels upcoming show for @entity280

@entity280 = @entity249 = SAEs

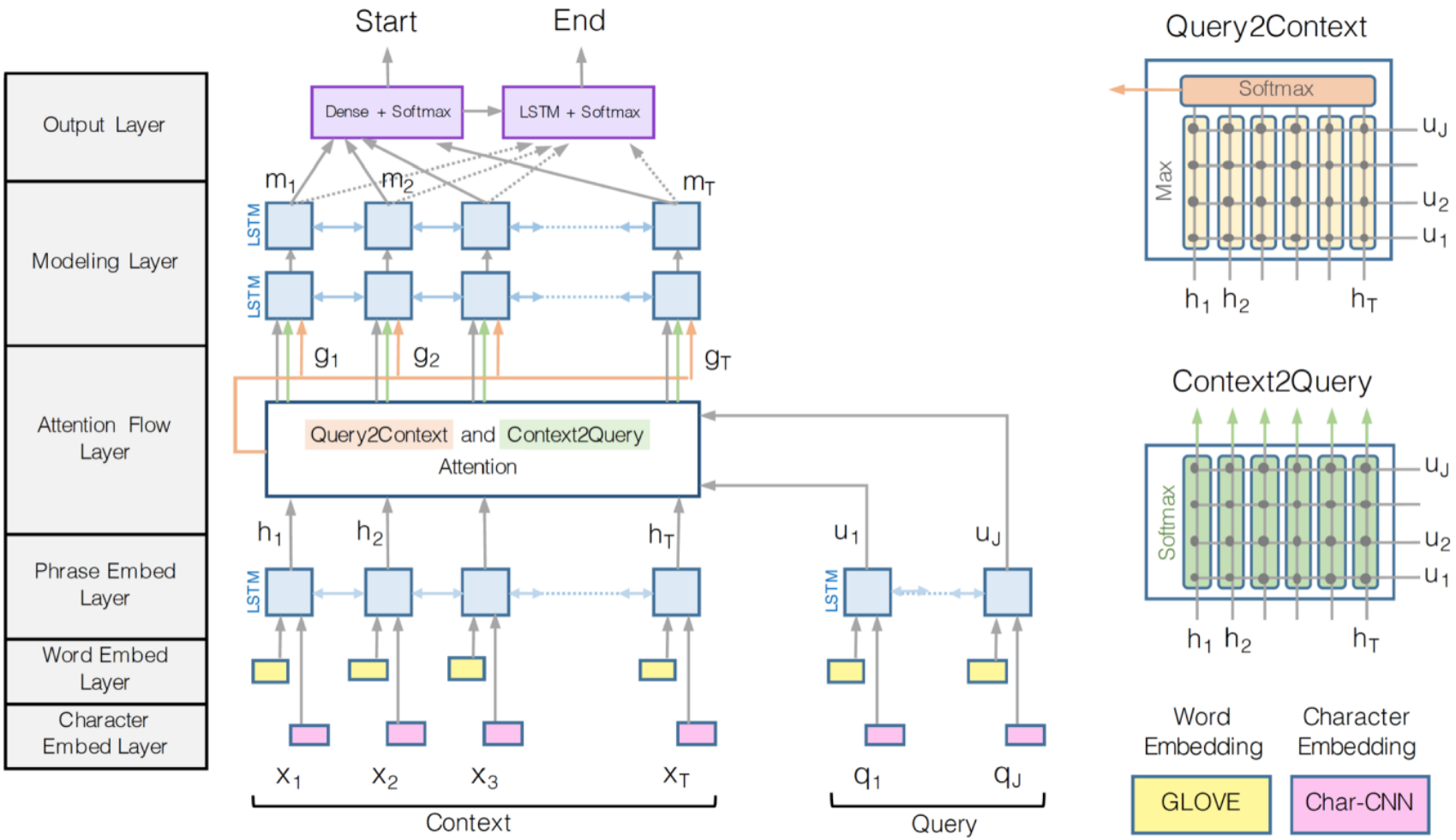
A @entity246

Ambiguous / hard

P … a small aircraft carrying @entity5 , @entity6 and @entity7 " the @entity12 " @entity3 crashed …

Q pilot error and snow were reasons stated for @placeholder plane crash

BiDAF: Bi-Directional Attention Flow for Machine Comprehension



"Bidirectional attention flow for machine comprehension", Seo et al., 2016.
https://arxiv.org/pdf/1611.01603.pdf

BiDAF: Bi-Directional Attention Flow for Machine Comprehension

- There are variants of and improvements to the BiDAF architecture over the years, but the central idea is **the Attention Flow layer**
- **Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Make similarity matrix (with **w** of dimension $6d$):

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^T [\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

- Context-to-Question (C2Q) attention:
  (which query words are most relevant to each context word)

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

BiDAF: Bi-Directional Attention Flow for Machine Comprehension

- **Attention Flow Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Question-to-Context (Q2C) attention:
  (the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max)

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \text{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^{2h}$$

- For each passage position, output of BiDAF layer is:

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h} \quad \forall i \in \{1, \ldots, N\}$$

BiDAF: Bi-Directional Attention Flow for Machine Comprehension

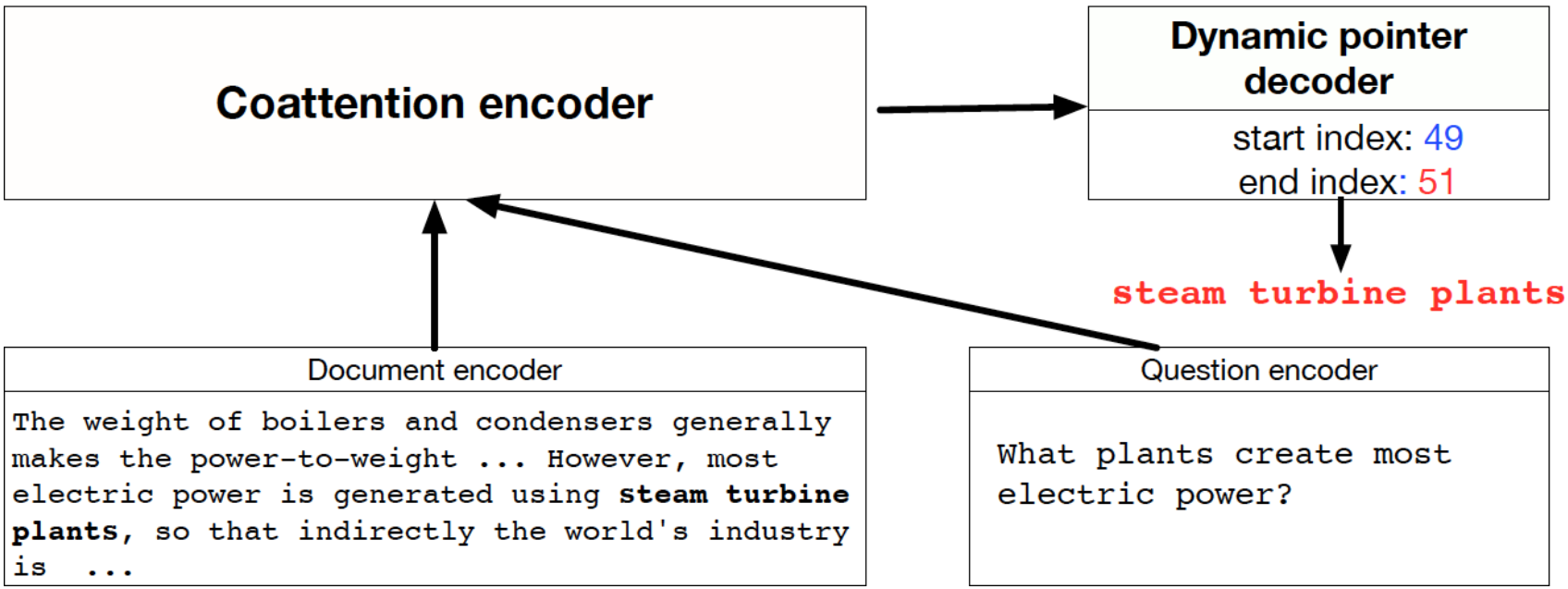
- There is then a "modelling" layer:
  - Another deep (2-layer) BiLSTM over the passage
- And answer span selection is more complex:
  - Start: Pass output of BiDAF and modelling layer concatenated to a dense FF layer and then a softmax
  - End: Put output of modelling layer M through another BiLSTM to give M_2 and then concatenate with BiDAF layer and again put through dense FF layer and a softmax

Recent, more advanced architectures

- Most of the work in 2016, 2017, and 2018 employed progressively more complex architectures with a multitude of variants of attention – often yielding good task gains

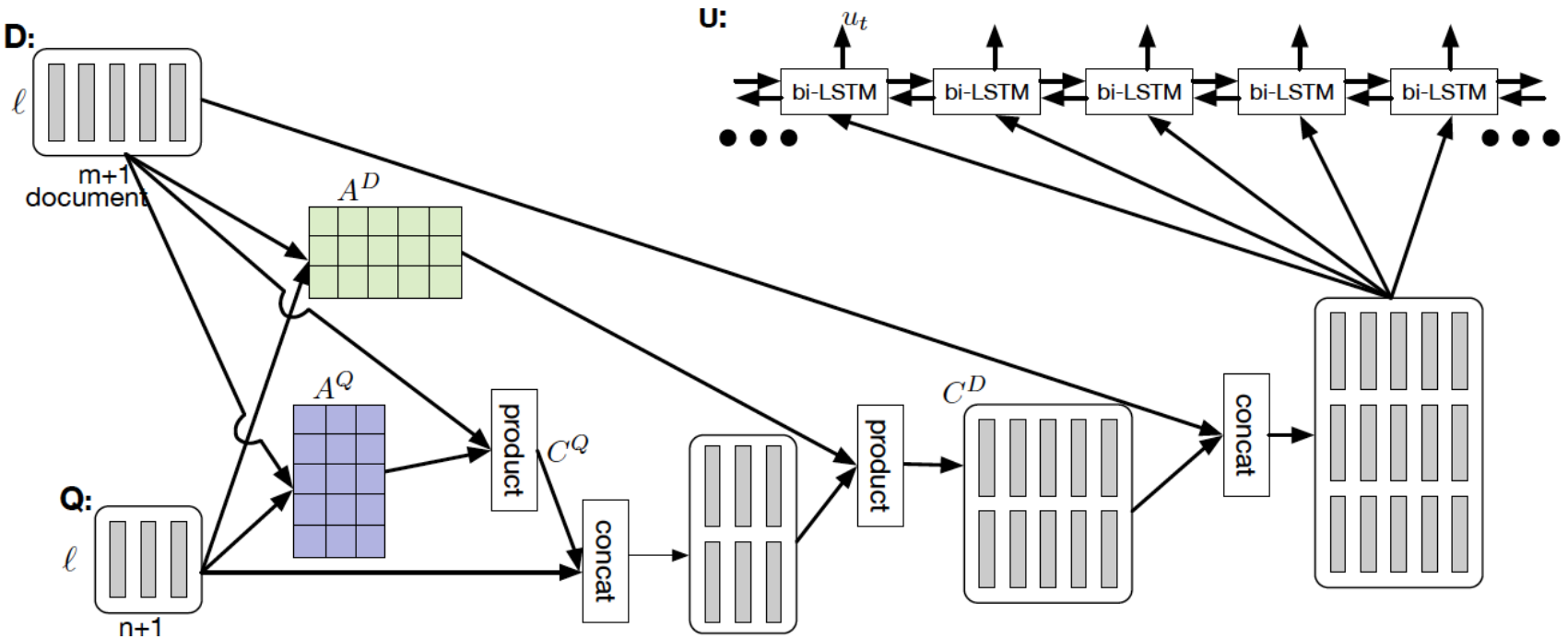
Dynamic Coattention Networks for Question Answering

- Flaw: Questions have input-independent representations
- Interdependence needed for a comprehensive QA model



"Dynamic coattention networks for question answering", Xiong et al., 2016.
https://arxiv.org/pdf/1611.01604.pdf

Coattention Encoder



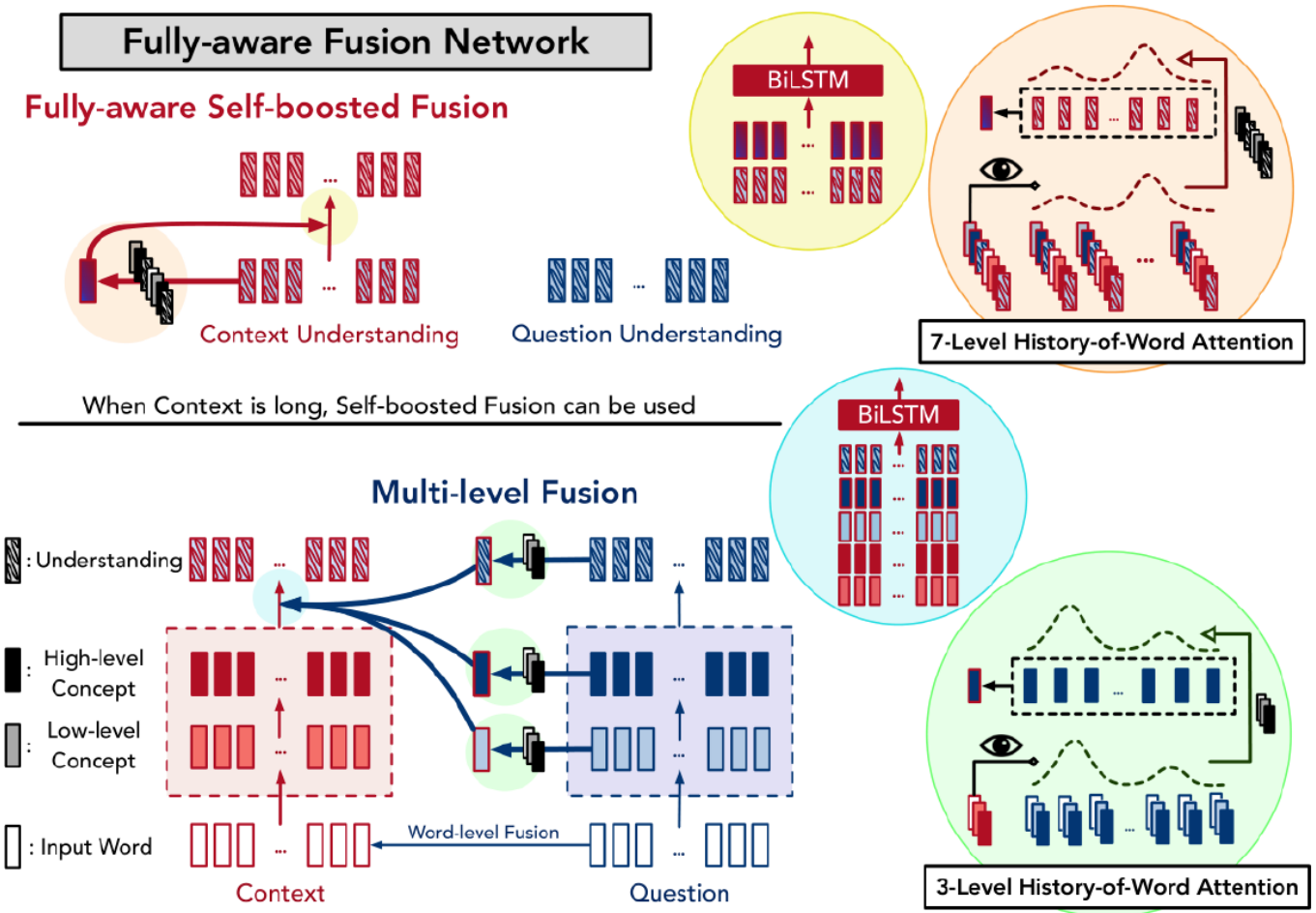
Coattention layer

- Coattention layer again provides a two-way attention between the context and the question
- However, coattention involves a second-level attention computation:
  - attending over representations that are themselves attention outputs
- We use the C2Q attention distributions α_i to take weighted sums of the Q2C attention outputs $\mathbf{b}_j$. This gives us second-level attention outputs $\mathbf{s}_i$:

$$\boldsymbol{s}_i = \sum_{j=1}^{M+1} \alpha_j^i \boldsymbol{b}_j \in \mathbb{R}^l \quad \forall i \in \{1, \ldots, N\}$$

FusionNet

Tries to combine many forms of attention

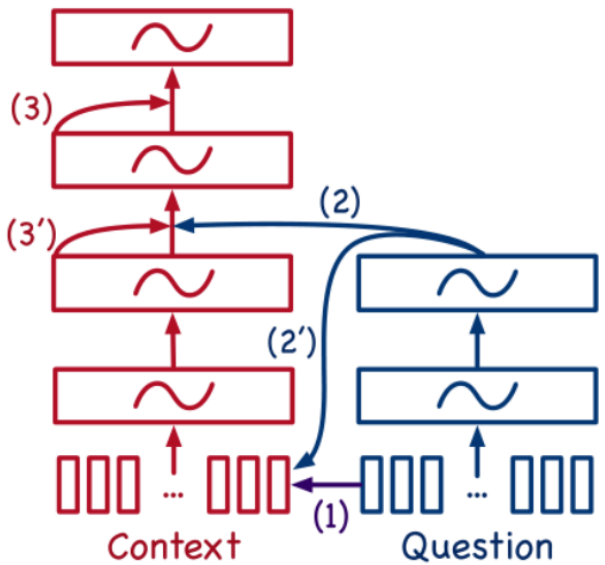


"FusionNet: Fusing via fully-aware attention with application to machine comprehension", Huang et al., 2017. https://arxiv.org/pdf/1711.07341.pdf

Recent, more advanced architectures

- Most of the work in 2016, 2017, and 2018 employed progressively more complex architectures with a multitude of variants of attention – often yielding good task gains

Architectures	(1)	(2)	(2')	(3)	(3')
Match-LSTM (Wang and Jiang, 2016)		✓			
DCN (Xiong et al., 2017)		✓			✓
FastQA (Weissenborn et al., 2017)	✓				
FastQAExt (Weissenborn et al., 2017)	✓	✓		✓	
BiDAF (Seo et al., 2017)		✓			✓
RaSoR (Lee et al., 2016)	✓		✓		
DrQA (Chen et al., 2017)	✓				
MPCM (Wang et al., 2016)	✓	✓			
Mnemonic Reader (Hu et al., 2017)	✓	✓		✓	
R-net (Wang et al., 2017b)		✓		✓	



(1) Word-level fusion, (2) high-level fusion, (2') high-level fusion (alternative), (3) self-boosted fusion, and (3') self-boosted fusion (alternative).