Natural Language Processing

Info 159/259

Lecture 22: Question Answering (April 13, 2021)

David Bamman, UC Berkeley

us congress

All | News | Images | Videos | Maps | More | Settings | Tools

About 731,000,000 results (0.96 seconds)

https://www.congress.gov

Congress.gov | Library of Congress

U.S. Congress legislation, Congressional Record debates, Members of Congress , legislative process educational resources presented by the Library of ...

Members of the US Congress · Most-Viewed Bills · Committees of the US Congress

https://www.house.gov

Homepage | house.gov

Speaker Nancy Pelosi addresses the Members of the 116th **Congress**. ... The Office of the Clerk gives you a front-row seat. **U.S.** Capitol dome at night. Find Your ...

Find Your Representative · Representatives · The House Explained · Committees

People also ask

What is the difference between US Congress and Senate?

How many US Congress are there?

What are the two houses of the US Congress?

Which party is Congress in USA?

Feedback

United States Congress

congress.gov

The United States Congress or U.S. Congress is the bicameral legislature of the federal government of the United States and consists of the House of Representatives and the Senate. The Congress meets in the United States Capitol in Washington, D.C. Wikipedia

Founded: March 4, 1789

Headquarters: Washington, D.C.

Senate last election: November 3, 2020

Senate next election: November 8, 2022

President of the Senate: Kamala Harris (D); since January 20, 2021

Senate president pro tempore: Patrick Leahy (D); since January 20, 2021

Google

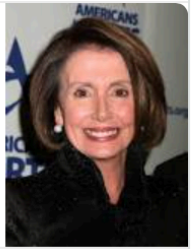
who is the speaker of the house

All | News | Images | Books | Shopping | More | Settings | Tools

About 1,350,000,000 results (0.69 seconds)

United States House of Representatives / Speaker of the House

Nancy Pelosi



People also search for

View 15+ more


Rudy Giuliani
Trending


Donald Trump


Dianne Feinstein
Trending


Adam Schiff


Devin Nunes
Trending


Chuck Schumer


Alexar Ocasi

Feedback

Claim this knowledge panel

8:49

Hey Siri who won the Alaska democratic primary?

Tap to Edit

According to the Associated Press, Joe Biden has won the Alaska Democratic primary with 55.3% of the vote.

2020 ELECTIONS COVERAGE

Alaska Democratic Party Primary

2020 Presidential Primaries Results

CANDIDATE	VOTE%	DELEGATES
Joe Biden	**55.3%**	**11**
Bernie Sanders	44.7%	4
Elizabeth Warren	—	—
Tulsi Gabbard	—	—

100% of precincts reporting

Updated at 07:27PM PDT
Election results from The Associated Press

Watson


THINK
DENKE
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$ 1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

Factoid questions

- Questions that can be answered with simple facts:
  - Who founded Virgin Airlines?
  - What is the average age of the onset of autism?
  - Where is Apple Computer based?
  - ~~How do I get to be a cook at the French Laundry?~~

Question Answering

- Knowledge-based question answering
- IR-based question answering

Knowledge-based QA

- Parse a question into a meaning representation, and execute that meaning representation against a database.

SQL	What record company did conductor Mikhail Snitko record for after 1996?
	SELECT RecordCompany WHERE (YearOfRecording > 1996) AND (Conductor = Mikhail Snitko)

Semantics

Lexical semantics is concerned with representing the meaning of words (and their relations)

Logical semantics is concerned with representing the meaning of sentences.

Meaning representation

- A representation of the meaning of a sentence needs to bridge linguistic aspects of the sentence with non-linguistic knowledge about the world.

Following directions

Linguistic	Non-linguistic
Verbs like *grab*, *release*	Actions a robot can execute
Nouns like *cup*, *block*	Specific entities in the world
transitive VP (V NP)	An action executed upon an object

Meaning representation

Geo	which state has the most rivers running through it?
	(argmax $0 (state:t $0) (count $1 (and (river:t $1) (loc:t $1 $0))))
ATIS	all flights from dallas before 10am
	(lambda $0 e (and (flight $0) (from $0 dallas:ci) (< (departure time $0) 1000:ti)))
SQL	What record company did conductor Mikhail Snitko record for after 1996?
	SELECT Record Company WHERE (Year of Recording > 1996) AND (Conductor = Mikhail Snitko)
Django	if length of bits is lesser than integer 3 or second element of bits is not equal to string 'as'
	if len(bits) < 3 or bits[1] != 'as':

Dong and Lapata (2018), "Coarse-to-Fine Decoding for Neural Semantic Parsing"

Representation of meaning

FIGHT(LUKE, VADER)

This is a *relation*; we define what it means

These are *constants*; we know what they uniquely identify in our database/knowlege base

Representation of meaning

- Relations: likes(x,y) is scoped over two variables
- We can represent the partial representation of meaning with lambda expressions:

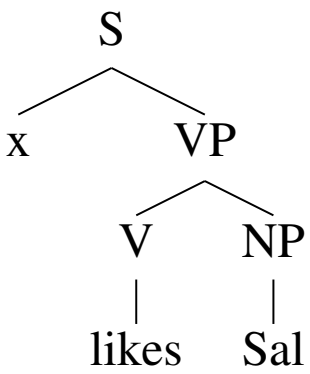
$\lambda x.likes(x,Sal)$

Expect one other argument to complete the meaning of this relation

Representation of meaning

λx.likes(x,Sal)

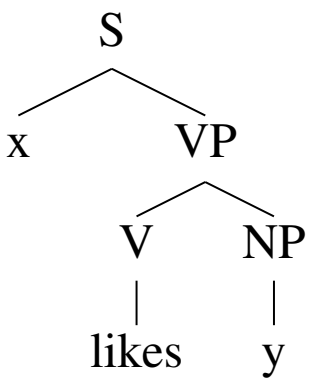
Lambda expressions let us tie semantics explicitly to phrases (subtrees in syntax)



Representation of meaning

λy.λx.likes(x,y)

Lambda expressions let us tie semantics explicitly to phrases (subtrees in syntax)



Learning from logical forms

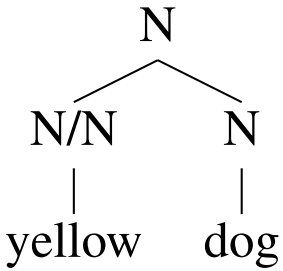
sentence	what states border texas
logical form	λx.state(x) ^ borders(x, texas)

CCG

- Combinatory categorial grammar is a phrase-structure grammar that ties syntactic representation to semantic structure.
- Two components:
    - Combinators that tell us how to compose syntactic types (S, VP, NP, etc.)
    - Mapping between syntactic operations and equivalent semantic operation
    - Lexicon mapping each word in a vocabulary to its syntactic/semantic information.

CCG Combinators

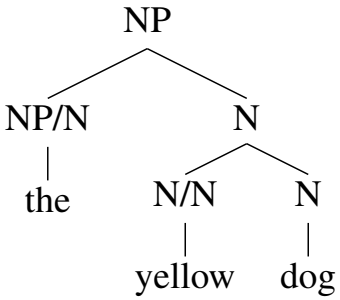
- Forward application combinator (X/Y → X)
- N/N → N



Smith 2017

CCG Combinators

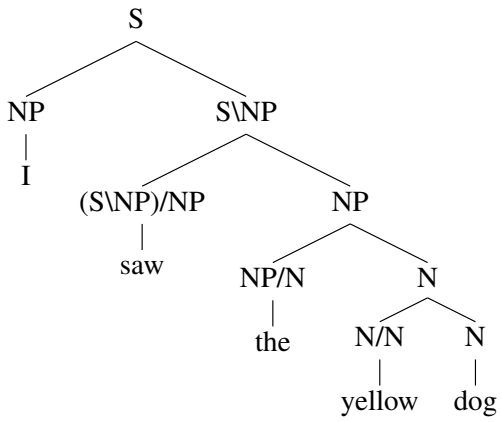
- Forward application combinator (X/Y → X)
  - N/N → N
  - NP/N → NP



Smith 2017

CCG Combinators

- Backward application combinator (X\Y → X)
  - S\NP → S



Smith 2017

CCG Lexicon

Utah	NP	**utah**
Idaho	NP	**idaho**
borders	(S\NP)/NP	**λx.λy.borders(y,x)**
adjoins	(S\NP)/NP	**λx.λy.adjoins(y,x)**
abuts	(S\NP)/NP	**λx.λy.abuts(y,x)**

CCG Combinators

- Each combinator tells us what to do with the corresponding semantics
- Forward application:

X/Y : f	Y : g	→	X f(g)	
(S\NP)/NP : **λx.λy.borders(y,x)**	NP : **idaho**	→	S\NP : **λx.λy.borders(y,x)(idaho)**	S\NP : **λy.borders(y,idaho)**

Smith 2017

Semantics

S
borders(utah,idaho)

S\NP
λy.borders(y,idaho)

NP	(S\NP)/NP	NP
utah	**λx.λy.borders(y,x)**	**idaho**
Utah	borders	Idaho

Semantics

Utah	NP	**utah**
Idaho	NP	**idaho**
borders	(S\NP)/NP	**λx.λy.borders(y,x)**
adjoins	(S\NP)/NP	**λx.λy.adjoins(y,x)**
abuts	(S\NP)/NP	**λx.λy.abuts(y,x)**

Semantics

Utah	NP	**utah**
Idaho	NP	**idaho**
borders	(S\NP)/NP	**λx.λy.borders(y,x)**
adjoins	(S\NP)/NP	**λx.λy.borders(y,x)**
abuts	(S\NP)/NP	**λx.λy.borders(y,x)**

Semantics

- Semantic parsing with CCG is simply syntactic parsing, assuming mapping from syntactic primitives to logical forms.
- But this encounters two problems:
    - We don’t have those manual mappings (task-specific).
    - We can’t parse anything not in our lexicon.

Learning from logical forms

- We can train a semantic parser in a number of ways:
    - Full derivational trees (CCGBank)
    - Logical forms (Zettlemoyer and Collins 2005)
    - Denotations (Berant et al. 2013)

Learning from trees

S
λx.state(x) ^ borders(x,texas)

(S/S\NP)
λg.λx.state(x) ^ g(x)

(S\NP)
λy.borders(y,texas)

(S/S\NP)/N	N	(S\NP)/NP	NP
λf.λg.λx.f(x) ^ g(x)	**λx.state(x)**	**λx.λy.borders(y,x)**	**texas**
what	states	border	texas

Learning from logical forms

sentence	what states border texas
logical form	λx.state(x) ^ borders(x, texas)

Two core ideas:

- We'll learn the lexicon (including the lambda expressions)
- We'll learn CCG parser from that lexicon, and treat the true tree as a latent variable

Learning from logical forms

- Calculate the joint probability of a logical form L and derivation T for sentence S as:

$$P(L, T \mid S; \theta) = \frac{\exp(f(L, T, S)^\top \theta)}{\sum_{L,T} \exp(f(L, T, S)^\top \theta)}$$

sums over all valid trees/logical forms for the sentence

feature
Utah := NP : utah
Utah := NP : idaho
borders := (S\NP)/NP : λx.λy.borders(y,x)
borders := (S\NP)/NP : λx.λy.borders(x,y)

f(L,T,S)

Learning from denotations

sentence	what states border texas
logical form	λx.state(x) ^ borders(x, texas)
denotation	**new_mexico, oklahoma, arkansas, louisiana**

sentence	number of dramas starring tom cruise
logical form	count(λx.genre(x,drama) ^ ∃y.performance(x,y) ^ actor(y,tom_cruise))
denotation	**28**

Learning from denotations

sentence	what states border texas
logical form	λx.state(x) ^ borders(x, texas)
denotation	**new_mexico, oklahoma, arkansas, louisiana**

sentence	number of dramas starring tom cruise
logical form	count(λx.genre(x,drama) ^ ∃y.performance(x,y) ^ actor(y,tom_cruise))
denotation	**28**

Learning from denotations

- How could we use the principles of learning from logical forms to learn from denotations?
- The meaning of a sentence is the set of possible worlds consistent with that statement.

Utah borders Idaho	TRUE
number of dramas starring tom cruise	28

Learning from denotations

- Basic idea: maximize the probability of the tree T/logical form z that, when executed against a knowledge base $\mathscr{K}$, yield the correct denotation y

$$\sum_{i=1}^{N} \log \sum_{T:[\![T.z]\!]_{\mathcal{K}}=y_i} P(T \mid S_i, \theta)$$

objective function

IR-based QA

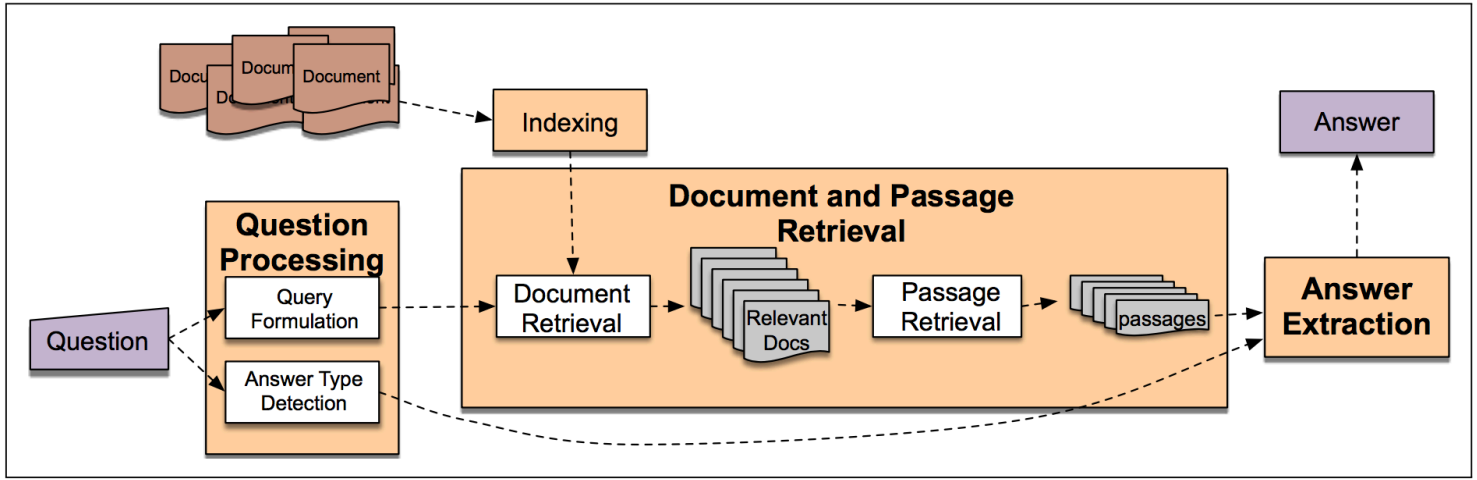


Figure 25.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

Answer type detection

- If we know the type of answer expected for a question, we can exclude answers that aren't of that type (resulting in higher precision)

Who founded Virgin Airlines?	→	PERSON
What Canadian city has the largest population?	→	CITY

Answer type detection

ENTITY	
animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma" in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

Li and Roth (2002); SLP3

Answer type detection

- Train an answer type detection system with supervised learning:
  - Words/embeddings
  - POS for each word
  - Question headword
    - Which city in China has the largest number of foreign financial companies?
    - What is the state flower of California?

Document retrieval

- Rather than running answer extraction on all documents, let's just focus on the subset that are likely to be relevant.
- This is a problem of information retrieval.

IR

- Represent query and each document by tf-idf weighted vector, rank by cosine similarity

$$cos(x, y) = \frac{\sum_{i=1}^{F} x_i y_i}{\sqrt{\sum_{i=1}^{F} x_i^2}\sqrt{\sum_{i=1}^{F} y_i^2}}$$

term	doc 1	doc 2	doc 3
alaska			4.5
primary			7.8
dog		4.8	3.4
friday			
nlp	6.4		
the	0.05	0.03	0.02
dem			8.1
tree	2.3	9.1	

Document retrieval

- Answer extraction typically operates on passages (not entire documents), so we can rank the passages as well using information about:
    - Number of entities of correct type
    - Number of question keywords
    - Longest exact sequence of keywords in passage
    - Rank of document
    - Ngram overlap

Answer extraction

- From a given passage, extract the answer.

Input passage

Talking Heads were an American rock band formed in 1975 in New York City and active until 1991.[8] The band comprised David Byrne (lead vocals, guitar), Chris Frantz (drums), Tina Weymouth (bass), and Jerry Harrison (keyboards, guitar). Described by the critic Stephen Thomas Erlewine as "one of the most critically acclaimed bands of the '80s,"[3] the group helped to pioneer new wave music by integrating elements of punk, art rock, funk, and world music with avant-garde sensibilities and an anxious, clean-cut image.[3]

Question

Where did the Talking Heads originate?

Datasets

Dataset	Size	Source	Notes
SQUAD	150K	Wikipedia	
NewsQA	100K	CNN	
HotpotQA	113K	Wikipedia	Requires reasoning over multiple documents
NarrativeQA	46K	Books/movie scripts	Requires reasoning over long documents, answer not a span

HotpotQA

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

Yang et al. (2018), "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering"

NarrativeQA

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.

Kočisky et al. (2017), "The NarrativeQA Reading Comprehension Challenge"

SQUAD

- 100,000 question-answer pairs created by crowdworkers.
- Task: given an input passage from Wikipedia, create a question whose answer is a span of text within it.

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

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

Rajpurkar et al. (2016), "SQuAD: 100,000+ Questions for Machine Comprehension of Text"

SQUAD2.0

- 100K questions from SQUAD + 50K unanswerable questions.
- To perform well, models must know when to abstain from answering a question.

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Rajpurkar et al. (2018), "Know What You Don't Know: Unanswerable Questions for SQuAD"

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Rajpurkar et al. (2016), "SQuAD: 100,000+ Questions for Machine Comprehension of Text"

Evaluation

- Exact match: proportion of questions for which the predicted answer is an exact match for the correct answer.
- F1: Calculate F1 between predicted answer and gold answer (both bag of tokens).

SQUAD2.0

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
4 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978

Input passage

Talking Heads were an American rock band formed in 1975 in New York City and active until 1991.[8] The band comprised David Byrne (lead vocals, guitar), Chris Frantz (drums), Tina Weymouth (bass), and Jerry Harrison (keyboards, guitar). Described by the critic Stephen Thomas Erlewine as "one of the most critically acclaimed bands of the '80s,"[3] the group helped to pioneer new wave music by integrating elements of punk, art rock, funk, and world music with avant-garde sensibilities and an anxious, clean-cut image.[3]

Question

Where did the Talking Heads originate?

- What is the start token position and end token position for the answer span in the input passage?

Featurized

- Answer type match: does the candidate answer contain a phrase with the answer type expected?

- Pattern match:

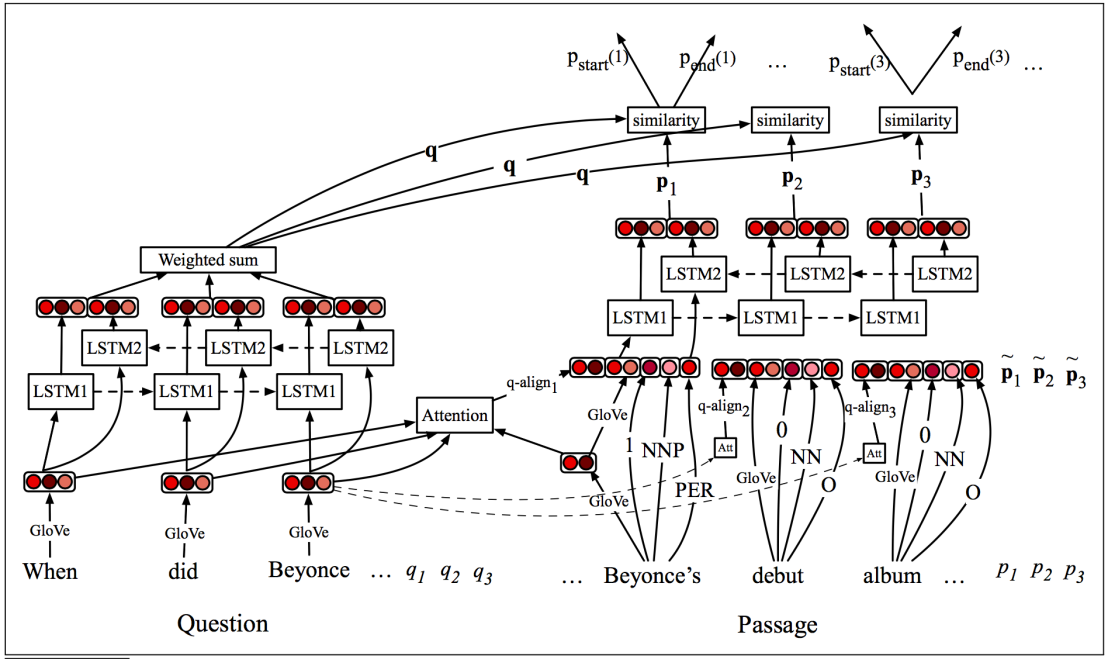
Pattern	Question	Answer
<AP> such as <QP>	What is autism?	“, developmental disorders such as autism”
<QP>, a <AP>	What is a caldera?	“the Long Valley caldera, a volcanic crater 19 miles long”

Figure 25.5 Some answer-extraction patterns using the answer phrase (AP) and question phrase (QP) for definition questions (Pasca, 2003).

Featurized

- Number of matched question keywords
- Keyword distance: distance between answer and query keywords
- Novelty factor: does the passage contain words that are not in the query?
- Length of longest sequence of question terms that occur in the candidate answer
- Apposition features: answer is in appositive construction to phrase containing several question words.
    - Question: What is a caldera?
    - Candidate: “the Long Valley caldera, a volcanic crater 19 miles long”

Neural



Chen et al. 2017; SLP3

Input passage

Talking Heads were an American rock band formed in 1975 in New York City and active until 1991.[8] The band comprised David Byrne (lead vocals, guitar), Chris Frantz (drums), Tina Weymouth (bass), and Jerry Harrison (keyboards, guitar). Described by the critic Stephen Thomas Erlewine as "one of the most critically acclaimed bands of the '80s,"[3] the group helped to pioneer new wave music by integrating elements of punk, art rock, funk, and world music with avant-garde sensibilities and an anxious, clean-cut image.[3]

Question

Where did the Talking Heads originate?

- What is the start token position and end token position for the answer span in the input passage?

Input passage

Talking Heads were an American rock band formed in 1975 in New York City and active until 1991.

Question

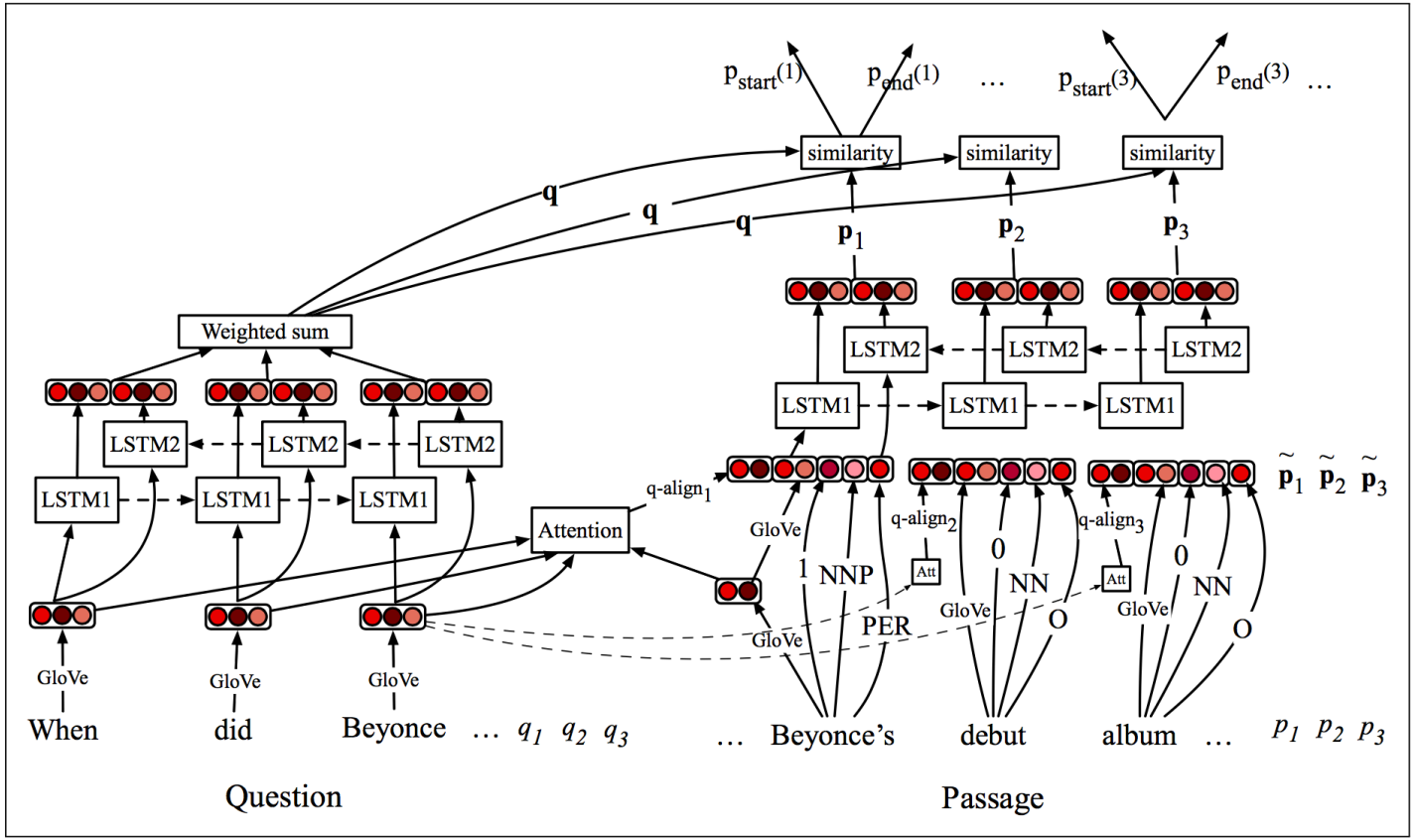
Where did the Talking Heads originate?

- What is the start token position and end token position for the answer span in the input passage?

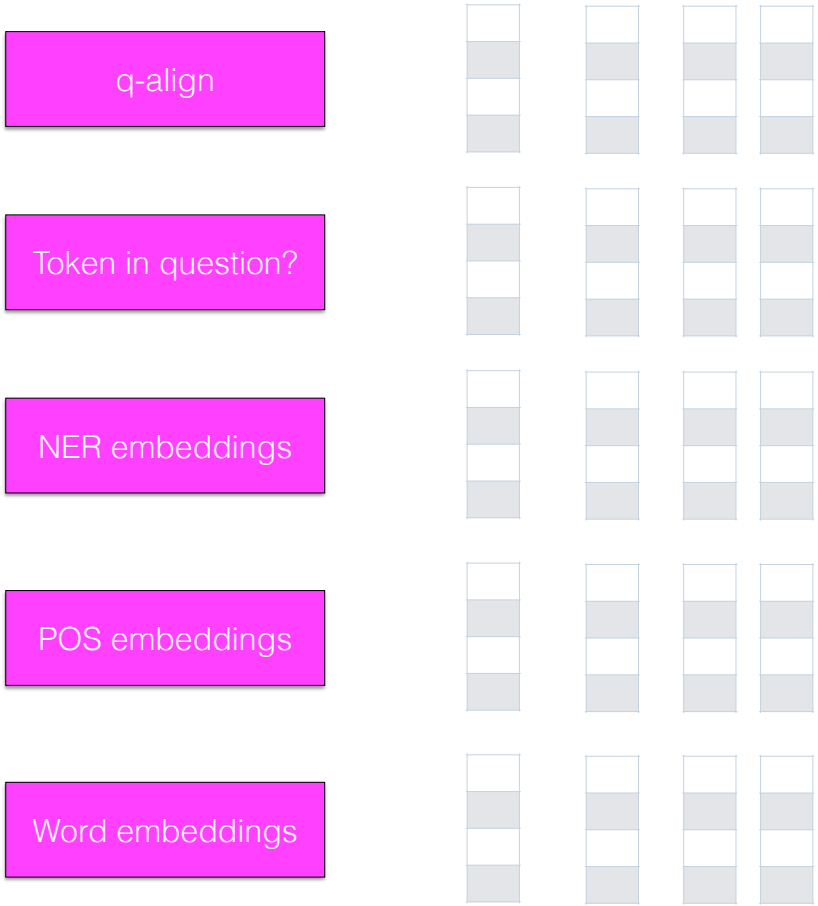
$p_{start}(i)$	0	0	0	0	0.10	0	0	0	0
$p_{end}(i)$	0	0	0	0	0.09	0	0	0	0
Text	Talking	Heads	were	an	American	rock	band	formed	in

$p_{start}(i)$	0	0	0.85	0	0.05	0	0	0	0
$p_{end}(i)$	0	0	0	0.16	0.75	0	0	0	0
Text	1975	in	New	York	City	and	active	until	1991

- Each token position i has two associated probabilities:
  - $p_{start}(i)$ — the probability of the answer starting at that token.
  - $p_{end}(i)$ — the probability of the answer ending at that token.



Chen et al. 2017; SLP3

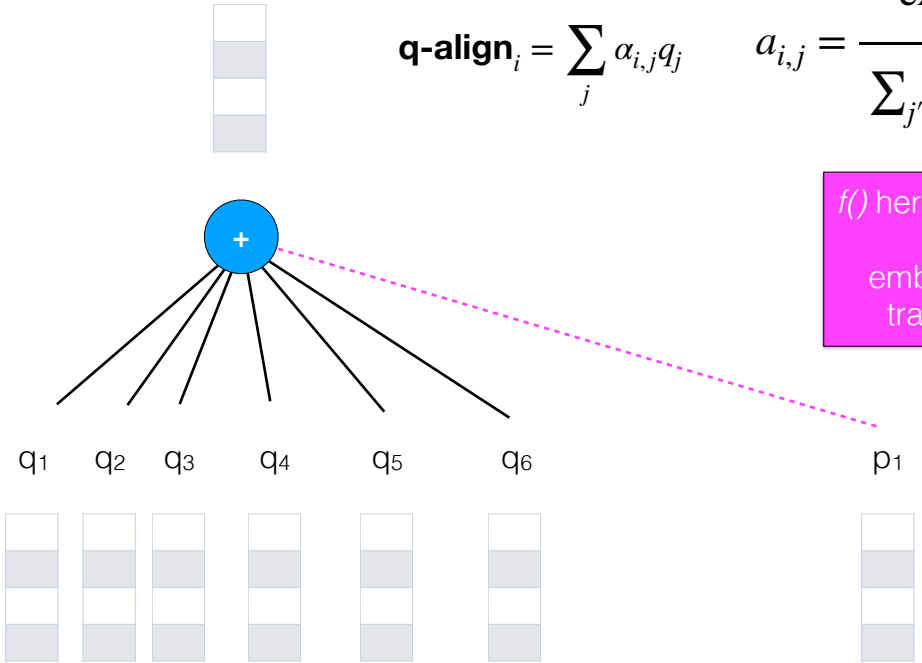


Talking Heads were an …

Each passage token attends over all question tokens

$$\textbf{q-align}_i = \sum_j \alpha_{i,j} q_j \qquad a_{i,j} = \frac{\exp\left(f(p_i)^\top f(q_j)\right)}{\sum_{j'} \exp\left(f(p_i)^\top f(q_j')\right)}$$

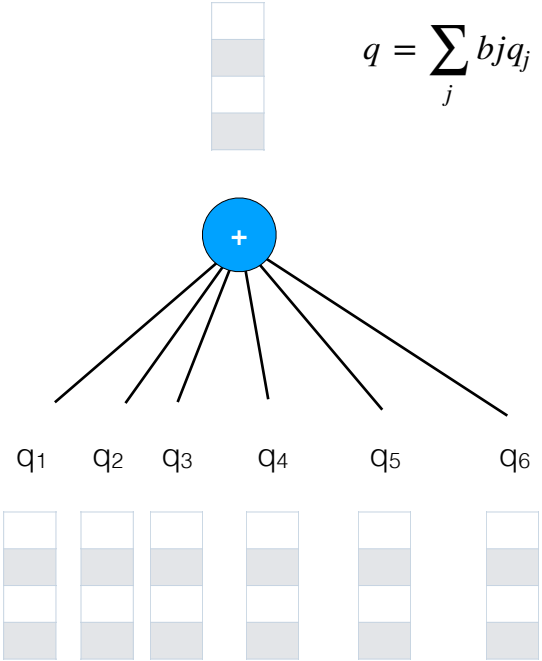
f() here is some differentiable function over the embeddings (e.g., linear transformation, FFNN)



Where did the Talking Heads originate?

Talking Heads were an …

Single representation (using attention) for the entire question



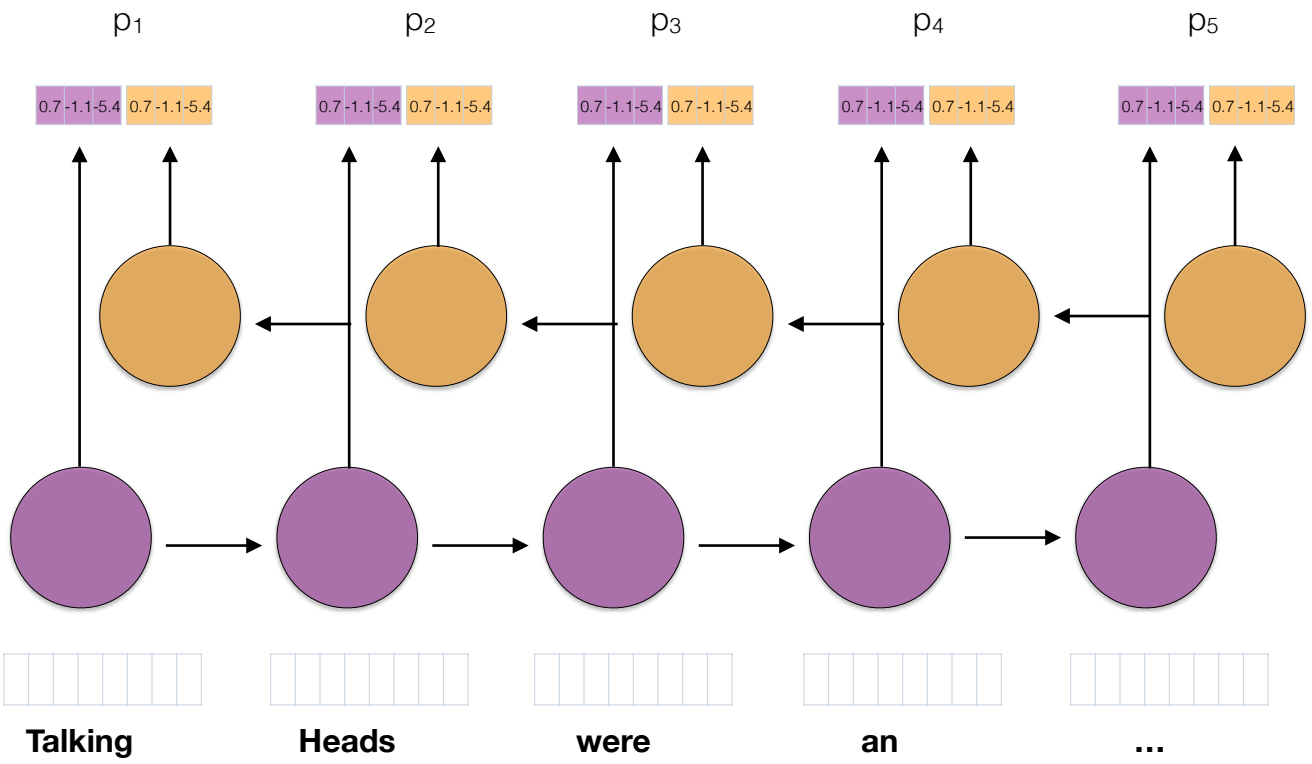
$$q = \sum_j bjq_j$$

$$b_j = \frac{\exp\left(w^\top q_j\right)}{\sum_j \exp\left(w^\top q_j\right)}$$

Where did the Talking Heads originate?

$$p_{\textbf{start}}(i) \propto \exp\left(p_i W_s q\right)$$

$$p_{\textbf{end}}(i) \propto \exp\left(p_i W_e q\right)$$



SQUAD2.0

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
4 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978

Question Answering

		Evaluated on					
		SQuAD	TriviaQA	NQ	QuAC	NewsQA	Avg Δ
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1	-17.2
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5	-29.9
	NQ	53.5	46.3	**73.5**	21.6	24.7	-20.4
	QuAC	39.4	33.1	33.8	**33.3**	13.8	-36.9
	NewsQA	52.1	38.4	41.7	20.4	**60.1**	-22.2

Table 4: F1 scores of each fine-tuned model evaluated on each test set

Sen and Safari (2020), “What do Models Learn from Question Answering Datasets?”