CS388: Natural Language Processing

Lecture 22: Question Answering 1

Greg Durrett

TEXAS
The University of Texas at Austin

Administrivia

- Jason Baldridge guest lecture next Thursday
- P2 back soon

Recall: Erasure Methods

- This secondary classifier's **weights** now give us highlights on the input

The movie is mediocre, maybe even bad.	Negative 99.8%
The movie is mediocre, maybe even ~~bad~~.	Negative 98.0%
The movie is ~~mediocre~~, maybe even bad.	Negative 98.7%
The movie is ~~mediocre~~, maybe even ~~bad~~.	Positive 63.4%
The movie is ~~mediocre~~, ~~maybe~~ even ~~bad~~.	Positive 74.5%
The ~~movie~~ is mediocre, maybe even ~~bad~~.	Negative 97.9%

The movie is mediocre, maybe even bad.

Wallace, Gardner, Singh
Interpretability Tutorial at EMNLP 2020

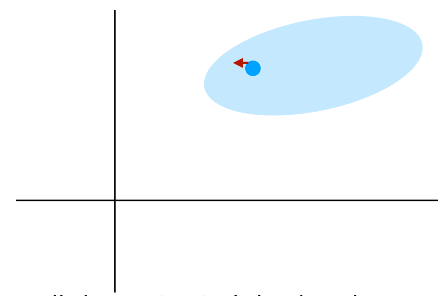
Recall: Gradient-based Methods

- Originally used for images

S_c = score of class c

I_0 = current image

$$w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$



- Higher gradient magnitude = small change in pixels leads to large change in prediction
- For words: "pixels" are coordinates of each word's vector, sum these up to get the importance of that word

Simonyan et al. (2013)

This Lecture

- Types of question answering/reading comprehension
- Span-based question answering on SQuAD
- SQuAD results

Reading Comprehension

Classical Question Answering

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

Q: *where was Barack Obama born*

```
λx. type(x, Location) ∧ born_in(Barack_Obama, x)
```

(also Prolog / GeoQuery, etc.)

- How to deal with open-domain data/relations? Need data to learn how to ground every predicate or need to be able to produce predicates in a zero-shot way

QA from Open IE

(a) **CCG parse** builds an underspecified semantic representation of the sentence.

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

$$N: \lambda x.former(x) \wedge municipalities(x) \qquad N\backslash N: \lambda f \lambda y.f(y) \wedge in(y, Brandenburg)$$

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

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

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

- Why use the KB at all? Why not answer questions directly from text?

Choi et al. (2015)

QA is very broad

- Factoid QA: *what states border Mississippi?, when was Barack Obama born?*
  - Lots of this could be handled by QA from a knowledge base, if we had a big enough knowledge base
- "Question answering" as a term is so broad as to be meaningless
  - *What is the meaning of life?*
  - *What is 4+5?*
  - *What is the translation of [sentence] into French?* [McCann et al., 2018]

What are the limits of QA?

- Focus on questions where the answer might appear in text — still hard!
  - *What were the main causes of World War II?* — requires summarization
  - *Can you get the flu from a flu shot?* — want IR to provide an explanation of the answer, not just yes/no
  - *How long should I soak dry pinto beans?* — could be written down in a KB but probably isn't
  - Today: QA when it requires retrieving the answer from a passage

Reading Comprehension

- "AI challenge problem": answer question given context
- Recognizing Textual Entailment (2006)
- MCTest (2013): 500 passages, 4 questions per passage
- Two questions per passage explicitly require cross-sentence reasoning

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

Richardson (2013)

Baselines

- N-gram matching: append question + each answer, return answer which gives highest n-gram overlap with a sentence

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

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

What did James pull off of the shelves in the grocery store? Pudding

rephrased: James pulled pudding off of the shelves in the grocery store

Richardson (2013)

Baselines

- N-gram matching: append question + each answer, return answer which gives highest n-gram overlap with a sentence
- Parsing: find direct object of "pulled" in the document where the subject is James

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

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

Richardson (2013)

Reading Comprehension

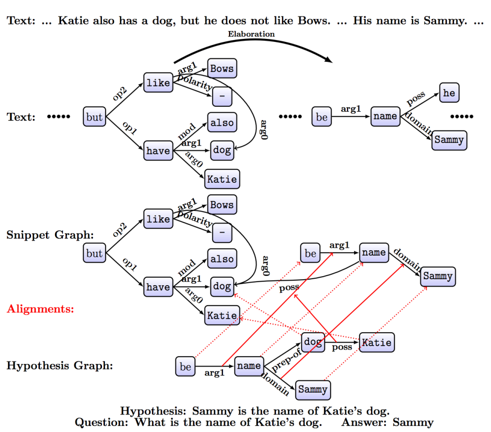
ngram sliding window

	MC160 Test	MC500 Test
Baseline (SW+D)	66.25	56.67
RTE	59.79‡	53.52
Combined	67.60	60.83‡

- Classic textual entailment systems don't work as well as n-grams
- Scores are low partially due to questions spanning multiple sentences
- Unfortunately not much data to train better methods on (2000 questions)

Richardson (2013)

Better Systems



- Match an AMR (abstract meaning representation) of the question against the original text
- 70% accuracy (roughly 10% better than baseline)

Sachan and Xing (2016)

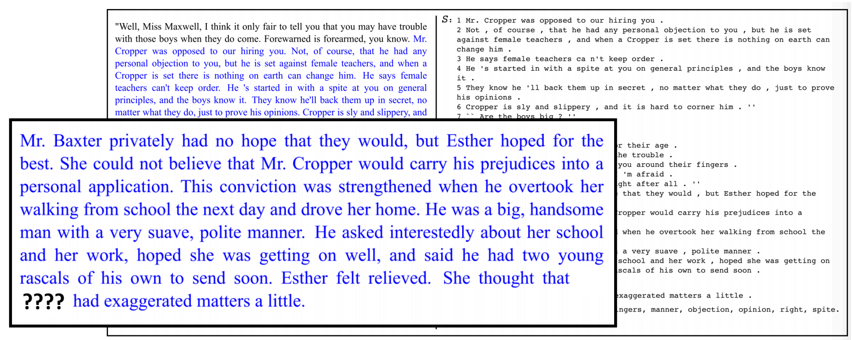
Dataset Explosion

- 30+ QA datasets released since 2015
- Question answering: questions are in natural language
  - Answers: multiple choice, require picking from the passage, or generate freeform answer (last is pretty rare)
  - Require human annotation
- "Cloze" task: word (often an entity) is removed from a sentence
  - Answers: multiple choice, pick from passage, or pick from vocabulary
  - Can be created automatically from things that aren't questions

Dataset Properties

- Axis 1: cloze task (fill in blank) vs. multiple choice vs. span-based vs. freeform generation
- Axis 2: what's the input?
  - One paragraph? One document? All of Wikipedia?
  - Some explicitly require linking between multiple sentences (MCTest, WikiHop, HotpotQA)
- Axis 3: what capabilities are needed to answer questions?
  - Finding simple information? Combining information across multiple sources? Commonsense knowledge?

Children's Book Test



- Children's Book Test: take a section of a children's story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

LAMBADA

Context: They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I 've always loved ____.
Target word: dancing

- GPT/BERT can in general do very well at cloze tasks because this is what they're trained to do
- Hard to come up with plausible alternatives: "cooking", "drawing", "soccer", etc. don't work in the above context

Paperno et al. (2016)

Multiple-Choice

- SWAG dataset was constructed to be difficult for ELMo
- BERT subsequently got 20+% accuracy improvements and achieved human-level performance
- Problem: distractors too easy

The person blows the leaves from a grass area using the blower. The blower...

a) puts the trimming product over her face in another section.
b) is seen up close with different attachments and settings featured.
c) continues to blow mulch all over the yard several times.
d) blows beside them on the grass.

- Let's focus on architectures for **retrieval from a passage**

Zellers et al. (2018)

Span-based Question Answering

SQuAD

- Single-document question-answering task where the answer is always a substring of the passage (= a paragraph from Wikipedia)
- Predict start and end indices of the answer in the passage

One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

What was Maria Curie the first female recipient of?
Ground Truth Answers: Nobel Prize Nobel Prize Nobel Prize

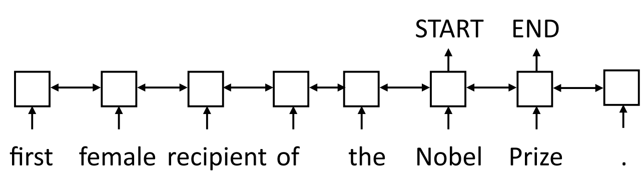
What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 1745 1745

Who was one of the most famous people born in Warsaw?
Ground Truth Answers: Maria Skłodowska-Curie Maria Skłodowska-Curie Maria Skłodowska-Curie

Rajpurkar et al. (2016)

SQuAD

What was Marie Curie the first female recipient of?



- Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query

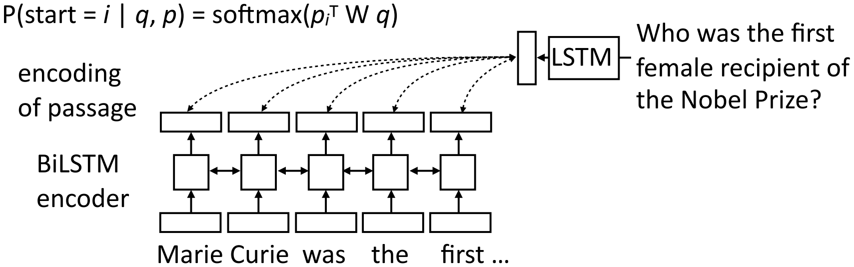
Rajpurkar et al. (2016)

Architectures

- Predict a distributions over start and end points of the answer

P(end | q, p) computed similarly

P(start = i | q, p) = softmax(p_i^T W q)



Training and Inference

- Train on labeled data with start and end points, maximize likelihood of correct decisions: $\log \sum_{i \in \text{gold starts}} p(\text{start} = i | p, q) + \log \sum_{i \in \text{gold ends}} p(\text{end} = i | p, q)$

In September 1958, Bank of America launched a new product called BankAmericard in Fresno. After a troubled gestation during which its creator resigned, BankAmericard went on to become the first successful credit card; that is, a financial instrument that was usable across a large number of merchants and also allowed cardholders to revolve a balance (earlier financial products could do one or the other but not both). In 1976, BankAmericard was renamed and spun off into a separate company known today as Visa Inc.

What was the name of the first successful credit card?

- Inference: maximize P(start) + P(end) with the constraint that (start, end) isn't too big a span

What do these models do?

Question: who caught a 16-yard pass on this drive ?
Answer: devin funchess

START

there would be no more scoring in the third quarter , but early in the fourth , the broncos drove to the panthers 41-yard line . on the next play , ealy knocked the ball out of manning 's hand as he was winding up for a pass , and then recovered it for carolina on the 50-yard line . a 16-yard reception by devin funchess and a 12-yard run by stewart then set up gano 's 39-yard field goal , cutting the panthers deficit to one score at 16â€“10 . the next three drives of the game would end in punts .

END

there would be no more scoring in the third quarter , but early in the fourth , the broncos drove to the panthers 41-yard line . on the next play , ealy knocked the ball out of manning 's hand as he was winding up for a pass , and then recovered it for carolina on the 50-yard line . a 16-yard reception by devin funchess and a 12-yard run by stewart then set up gano 's 39-yard field goal , cutting the panthers deficit to one score at 16â€“10 . the next three drives of the game would end in punts .

What do these models do?

Question: how many victorians are non - religious ?
Answer: 20 %

START

about 61.1 % of victorians describe themselves as christian . roman catholics form the single largest religious group in the state with 26.7 % of the victorian population , followed by anglicans and members of the uniting church . buddhism is the state 's largest non - christian religion , with 168,637 members as of the most recent census . victoria is also home of 152,775 muslims and 45,150 jews . hinduism is the fastest growing religion . around 20 % of victorians claim no religion . amongst those who declare a religious affiliation , church attendance is low .

END

about 61.1 % of victorians describe themselves as christian . roman catholics form the single largest religious group in the state with 26.7 % of the victorian population , followed by anglicans and members of the uniting church . buddhism is the state 's largest non - christian religion , with 168,637 members as of the most recent census . victoria is also home of 152,775 muslims and 45,150 jews . hinduism is the fastest growing religion . around 20 % of victorians claim no religion . amongst those who declare a religious affiliation , church attendance is low .

Why did this take off?

- SQuAD was **big**: >100,000 questions at a time when deep learning was exploding
- SQuAD was **pretty easy**: year-over-year progress for a few years until the dataset was essentially solved
- SQuAD had **room to improve**: ~50% performance from a logistic regression baseline (classifier with 180M features over constituents)

Bidirectional Attention Flow

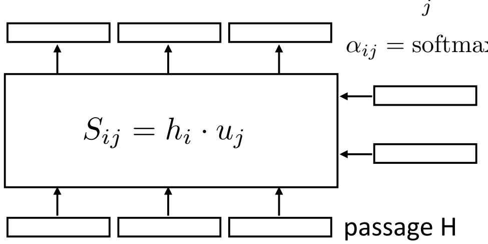
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

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

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

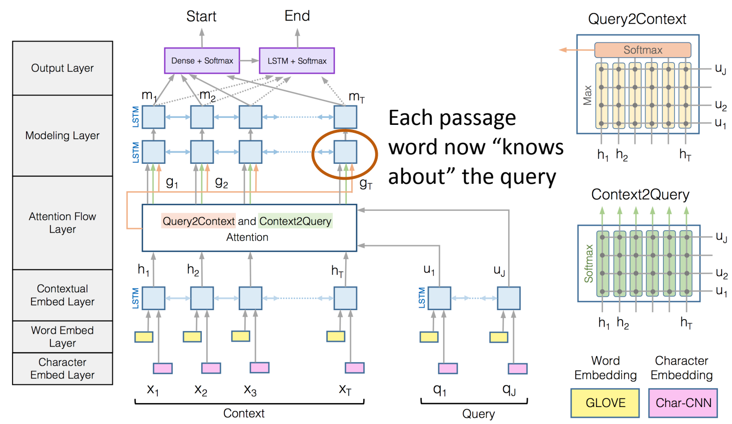
$$\alpha_{ij} = \text{softmax}_j(S_{ij})$$

- dist over query



Seo et al. (2016)

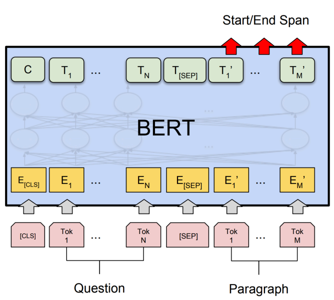
Bidirectional Attention Flow



Each passage word now "knows about" the query

Seo et al. (2016)

QA with BERT



What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of …

- Predict start and end positions in passage
- No need for cross-attention mechanisms!

Devlin et al. (2019)

QA with BERT

- How does this work?

What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of the Nobel Prize

Devlin et al. (2019)

SQuAD Results

SQuAD SOTA: Fall 18

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

- BiDAF: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF…)

SQuAD SOTA: Spring 19

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886
5 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
6 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
7 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615

- SQuAD 2.0: harder dataset because some questions are unanswerable
- Industry contest

SQuAD SOTA: Fall 19

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071

- Performance is very saturated
- Harder QA settings are needed!

TriviaQA

- Totally figuring this out is very challenging
- Coref:
  the failed campaign
  movie of the same name
- Lots of surface clues: 1961, campaign, etc.
- Systems can do well without really understanding the text

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Joshi et al. (2017)

What are these models learning?

- "Who…": knows to look for people
- "Which film…": can identify movies and then spot keywords that are related to the question
- Unless questions are made super tricky (target closely-related entities who are easily confused), they're usually not so hard to answer

What are these models learning?

(Answer = Stanford University)

Question: Where did the Broncos practice for the Super Bowl ?
Passage: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott . The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott .

(d) Erasure exact search optima.

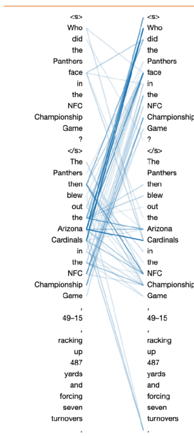
Question: Where did the Broncos practice for the Super Bowl ?
Passage: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott . The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott .

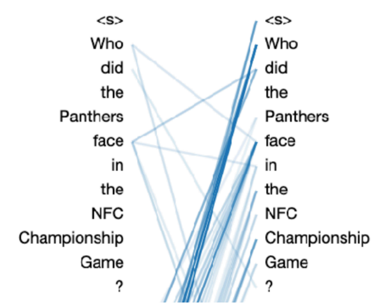
(a) Integrated Gradient (Sundararajan et al., 2017).

- Are these good explanations?

De Cao et al. (2020)

What are these models learning?

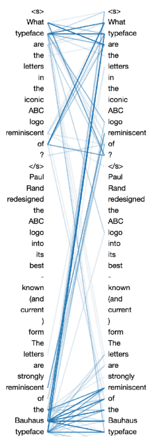


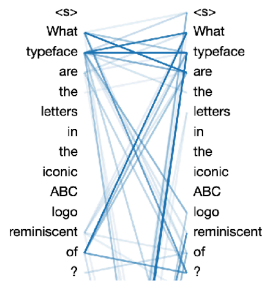


Pairwise explanation method: explains predictions in terms of associations between words

Ye, Nair, Durrett (2021)

What are these models learning?





ABC isn't used at all! The model is mostly using the fact that only one typeface is in the context

Ye, Nair, Durrett (2021)

Takeaways

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