CS388: Natural Language Processing

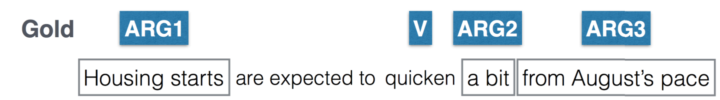
Lecture 21: Question Answering 1

Greg Durrett

TEXAS
The University of Texas at Austin

Recall: SRL

- Identify predicate, disambiguate it, identify that predicate's arguments
- Verb roles from Propbank (Palmer et al., 2005)



quicken:
Arg0-PAG: *causer of speed-up*
Arg1-PPT: *thing becoming faster* (vnrole: 45.4-patient)
Arg2-EXT: *EXT*
Arg3-DIR: *old speed*
Arg4-PRD: *new speed*

Figure from He et al. (2017)

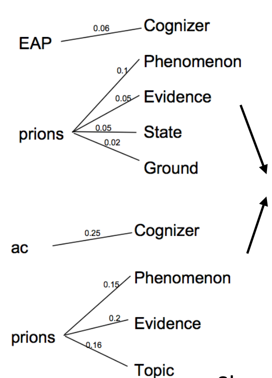
Recall: SRL for QA

- Question and several answer candidates

Q: *Who discovered prions?*

AC1: *In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions…*

AC2: *Prions were researched by…*



Score by matching expected answer phrase (EAP) against answer candidate (AC)

Shen and Lapata (2007)

This Lecture

- Types of question answering/reading comprehension
- Memory networks
- CNN/Daily Mail task: Attentive Reader
- SQuAD task: Bidirectional Attention Flow

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: $\lambda x. former(x) \wedge municipalities(x)$ $>$

$N\backslash N$: $\lambda f \lambda y. f(y) \wedge in(y, Brandenburg)$ $>$

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

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

$\mathbf{l}_0 = \lambda x. former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$

$\mathbf{l}_1 = \lambda x. former(x) \wedge municipalities(x) \wedge in(x, \mathtt{Brandenburg})$

$\mathbf{l}_2 = \lambda x. former(x) \wedge municipalities(x) \wedge \mathtt{location.containedby}(x, \mathtt{Brandenburg})$

$\mathbf{l}_3 = \lambda x. former(x) \wedge \mathtt{OpenRel}(x, \mathtt{Municipality}) \wedge \mathtt{location.containedby}(x, \mathtt{Brandenburg})$

$\mathbf{l}_4 = \lambda x. \mathtt{OpenType}(x) \wedge \boxed{\mathtt{OpenRel}(x, \mathtt{Municipality})} \wedge \mathtt{location.containedby}(x, \mathtt{Brandenburg})$

- Why use the KB at all? Why not answer questions directly from text? Like information retrieval!

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
  - *Is P=NP?*
  - *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 plausibly appear in text… but this is still too broad
  - *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
  - *What temperature should I cook chicken to?* — could be written down in a KB but probably isn't
  - Today: can we do 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
- Parsing: find direct object of "pulled" in the document where the subject is James
- Don't need any complex semantic representations

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

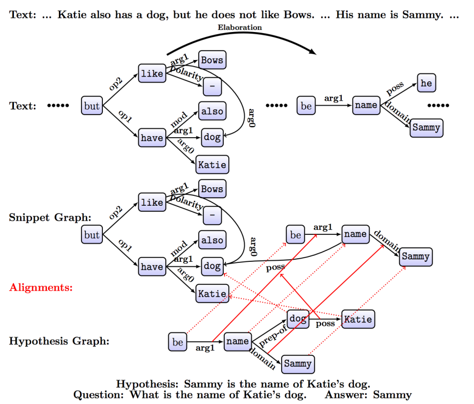
ngram sliding window

	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)

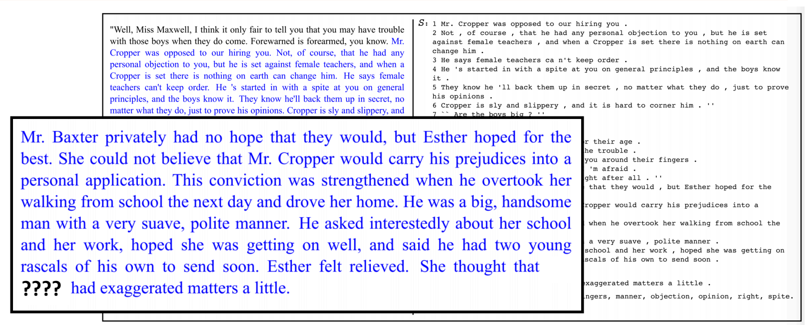
Dataset Explosion

- 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?

Children's Book Test

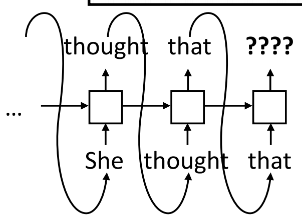


- Children's Book Test: take a section of a children's story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

LSTM Language Models

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that **????** had exaggerated matters a little.



- Predict next word with LSTM LM
- Context: either just the current sentence (query) or the whole document up to this point (query+context)

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)

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.

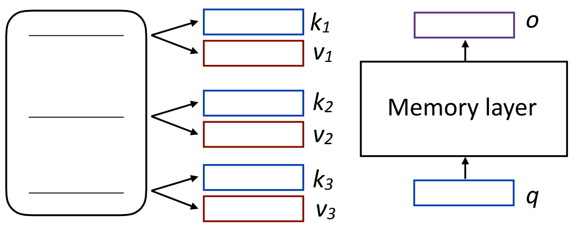
- Let's look at architectures for retrieval from a passage

Zellers et al. (2018)

Memory Networks

Memory Networks

- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output



$$o = \sum_i \alpha_i v_i$$

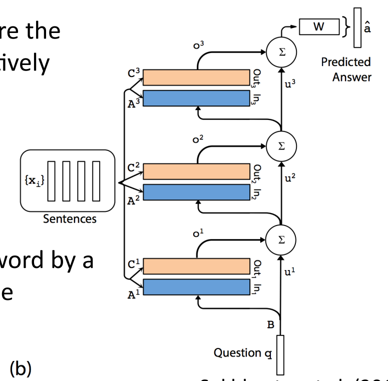
$$\alpha = \mathrm{softmax}(e)$$

$$e_i = q \cdot k_i$$

Sukhbaatar et al. (2015)

Memory Networks

- Three layers of memory network where the query representation is updated additively based on the memories at each step
- How to encode the sentences?
  - Bag of words (average embeddings)
  - Positional encoding: multiply each word by a vector capturing position in sentence



(b)

Sukhbaatar et al. (2015)

bAbI

- Evaluation on 20 tasks proposed as building blocks for building "AI-complete" systems
- Various levels of difficulty, exhibit different linguistic phenomena
- Small vocabulary, language isn't truly "natural"

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 13: Compound Coreference
Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning
In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Weston et al. (2014)

Evaluation: bAbI

	Baseline			MemN2N				
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint
Mean error (%)	6.7	51.3	40.2	25.1	20.3	25.8	15.6	13.3
Failed tasks (err. > 5%)	4	20	18	15	13	17	11	11

- 3-hop memory network does pretty well, better than LSTM at processing these types of examples

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow	Prediction: yellow			

Evaluation: Children's Book Test

METHODS	NAMED ENTITIES
HUMANS (QUERY)(*)	0.520
HUMANS (CONTEXT+QUERY)(*)	***0.816***
MAXIMUM FREQUENCY (CORPUS)	0.120
MAXIMUM FREQUENCY (CONTEXT)	0.335
SLIDING WINDOW	0.168
WORD DISTANCE MODEL	0.398
KNESER-NEY LANGUAGE MODEL	0.390
KNESER-NEY LANGUAGE MODEL + CACHE	0.439
LSTMS (QUERY)	0.408
LSTMS (CONTEXT+QUERY)	0.418
CONTEXTUAL LSTMS (WINDOW CONTEXT)	0.436
MEMNNS (LEXICAL MEMORY)	0.431
MEMNNS (WINDOW MEMORY)	0.493
MEMNNS (SENTENTIAL MEMORY + PE)	0.318
MEMNNS (WINDOW MEMORY + SELF-SUP.)	**0.666**

- Outperforms LSTMs substantially with the right supervision

Memory Network Takeaways

- Memory networks provide a way of attending to abstractions over the input
- Useful model for attending to multiple parts of an input
- What can we do with more basic attention?

CNN/Daily Mail: Attentive Reader

CNN/Daily Mail

- Single-document, (usually) single-sentence cloze task
- Formed based on article summaries — information should mostly be present, makes it easier than Children's Book Test
- Need to process the question, can't just use LSTM LMs

Passage

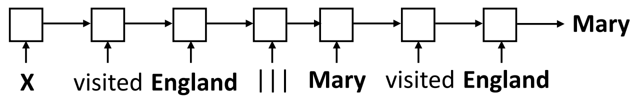
(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to 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 " books at @entity28 imprint @entity26 .

Question	Answer
characters in " @placeholder " movies have gradually become more diverse	@entity6

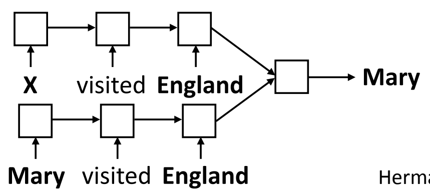
Hermann et al. (2015), Chen et al. (2016)

CNN/Daily Mail

- LSTM reader: encode question, encode passage, predict entity



- Can also use textual entailment-like models

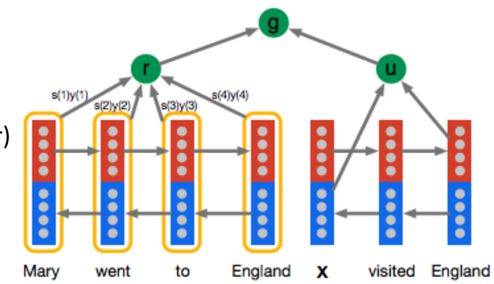


Multiclass classification problem over entities in the document

Hermann et al. (2015), Chen et al. (2016)

CNN/Daily Mail

- Attentive reader:
  u = encode query
  s = encode sentence
  r = attention(u -> s)
  prediction = f(candidate, u, r)
- Uses fixed-size representations for the final prediction, multiclass classification



Hermann et al. (2015)

CNN/Daily Mail

- Chen et al (2016): small changes to the attentive reader
- Additional analysis of the task found that many of the remaining questions were unanswerable or extremely difficult

	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	30.5	33.2	25.6	25.5
Exclusive frequency	36.6	39.3	32.7	32.8
Frame-semantic model	36.3	40.2	35.5	35.5
Word distance model	50.5	50.9	56.4	55.5
Deep LSTM Reader	55.0	57.0	63.3	62.2
Uniform Reader	39.0	39.4	34.6	34.4
Attentive Reader	61.6	63.0	**70.5**	**69.0**
Stanford Attentive Reader	76.2	76.5	**79.5**	**78.7**

Hermann et al. (2015), Chen et al. (2016)

SQuAD: Bidirectional Attention Flow

SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- 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

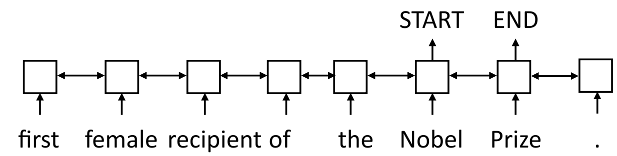
What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 | 1745 | 1745

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)

Bidirectional Attention Flow

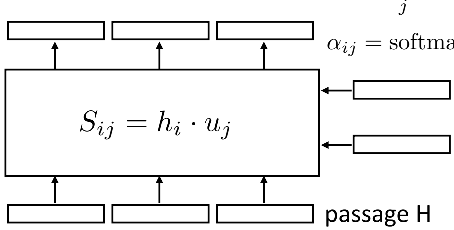
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

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

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

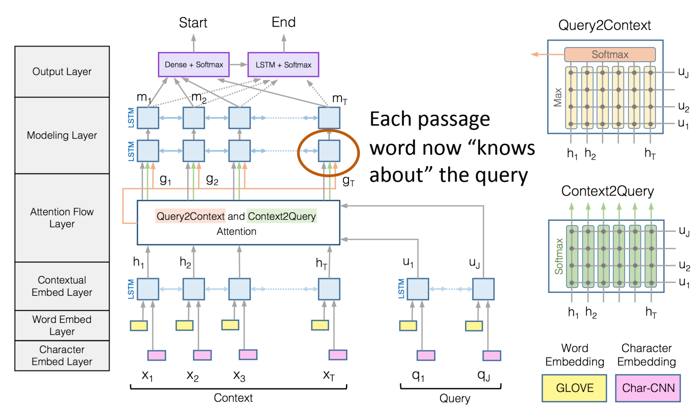
$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- dist over query



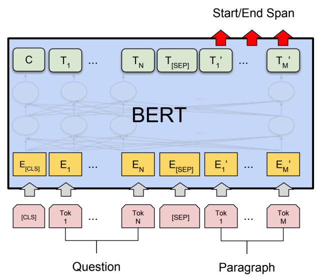
Seo et al. (2016)

Bidirectional Attention Flow



Seo et al. (2016)

QA with BERT



What was Marie Curie the first female recipient of ? [SEP] One of the most famous people born in Warsaw was Marie …

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

Devlin et al. (2019)

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: Today

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

Latest Datasets

- DROP
- SQuAD 2.0
- SQuAD 2.0
- Multi-hop: next time

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