A guide to the dataset explosion in QA, NLI, and commonsense reasoning

COLING 2020, Tutorial 6: Anna Rogers

COLING 2020

Anna Rogers
University of Copenhagen
arogers@sodas.ku.dk
@annargrs

Anna Rumshisky
University of Massachusetts Lowell
arum@cs.uml.edu
@arumshisky

Anna Rogers

Post-doctoral Associate
Center for Social Data Science

UNIVERSITY OF
COPENHAGEN

https://annargrs.github.io
arogers@sodas.ku.dk
@annargrs

Anna Rumshisky

Associate Professor
Text Machine Lab

UMASS LOWELL

http://text-machine.cs.uml.edu/
arum@cs.uml.edu
@arumshisky

Outline

High-level reasoning tasks in NLP system evaluation

The Dataset Explosion

- Question answering
- Commonsense reasoning
- Natural Language Inference (Anna Rumshisky)

Reality check

(Some) solutions

Open problems

What are "high-level reasoning tasks"? V1

Low-level

- part-of-speech tagging
- syntactic parsing
- named entity recognition
- ...

High-level

- question answering
- natural language inference
- commonsense reasoning
- ...

What are "high-level reasoning tasks"? V2

Low-level

- part-of-speech tagging
- syntactic parsing
- named entity recognition
- ...

Mid-level

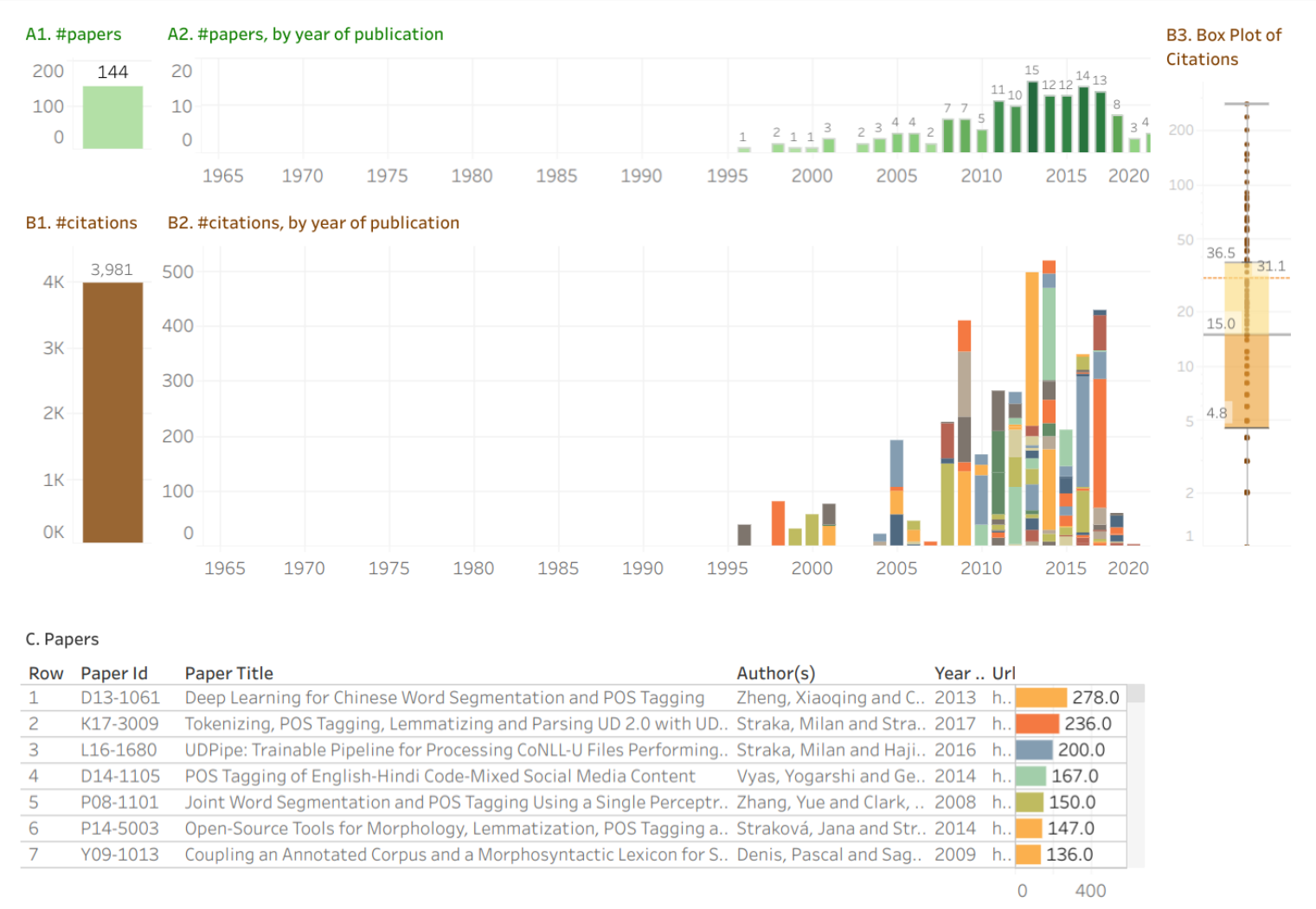
- sentiment analysis
- coreference resolution
- discourse relation extraction
- ...

High-level

- question answering
- natural language inference
- commonsense reasoning
- ...

High-level NLP tasks are gaining popularity!

NLP publications: POS-tagging



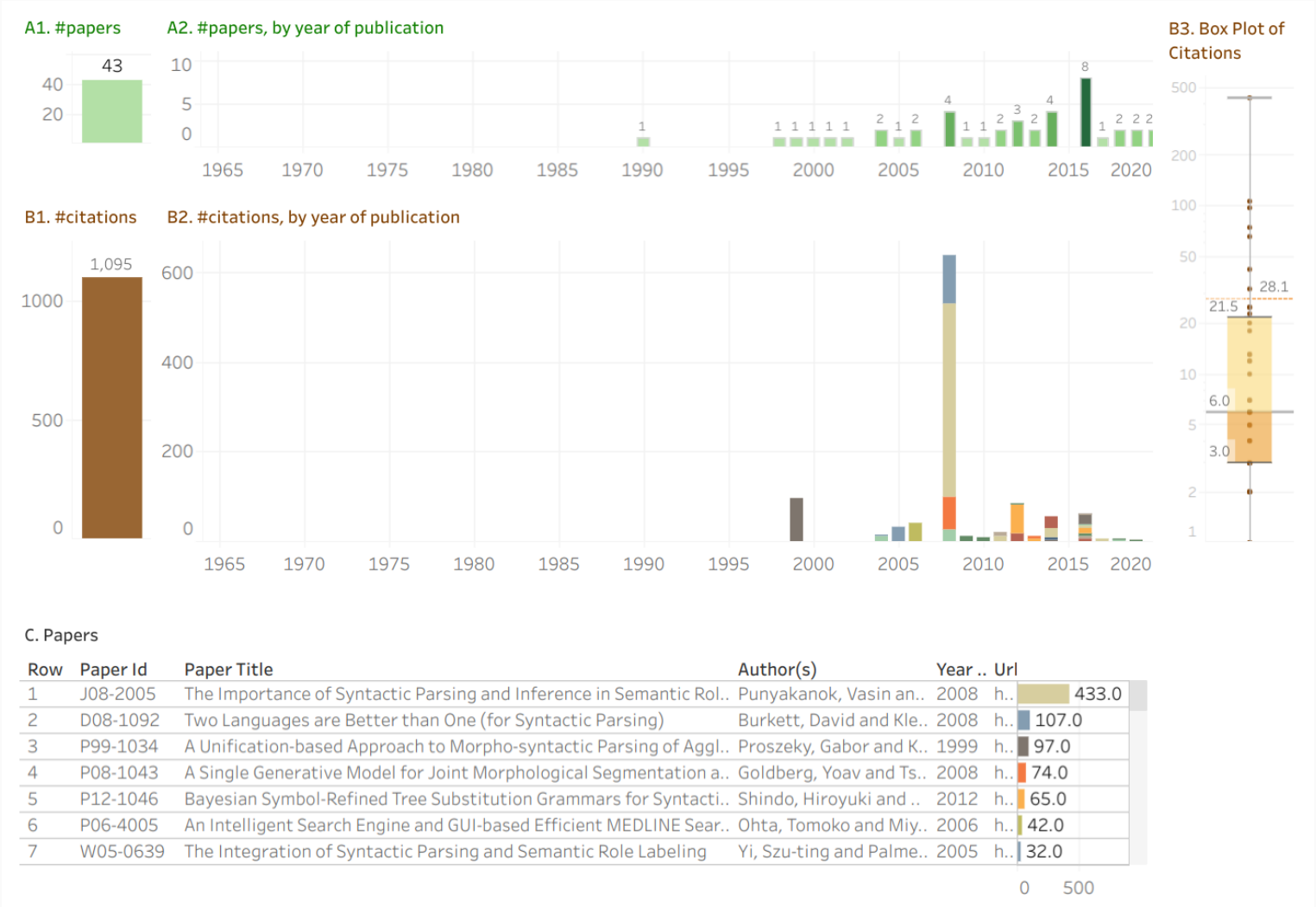
C. Papers

Row	Paper Id	Paper Title	Author(s)	Year ..	Url	
1	D13-1061	Deep Learning for Chinese Word Segmentation and POS Tagging	Zheng, Xiaoqing and C..	2013	h..	278.0
2	K17-3009	Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UD..	Straka, Milan and Stra..	2017	h..	236.0
3	L16-1680	UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing..	Straka, Milan and Haji..	2016	h..	200.0
4	D14-1105	POS Tagging of English-Hindi Code-Mixed Social Media Content	Vyas, Yogarshi and Ge..	2014	h..	167.0
5	P08-1101	Joint Word Segmentation and POS Tagging Using a Single Perceptr..	Zhang, Yue and Clark, ..	2008	h..	150.0
6	P14-5003	Open-Source Tools for Morphology, Lemmatization, POS Tagging a..	Straková, Jana and Str..	2014	h..	147.0
7	Y09-1013	Coupling an Annotated Corpus and a Morphosyntactic Lexicon for S..	Denis, Pascal and Sag..	2009	h..	136.0

0 400

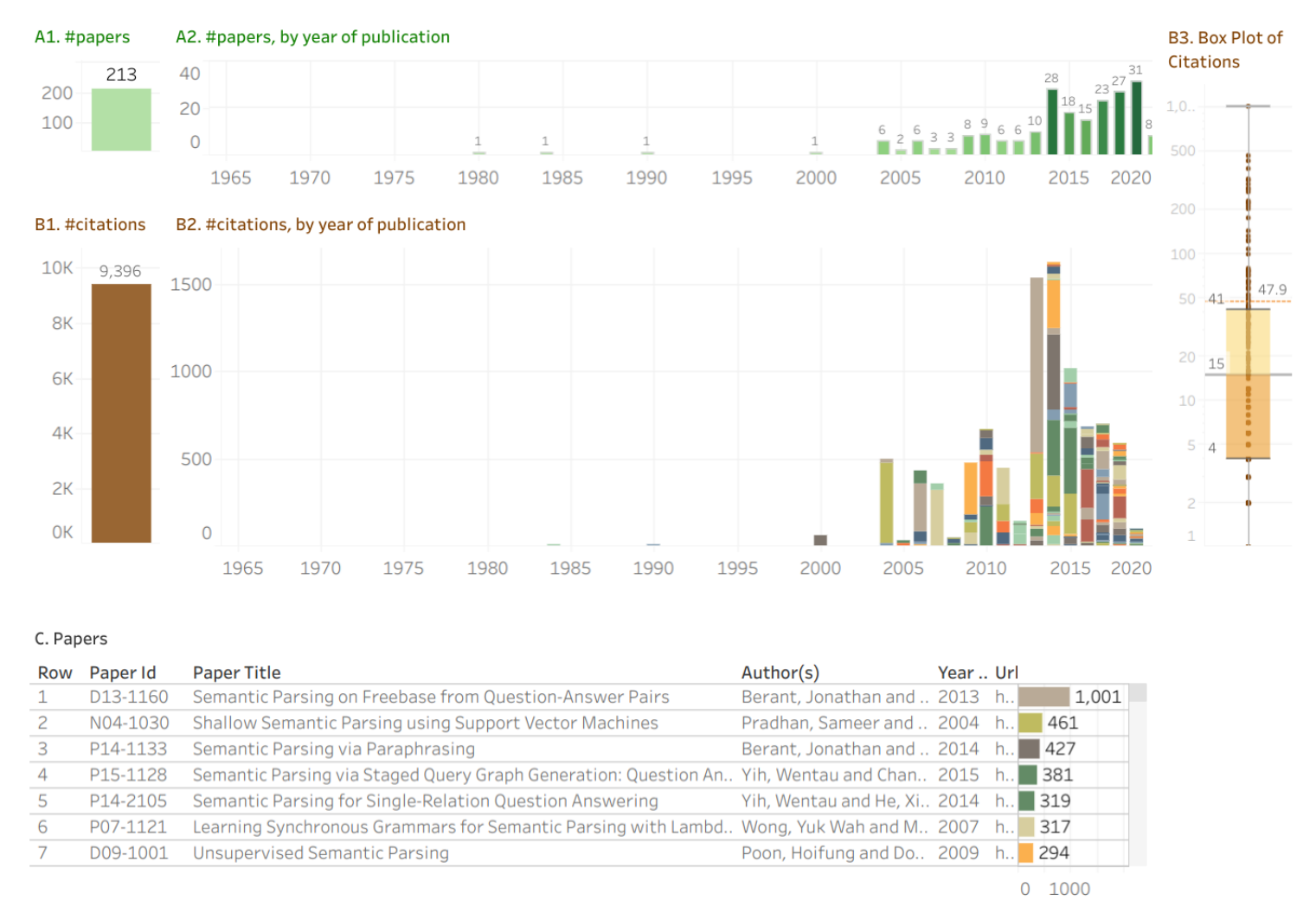
Mohammad (2020)

NLP publications: syntactic parsing



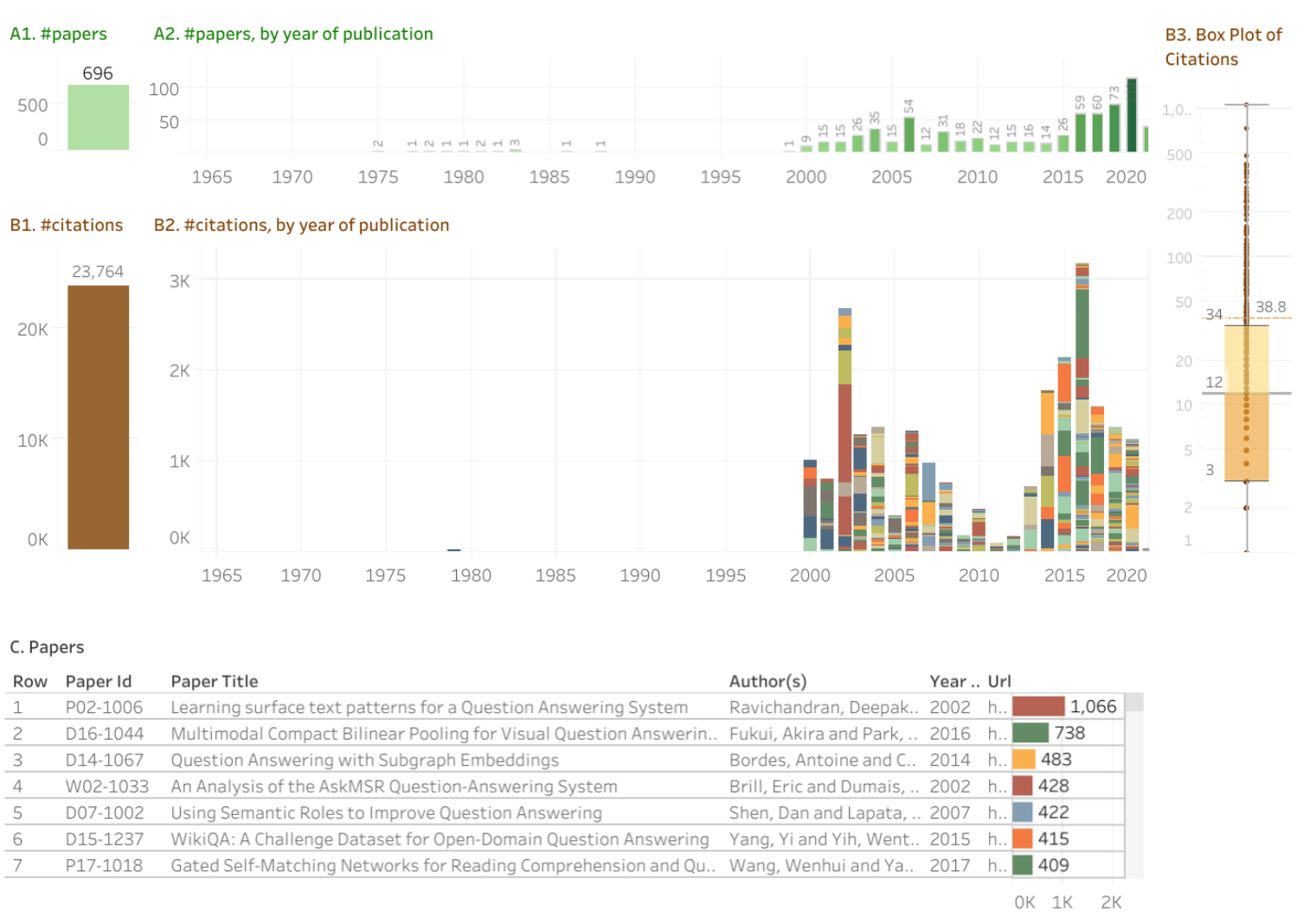
Mohammad (2020)

NLP publications: semantic parsing



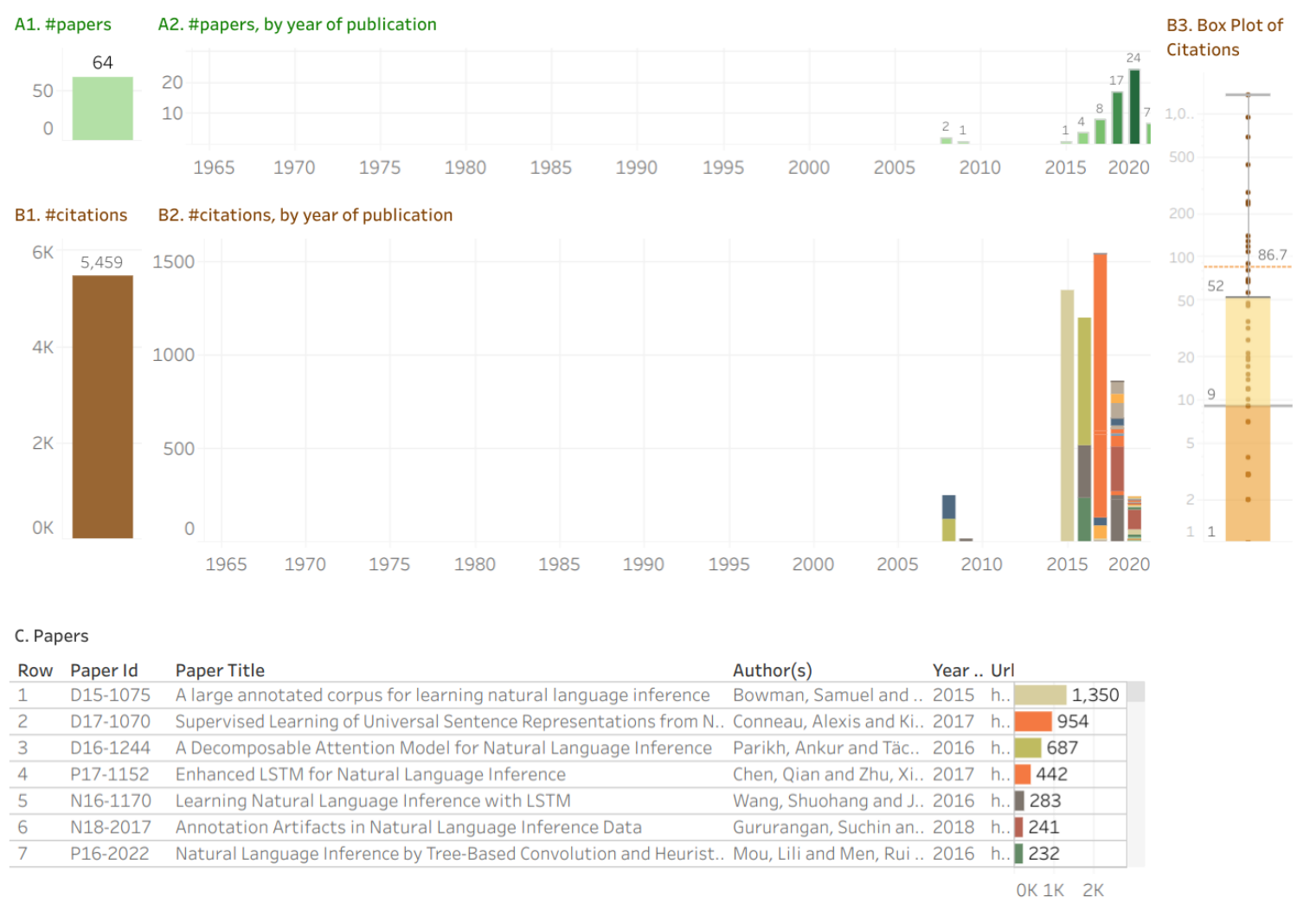
Mohammad (2020)

NLP publications: question answering



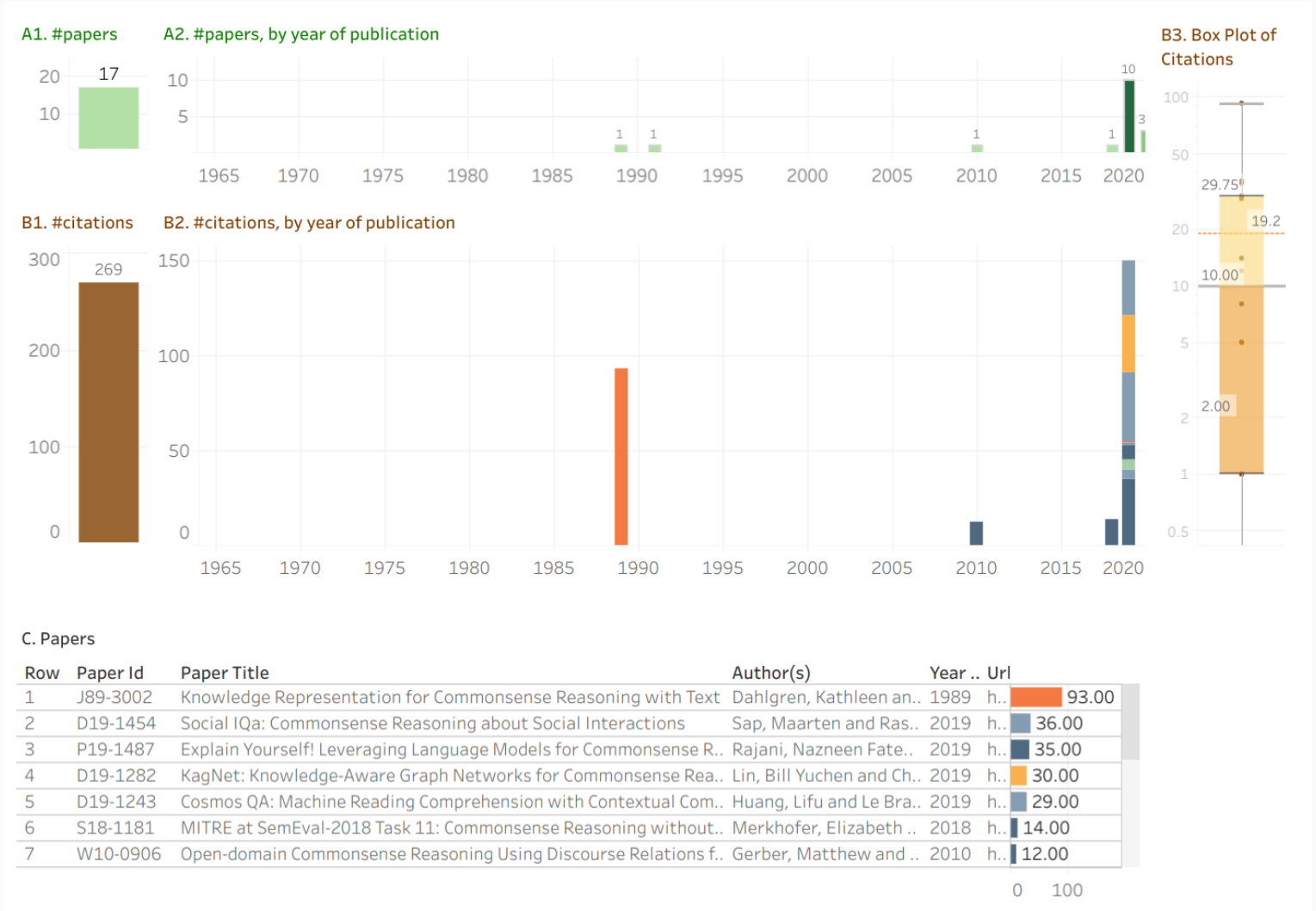
Mohammad (2020)

NLP publications: natural language inference



Mohammad (2020)

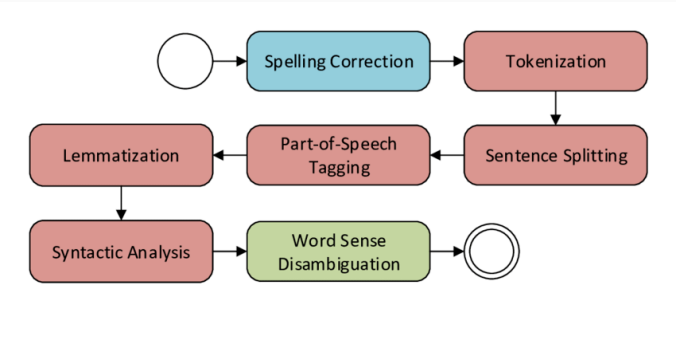
NLP publications: commonsense reasoning



Row	Paper Id	Paper Title	Author(s)	Year ..	Url	
1	J89-3002	Knowledge Representation for Commonsense Reasoning with Text	Dahlgren, Kathleen an..	1989	h..	93.00
2	D19-1454	Social IQa: Commonsense Reasoning about Social Interactions	Sap, Maarten and Ras..	2019	h..	36.00
3	P19-1487	Explain Yourself! Leveraging Language Models for Commonsense R..	Rajani, Nazneen Fate..	2019	h..	35.00
4	D19-1282	KagNet: Knowledge-Aware Graph Networks for Commonsense Rea..	Lin, Bill Yuchen and Ch..	2019	h..	30.00
5	D19-1243	Cosmos QA: Machine Reading Comprehension with Contextual Com..	Huang, Lifu and Le Bra..	2019	h..	29.00
6	S18-1181	MITRE at SemEval-2018 Task 11: Commonsense Reasoning without..	Merkhofer, Elizabeth ..	2018	h..	14.00
7	W10-0906	Open-domain Commonsense Reasoning Using Discourse Relations f..	Gerber, Matthew and ..	2010	h..	12.00

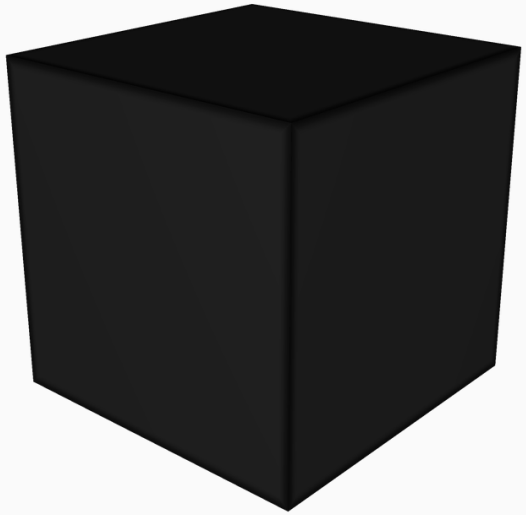
Mohammad (2020)

Why are these tasks so popular? A1: end-to-end systems!

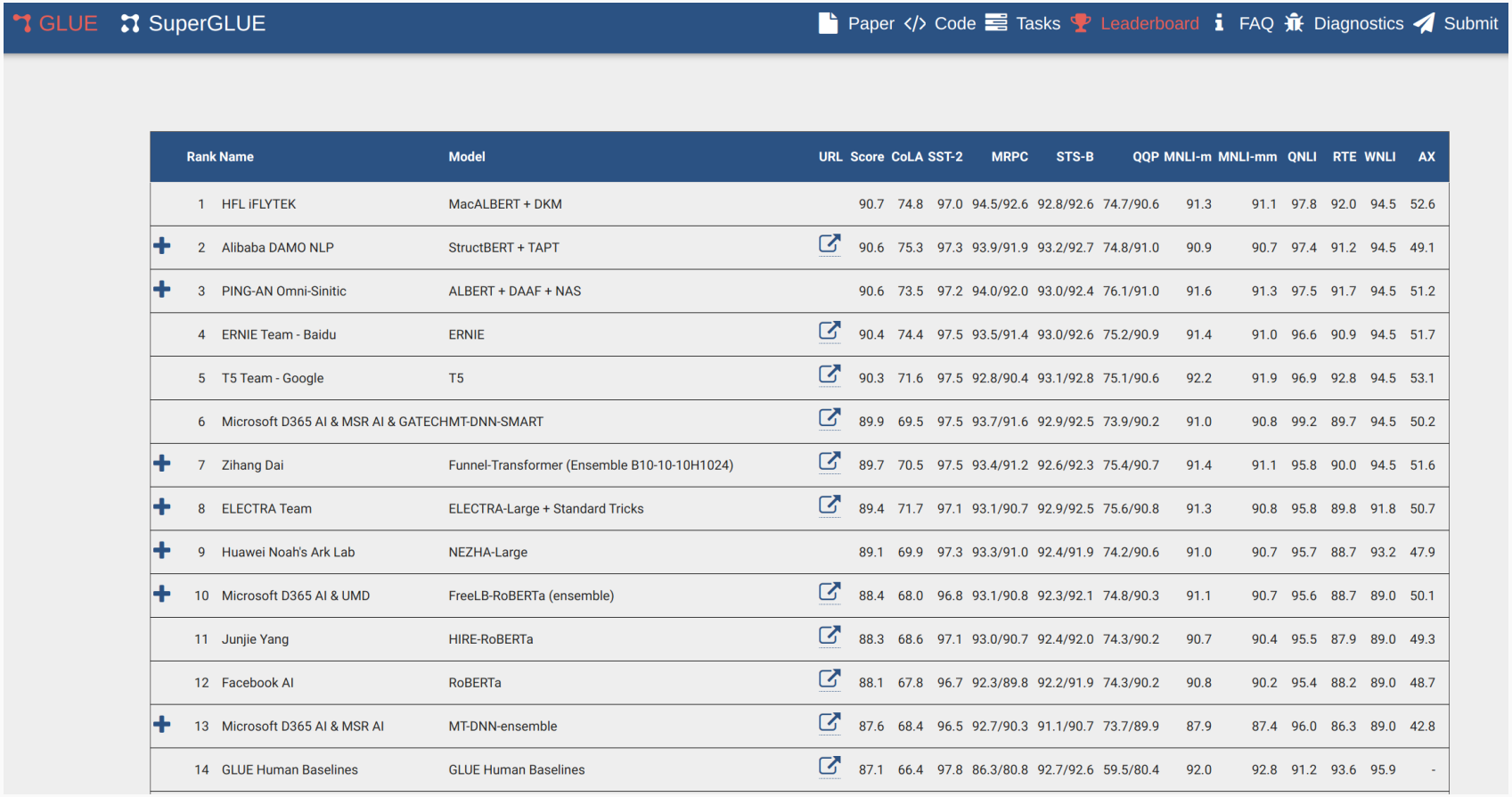


(Schouten et al., 2017)

or



Why are these tasks so popular? A2: leaderboards!



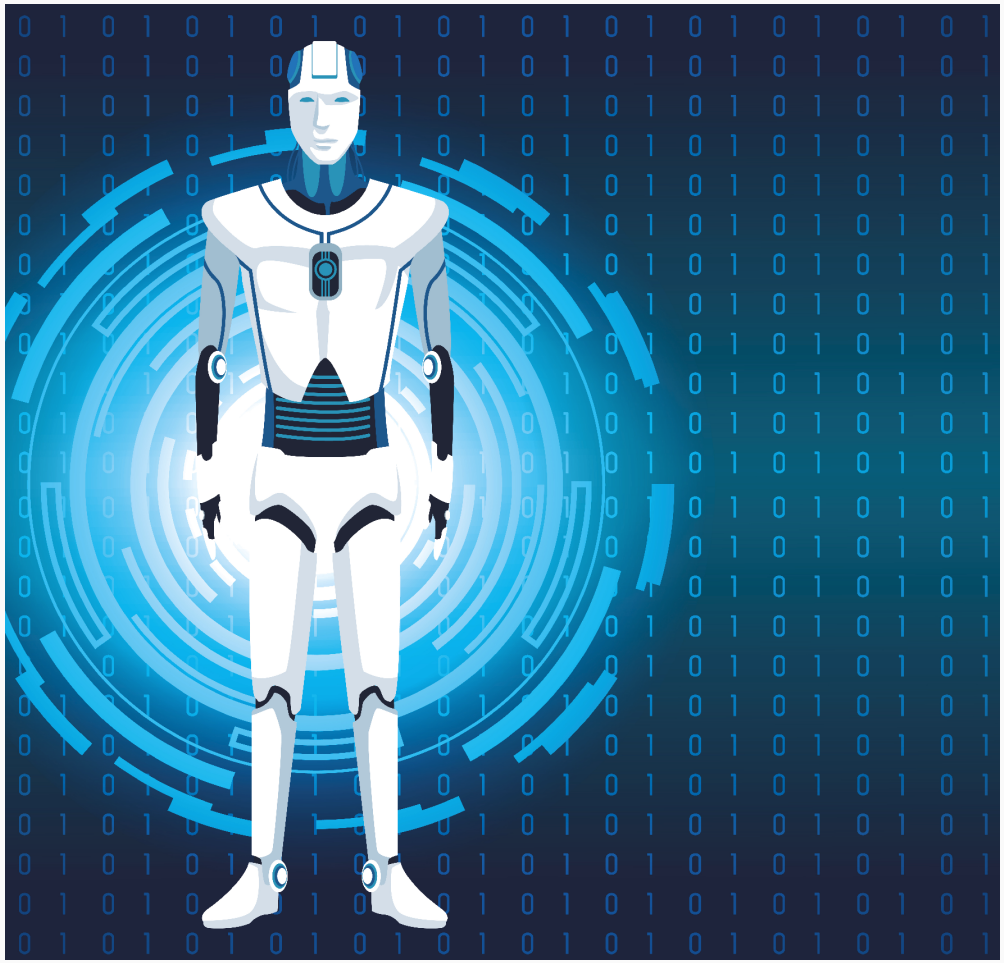
GLUE SuperGLUE Paper Code Tasks Leaderboard FAQ Diagnostics Submit

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
2	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	90.7	97.4	91.2	94.5	49.1
3	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
4	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6	90.9	94.5	51.7
5	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
6	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
7	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
8	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
9	Huawei Noah's Ark Lab	NEZHA-Large		89.1	69.9	97.3	93.3/91.0	92.4/91.9	74.2/90.6	91.0	90.7	95.7	88.7	93.2	47.9
10	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
11	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
12	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
13	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
14	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

Wang et al. (2018)

Why are these tasks so popular? A3: conversational agents!

Why are these tasks so popular? A4: they look cool!



Outline

High-level reasoning tasks in NLP system evaluation

The Dataset Explosion

- Question answering
- Commonsense reasoning
- Natural Language Inference (Anna Rumshisky)

Reality check

(Some) solutions

Open problems

Tasks vs formats

Everything is question answering!

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

(McCann et al., 2018)

Everything is question answering!

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

(McCann et al., 2018)

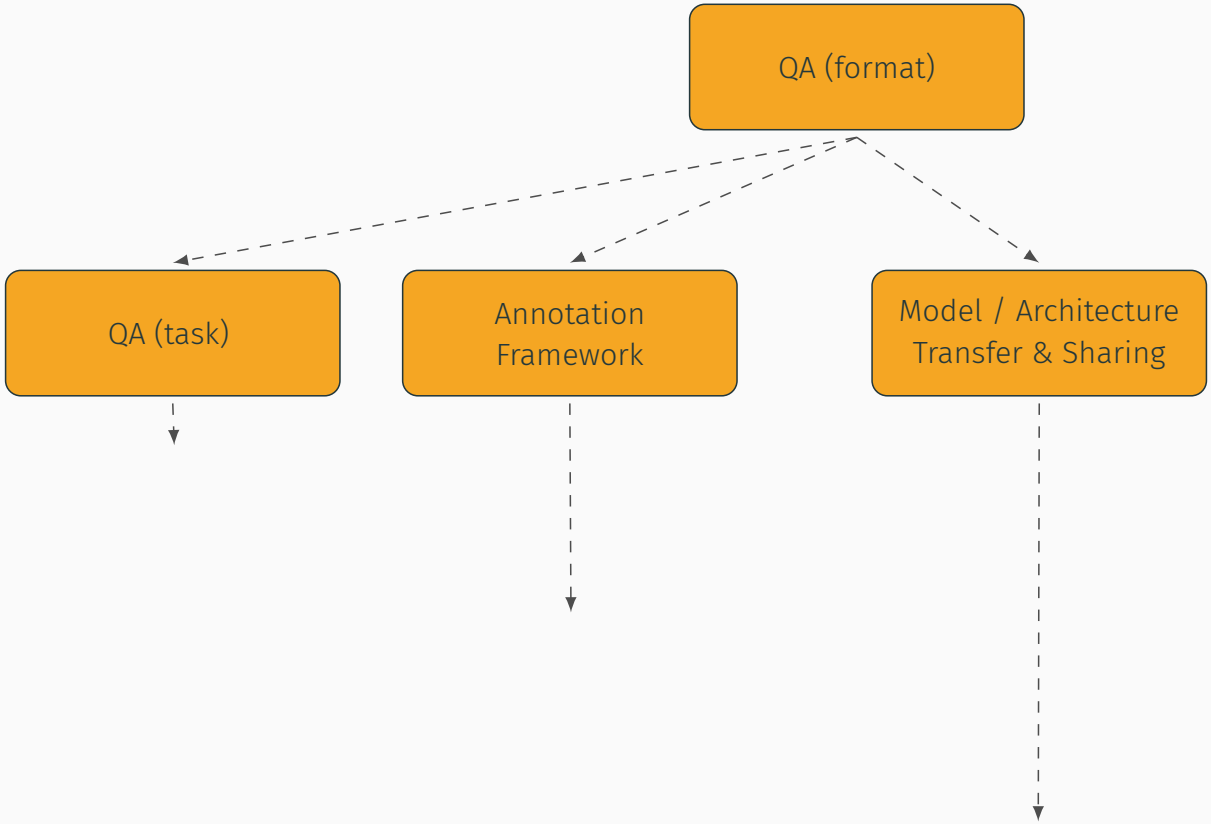
Everything is NLI!

Order
▶ *We waited until 2:25 PM and then left.* *The waiting started before the leaving started.*
▶ *Reggie said he will pay us soon.* *The paying ended before the saying started.*
Duration
▶ *The greeter said there was about 15 mins waiting.* *The saying did take or will take shorter than an hour.*
▶ *Randy , this is the issue I left you the voice mail on.* *The leaving did take or will take longer than a day.*

(Vashishtha et al., 2020)

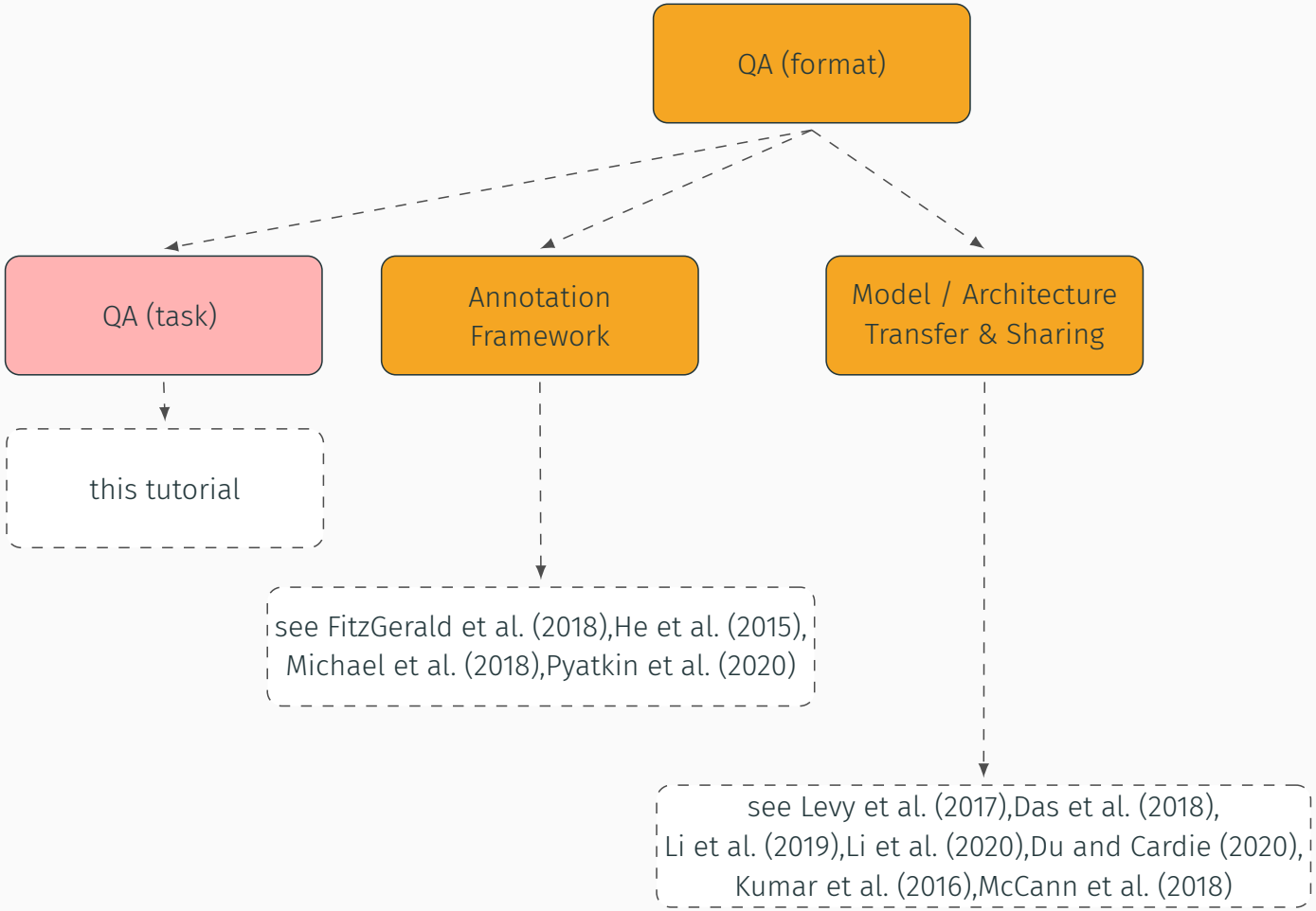
Let's rephrase?

Question Answering is (also) a format



inspired by: (Gardner et al., 2019)

Question Answering is (also) a format



inspired by: (Gardner et al., 2019)

When is QA a format?

Classification
What is the sentiment?

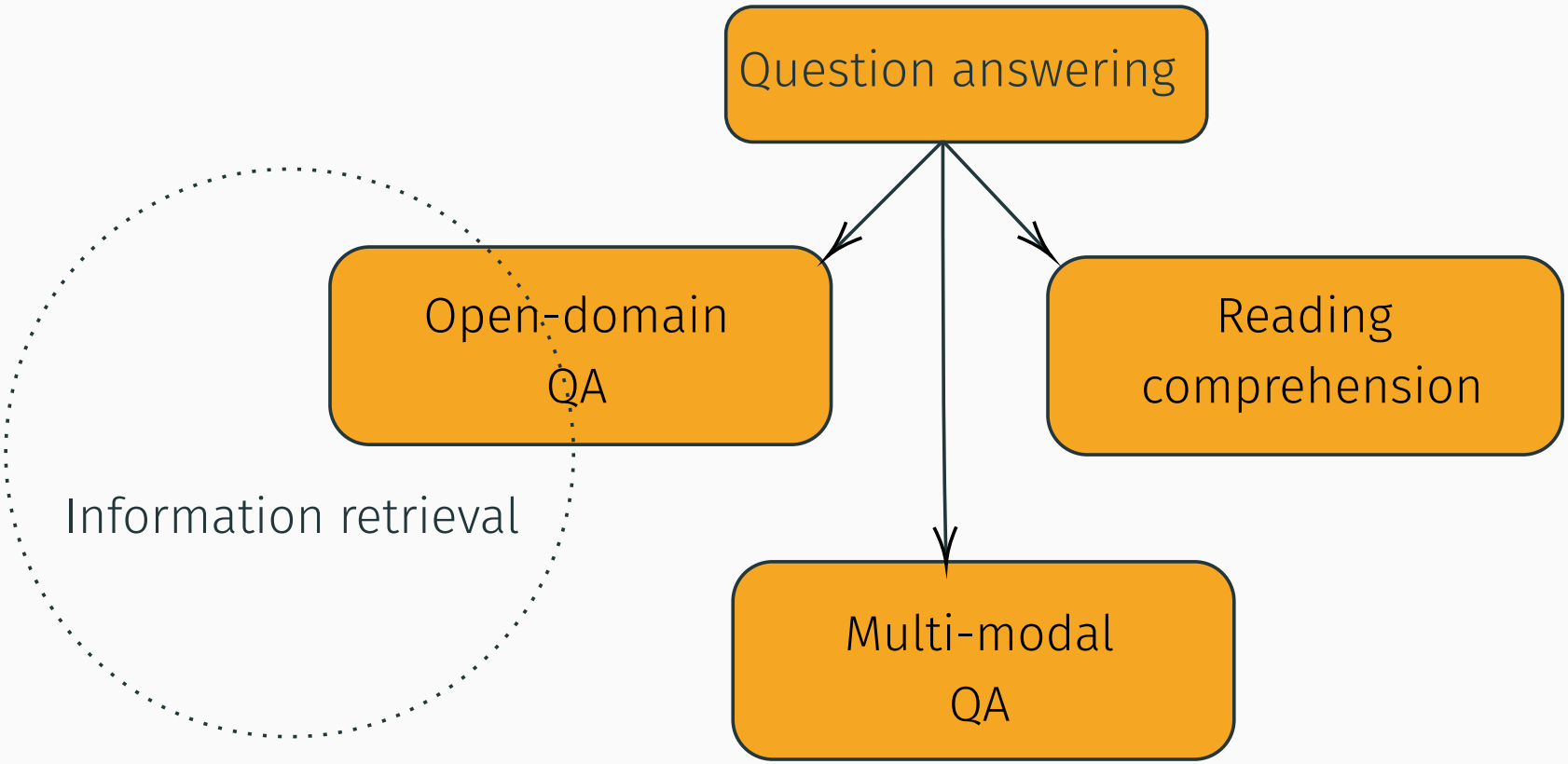
Template-filling
When was <PERSON> born?

Open-ended
(too many templates and/or variables)

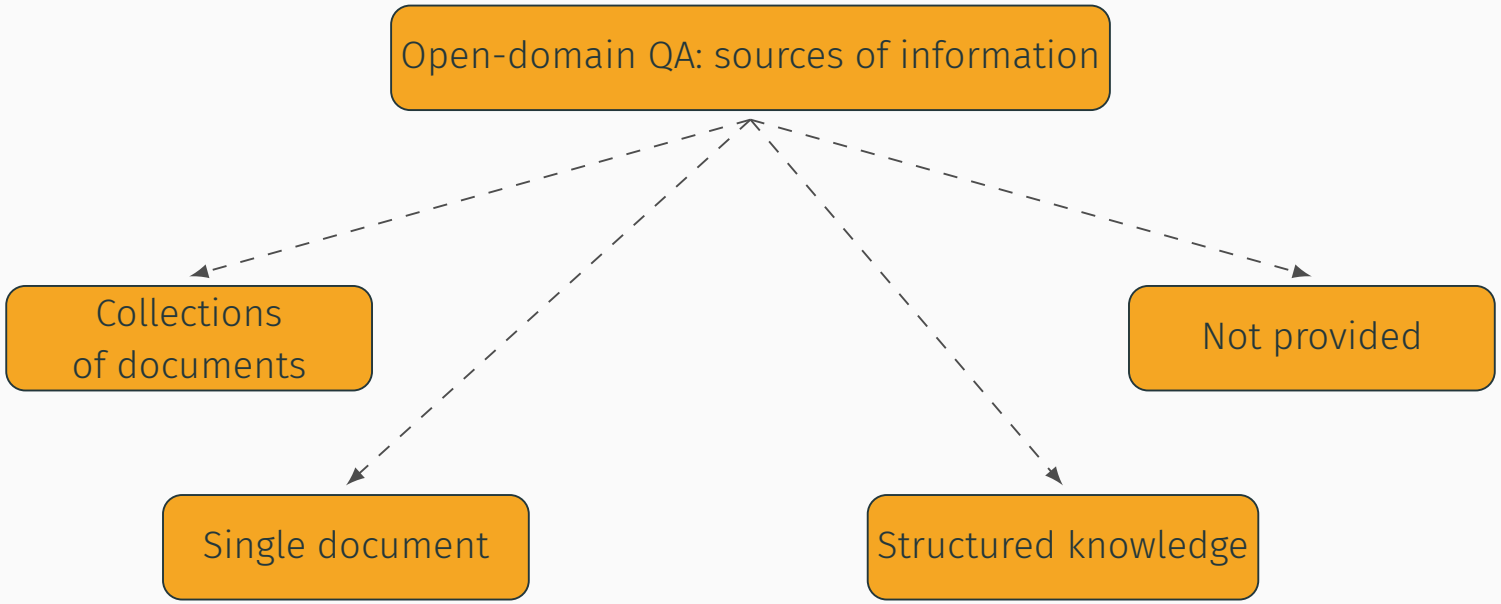
(Gardner et al., 2019)

QA: sources of information

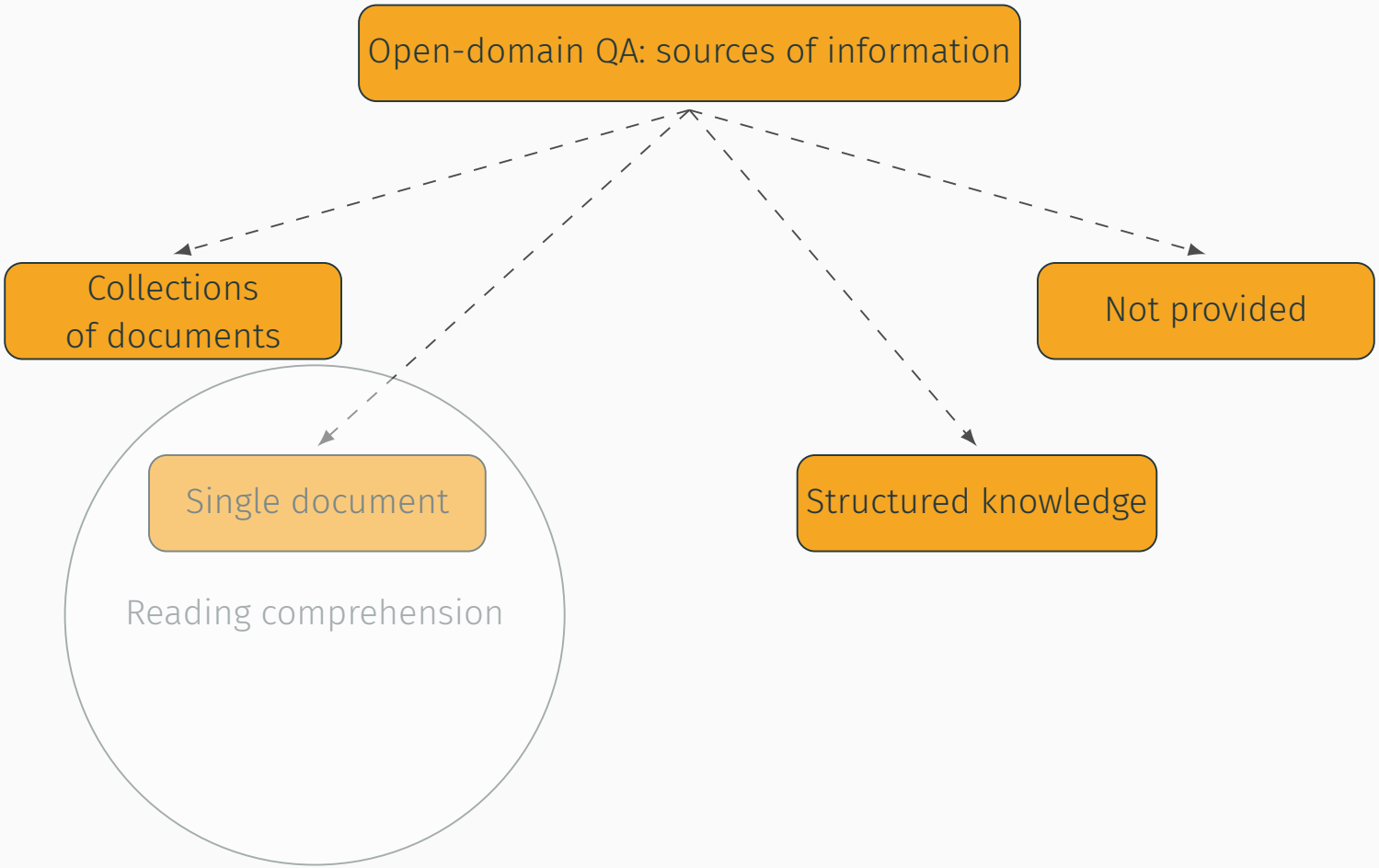
Question answering subfields



Area: Open-domain question answering

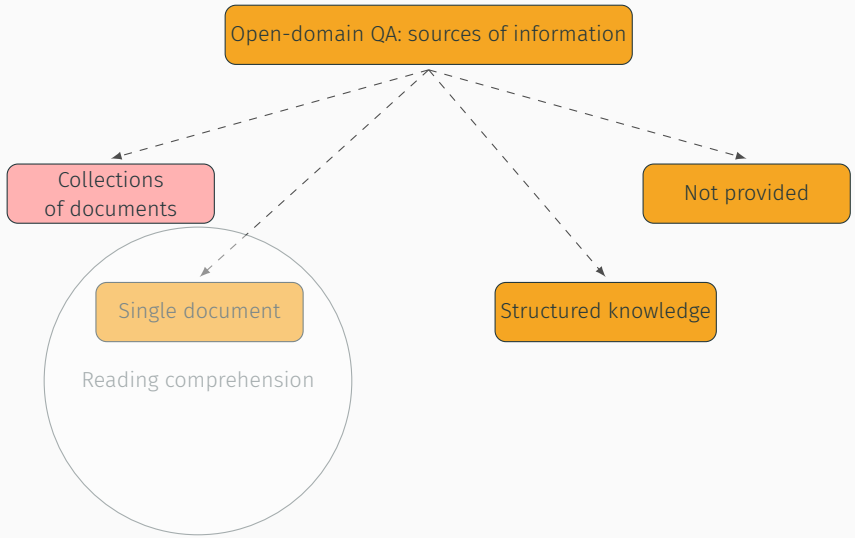


Area: Open-domain question answering



QA on collections of documents: datasets

- TriviaQA (Joshi et al., 2017);
- SearchQA (Dunn et al., 2017);
- MS MARCO (Bajaj et al., 2016);
- AmazonQA (Gupta et al., 2019);
- TrecQA-based data by Tsai et al. (2015);
- Chinese: WebQA; (Li et al., 2016);
- ...



TriviaQA: example

Q: Who was the man behind The Chipmunks?

A: David Seville

Context 1: "Alvin and the Chipmunks (2007) - IMDb IMDb 17 January 2017 4:34 PM, UTC NEWS There was an error trying to load your rating for this title. Some parts of this page won't work property. Please reload or try later. X Beta I'm Watching This! Keep track of everything you watch; tell your friends. Error Alvin and the Chipmunks (2007) PG | A struggling songwriter named Dave Seville finds success when he comes across a trio of singing chipmunks ..."

Context 2: "The Chipmunks - Biography | Billboard The Chipmunks Alvin Simon Theodore Ross Bagdasarian David Seville Possibly the most popular TV and musical cartoon of all time, the Chipmunks enjoyed several periods of prosperity – beginning with the '60s era of adolescent Baby Boomers, cresting in the '80s, when the Boomers' children were growing up, and riding the wave clear into the new millennium. The man who brought the Chipmunks to life, Ross Bagdasarian, was born on January 27, 1919, in Fresno, California...."

...

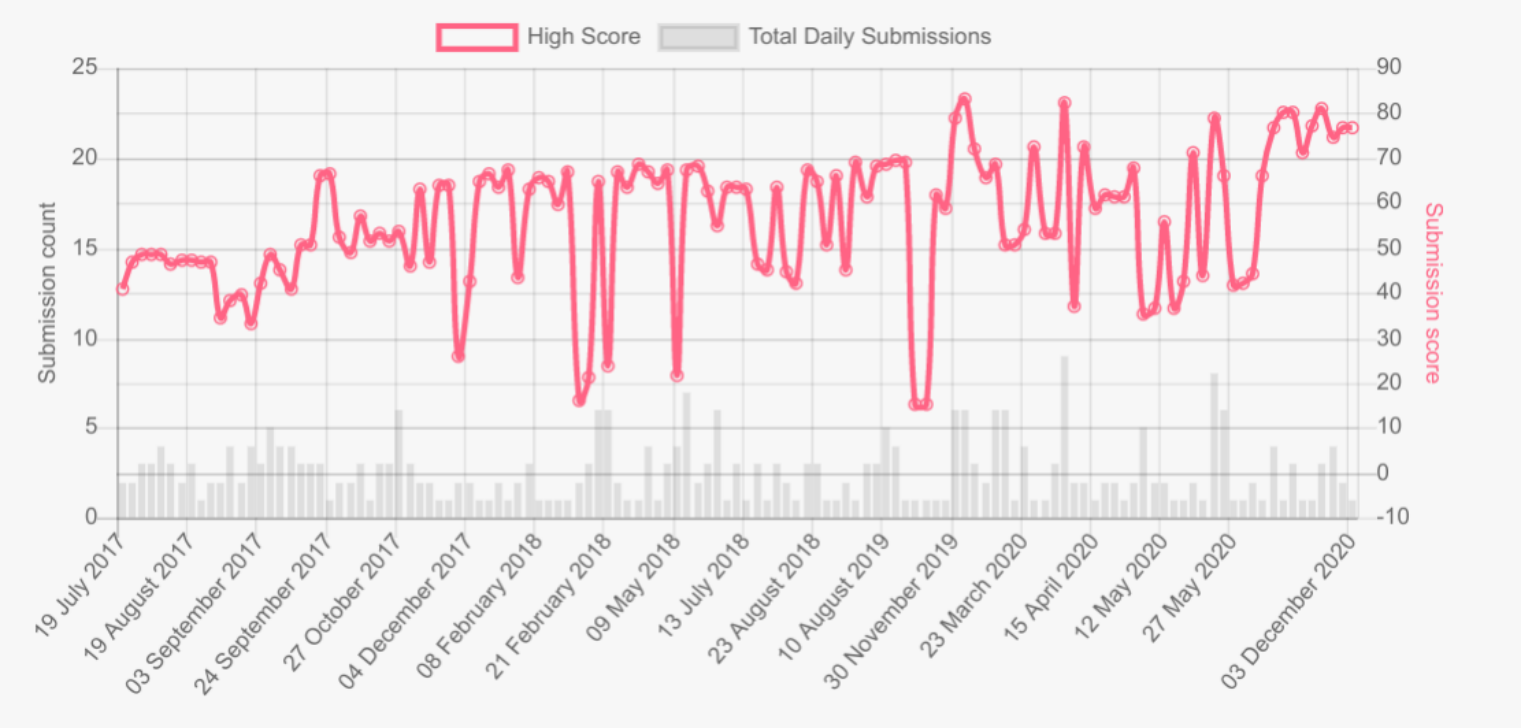
Context 10: "Alvin and the Chipmunks: The Squeakquel | Channel Awesome | Fandom powered by Wikia Alvin and the Chipmunks: The Squeakquel 2,694pages on Alvin and the Chipmunks: The Squeakquel Released (For the Nostalgia Critic's review of the movie, go here) ...."

Trivia QA: data

- Collection methodology: Questions authored by trivia enthusiasts are automatically paired with Wikipedia passages and web snippets, assuming that the presence of the answer string in the text indicates the presence of the answer. This holds about 75% of the time. A small subset of texts is manually validated to contain the answer.
- Challenges: lexical variation (synonims), lexical variation +world knowledge, syntactic variation, multi-hop reasoning, processing lists and tables
- Dataset size: 95K Q+A, 650K Q+A+evidence triplets, 1975 verified triplets

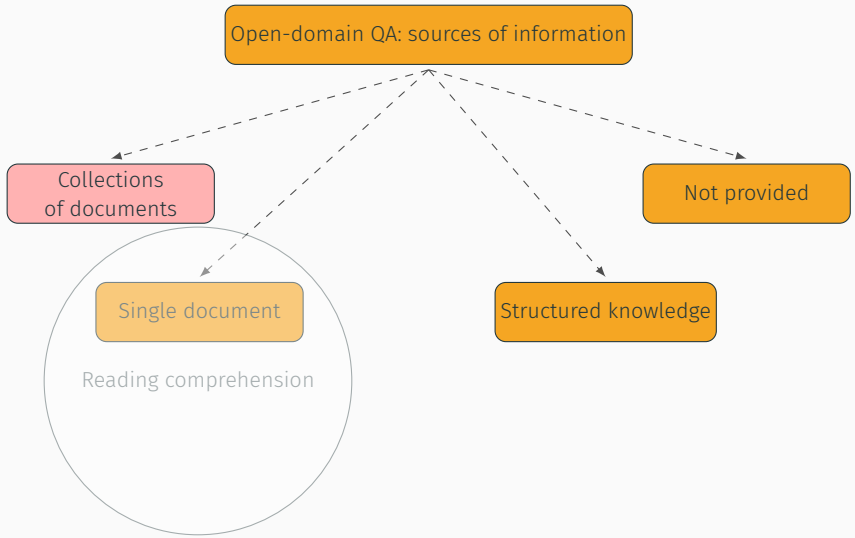
TriviaQA: status

Human performance: 79.7% on the Wikipedia domain, and 75.4% on the web domain



Special case: multi-hop QA

- HotPotQA (Yang et al., 2018a);
- QAngaroo (Welbl et al., 2018);
- ComplexWebQuestions (Talmor and Berant, 2018);
- HybridQA (Chen et al., 2020b);
- ...



HotPotQA: example (Yang et al., 2018a)

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

HotPotQA: data (Yang et al., 2018a)

- Collection methodology: questions are written by crowdworkers based on *several* wikipedia excerpts, identifying supporting facts
- Challenges: multi-hop reasoning, comparative questions
- Dataset size: 113K

HotPotQA: status(Yang et al., 2018a)

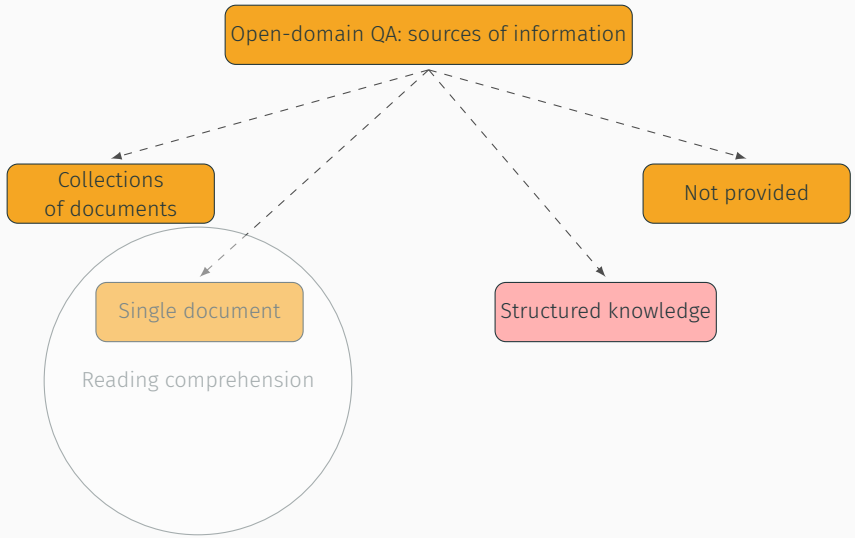
Human performance EM/F1: Answer (83.6/91.4), Supporting facts (61.5/90.04), Joint (52.3/82.55)

In the *distractor* setting, a question-answering system reads 10 paragraphs to provide an answer (Ans) to a question. They must also justify these answers with supporting facts (Sup).

	Model	Code	Ans		Sup		Joint	
			EM	F_1	EM	F_1	EM	F_1
1 Sep 6, 2020	SpiderNet-large (single model) *Kingsoft AI Lab*		70.15	**83.02**	**63.82**	88.85	**47.54**	**74.88**
2 Nov 23, 2020	Anonymous (single model) *Anonymous*		**70.24**	82.36	62.26	88.46	46.81	74.27
3 Dec 1, 2019	HGN-large (single model) *Anonymous*		69.22	82.19	62.76	88.47	47.11	74.21
4 Nov 15, 2020	AMGN (single model) *Anonymous*		69.89	82.79	62.67	88.12	46.59	74.20
5 Jun 10, 2020	BFR-Graph (single model) *Anonymous*		70.06	82.20	61.33	88.41	45.92	74.13
6 May 11, 2020	GSAN-large (single model) *Anonymous*		68.57	81.62	62.36	88.73	46.06	73.89
7 Oct 6, 2020	FFReader-large (single model) *Anonymous*		68.89	82.16	62.10	88.42	45.61	73.78
8 May 28, 2020	ETC-large (single model) *Anonymous*		68.12	81.18	63.25	**89.09**	46.40	73.62
9 May 28, 2020	Longformer (single model) *Anonymous*		68.00	81.25	63.09	88.34	45.91	73.16

Task: QA on structured knowledge

- FreebaseQA (Jiang et al., 2019)
- Event-QA (Costa et al., 2020)
- WikiTableQuestions (Pasupat and Liang, 2015)
- WikiOps (Cho et al., 2018)
- WikiReading (Hewlett et al., 2016)
- SimpleQuestions (Bordes et al., 2015)
- WikiSQL (Zhong et al., 2017)
- Russian RuBQ (Korablinov and Braslavski, 2020), Chinese TableQA (Sun et al., 2020), Korean TableQA (Park et al.)
- ...



SimpleQuestions: example

What American cartoonist is the creator of Andy Lippincott?	(andy_lippincott, character_created_by, garry_trudeau)
Which forest is Fires Creek in?	(fires_creek, containedby, nantahala_national_forest)
What is an active ingredient in childrens earache relief ?	(childrens_earache_relief, active_ingredients, capsicum)
What does Jimmy Neutron do?	(jimmy_neutron, fictional_character_occupation, inventor)
What dietary restriction is incompatible with kimchi?	(kimchi, incompatible_with_dietary_restrictions, veganism)

Table 1: **Examples of simple QA**. Questions and corresponding facts have been extracted from the new dataset SimpleQuestions introduced in this paper. Actual answers are underlined.

- Collection methodology: crowdworkers asked to write questions involving the subject and the relationship of a KB fact, with object as the correct answer
- Challenges: large-scale data
- Dataset size: 100K

(Bordes et al., 2015)

FreebaseQA: example

Components	Example 1	Example 2
Question [Answer]	Which 18th century author wrote Clarissa (or The History of a Young Lady), said to be the longest novel in the English language? [Samuel Richardson]	What is the correct name of the character voiced by Angela Lansbury in Beauty and The Beast? [Mrs Potts]
Subject (Freebase ID)	Clarissa (m.05s1st)	Angela Lansbury (m.0161h5)
Predicate	book.written-work.author	film.actor.dubbing-performances
Secondary Predicate	-	film.dubbing-performance.character
Object/Answer (ID)	Samuel Richardson (m.0hb27)	Mrs Potts (m.02vw823)

- Collection methodology: questions and answers collected from trivia websites are auto-matched with freebase subject-predicate-object triples, and matches are verified by crowdworkers
- Challenges: trivia questions more diverse and complex than existing data for querying KBs
- Dataset size: 28,348 unique questions

(Pasupat and Liang, 2015)

FreebaseQA: status

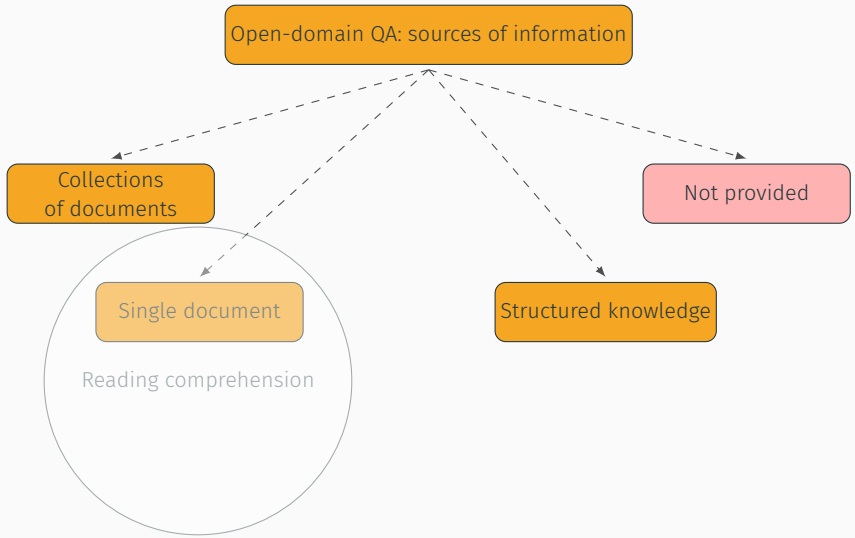
Human performance: ?

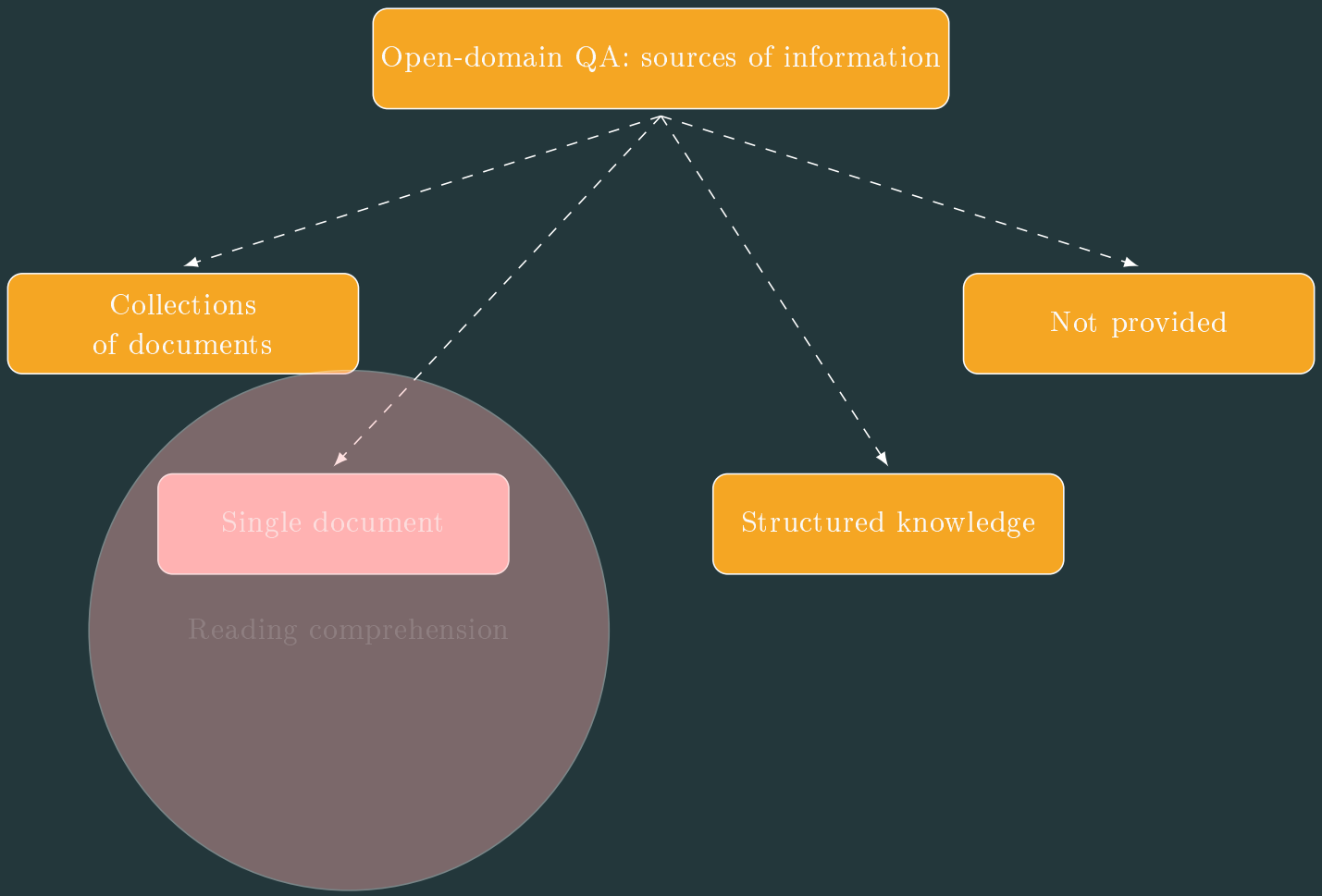
Training dataset	Test dataset	BuboQA	HR-BiLSTM	KBQA-Adapter	KEQA
FBQ	F917	17.29	**36.31**	35.73	36.02
	FBQ	**38.25**	28.40	28.78	28.73
	SQ	23.77	38.55	39.19	**42.97**
	WQ	29.10	30.27	31.43	**33.18**
SQ	F917	40.92	56.20	**59.37**	45.24
	FBQ	**20.08**	17.84	18.13	14.03
	SQ	74.81	72.30	72.01	**75.35**
	WQ	**41.79**	35.27	36.32	40.40
WQ	F917	12.68	29.97	29.39	**32.85**
	FBQ	7.94	7.61	8.37	**8.90**
	SQ	16.46	33.18	35.32	**38.01**
	WQ	61.23	49.94	49.36	**65.19**

(Han et al., 2020), see the paper for task status discussion

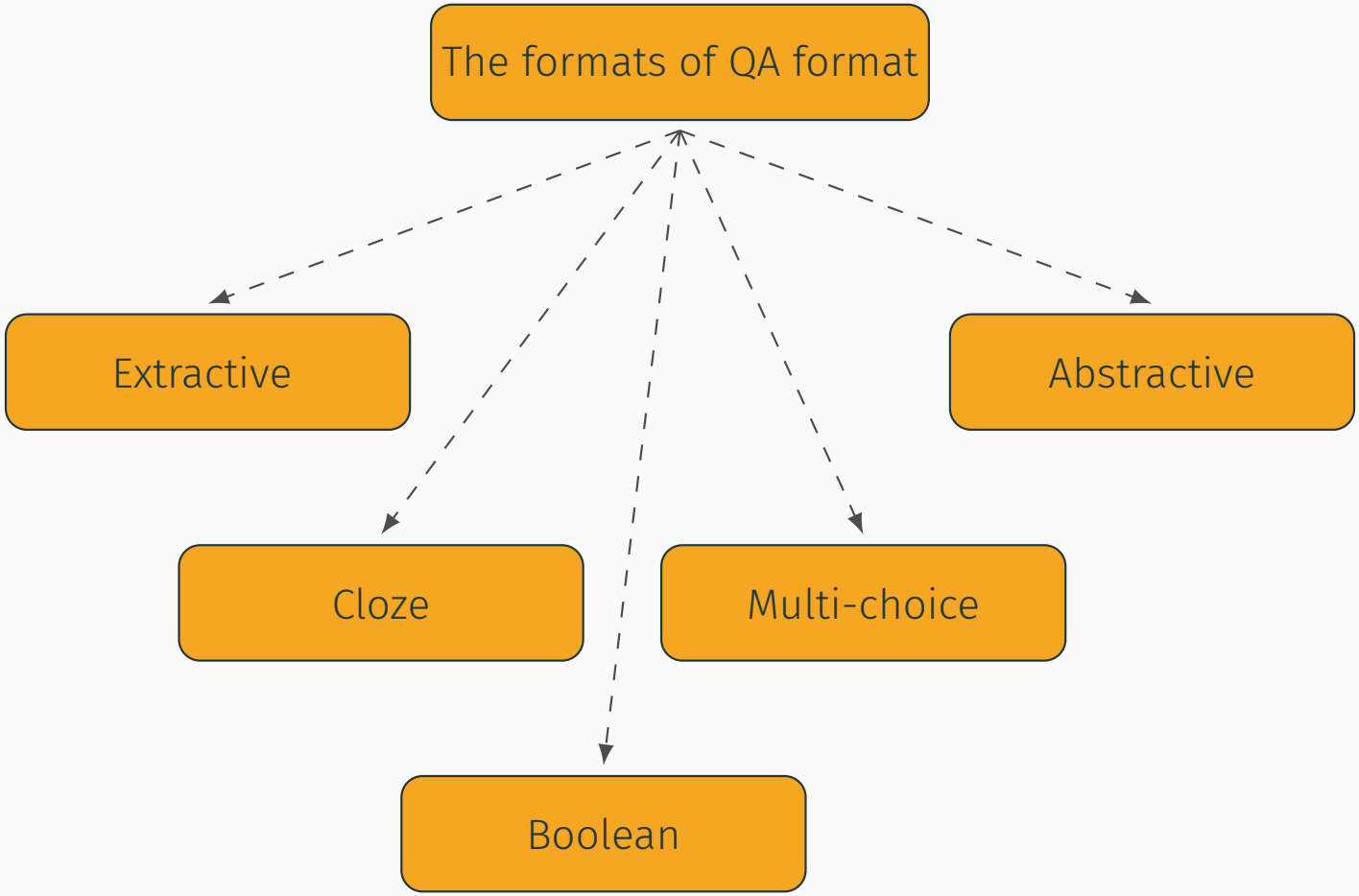
Task: QA without provided evidence

- retrieving evidence candidates with IR at inference time
- integrating KBs, e.g. with memory networks (Bordes et al., 2015)
- directly querying text data, e.g. latent retrieval (Lee et al., 2019)
- pre-trained model weights (Brown et al., 2020)



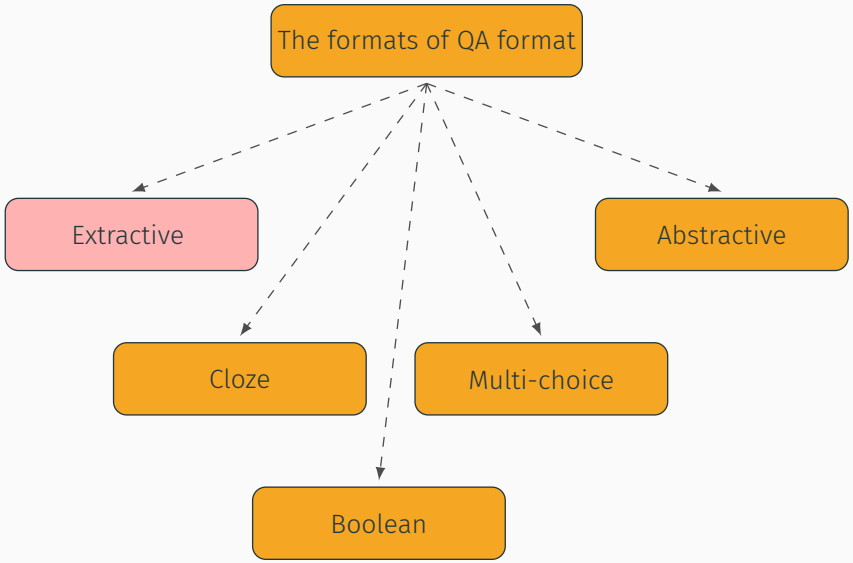

Open-domain QA: sources of information
Collections of documents
Not provided
Single document
Structured knowledge
Reading comprehension

The formats of QA format



Extractive QA

- SQuAD (Rajpurkar et al., 2016, 2018)
- Natural Questions (Kwiatkowski et al., 2019)
- HotpotQA (Yang et al., 2018a)
- NewsQA (Trischler et al., 2016)
- French FQuAD (d'Hoffschmidt et al., 2020), Chinese DRCD (Shao et al., 2019), Russian SberQuAD (Efimov et al., 2020), multilingual xQuAD (Artetxe et al., 2019), TYDI QA (Clark et al.), MLQA (Lewis et al., 2020) etc.
- ...



Extractive QA: SQuAD (Rajpurkar et al., 2016)

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

SQuAD data(Rajpurkar et al., 2016)

- Collection methodology: crowdsourced questions + answer spans, the writers see the full text (wikipedia excerpt)
- Challenges: unanswerable questions
- Dataset size: 100K answerable + 50K unanswerable questions

SQuAD status (Rajpurkar et al., 2016)

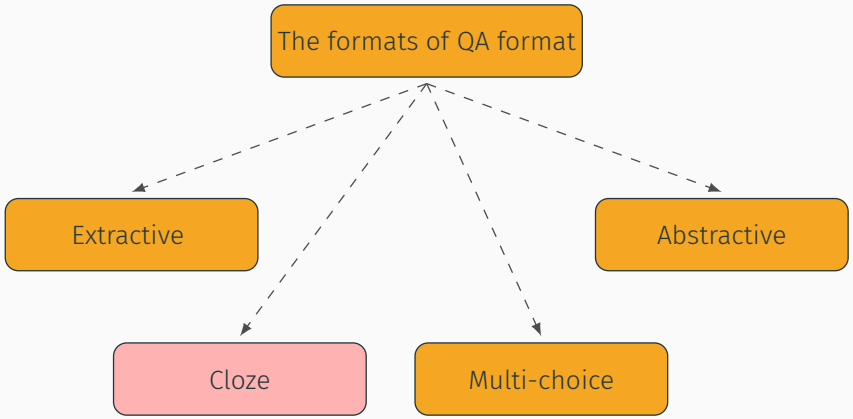
Leaderboard

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 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839

Cloze-style QA

- CBT (Hill et al., 2015a)
- CNN/Daily Mail (Hermann et al., 2015)
- WikiLinks Rare Entity (Long et al., 2017)
- BookTest (Bajgar et al., 2017)
- Who Did What (Onishi et al., 2016)
- CLOTH (Xie et al., 2018)
- ...



CNN/DailyMail: example (Hermann et al., 2015)

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

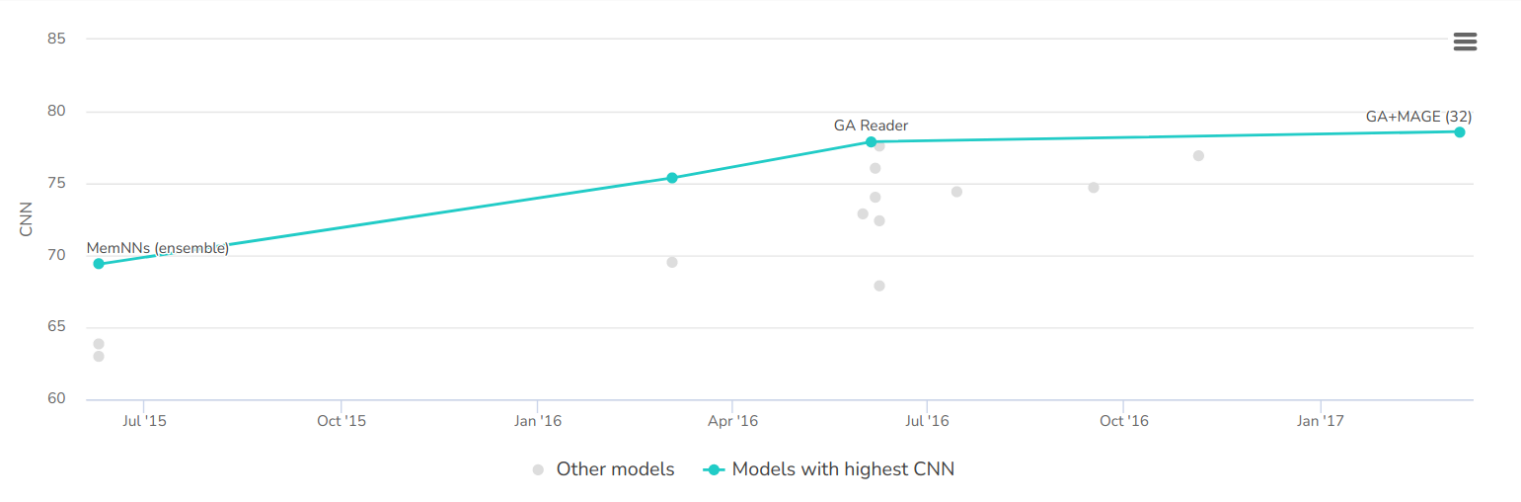
Table 3: Original and anonymised version of a data point from the Daily Mail validation set. The anonymised entity markers are constantly permuted during training and testing.

CNN/DailyMail: data (Hermann et al., 2015)

- Collection methodology: news articles were collected from news sites together with professional summaries, and sentences from summaries were converted to cloze questions
- Challenges: complex questions not biased by writers seeing the target text, not relying on world knowledge
- Dataset size: over 1M query-document-answer triplets

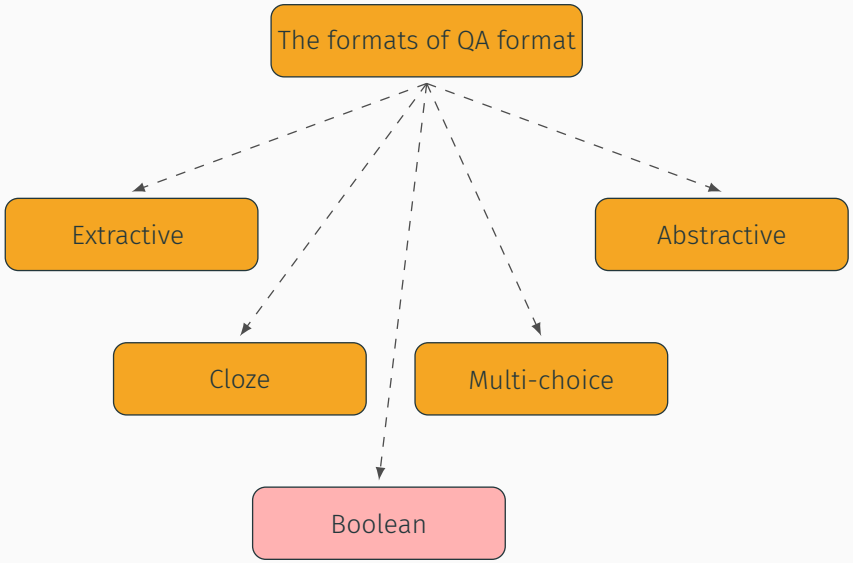
CNN/DailyMail: status (Hermann et al., 2015)

Human performance: n/a



Boolean QA

- BoolQ (Clark et al., 2019)
- ReCo (Chinese), Wang et al. (2020)
- partly: Natural Questions (Kwiatkowski et al., 2019), CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), HotPotQA (Yang et al., 2018a) and others



BoolQ: example (Clark et al., 2019)

Q:	Has the UK been hit by a hurricane?
P:	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A:	Yes. [An example event is given.]
Q:	Does France have a Prime Minister and a President?
P:	... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A:	Yes. [Both are mentioned, so it can be inferred both exist.]
Q:	Have the San Jose Sharks won a Stanley Cup?
P:	... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...
A:	No. [They were in the finals once, and lost.]

Figure 1: Example yes/no questions from the BoolQ dataset. Each example consists of a question (**Q**), an excerpt from a passage (**P**), and an answer (**A**) with an explanation added for clarity.

BoolQ: data (Clark et al., 2019)

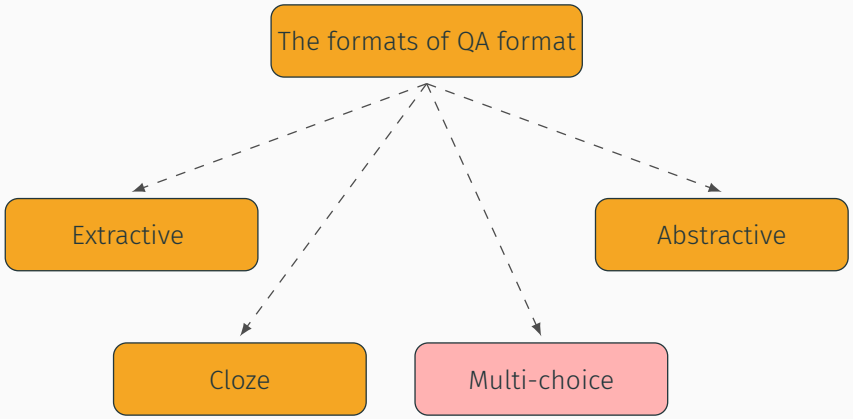
- Collection methodology: Google queries that are boolean questions are filtered, matched with wikipedia text and answered by crowdworkers
- Challenges: natural questions not biased by writers seeing the target text
- Dataset size: 16K

BoolQ: status (Clark et al., 2019)

	Rank	Name	Model	URL	Score	BoolQ
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0
+	2	T5 Team - Google	T5		89.3	91.2
+	3	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8
+	4	Alibaba PAI&ICBU	PAI Albert		86.1	88.1
+	5	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2
	6	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1
	7	Facebook AI	RoBERTa		84.6	87.1

Multi-choice QA

- RACE (Lai et al., 2017)
- ARC (Clark et al., 2018a)
- MCTest (Richardson et al., 2013)
- CLEF QA (Pe textasciitilde nas et al., 2014)
- QuAIL (Rogers et al., 2020)
- MultiRC (Khashabi et al., 2018)
- IJCNLP-2017 Task 5 (Chinese) Guo et al. (2017)
- ...



RACE: example (Lai et al., 2017)

Passage:
In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
"Really? How do you know that?" the gentleman said in surprise.
"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."
The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.
"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.
"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made _.
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because _ .
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that _ .
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by _ .
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made _ .
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

RACE: data (Lai et al., 2017)

- Collection methodology: Expert-written questions from English exams for middle/high school Chinese students
- Challenges: designed to test human comprehension
- Dataset size: 100K questions, 28K passages

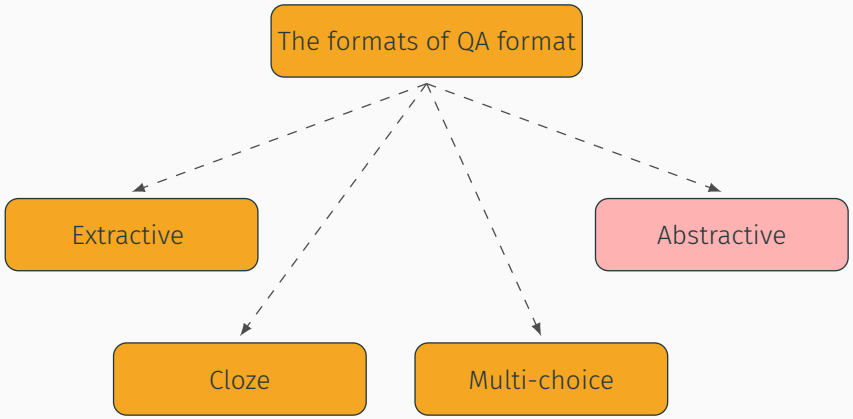
RACE: status (Lai et al., 2017)

Leaderboard

Model	Report Time	Institute	RACE	RACE-M	RACE-H
Human Ceiling Performance	Apr 15, 2017	CMU	*94.5*	*95.4*	*94.2*
Amazon Mechanical Turker	Apr 15, 2017	CMU	73.3	85.1	69.4
ALBERT-SingleChoice + transfer learning (ensemble)	Nov 06, 2020	Tencent Cloud Xiaowei & Tencent Cloud TI-ONE	**91.4**	**93.6**	**90.5**
Megatron-BERT (ensemble)	Mar 13, 2020	NVIDIA Research	90.9	93.1	90.0
ALBERT-SingleChoice + transfer learning	Nov 06, 2020	Tencent Cloud Xiaowei & Tencent Cloud TI-ONE	90.7	92.8	89.8
ALBERT + DUMA (ensemble)	Mar 18, 2020	SJTU & Huawei Noah's Ark Lab	89.8	92.6	88.7
Megatron-BERT	Mar 13, 2020	NVIDIA Research	89.5	91.8	88.6
ALBERT (ensemble)	Sep 26, 2019	Google Research & TTIC	89.4	91.2	88.6
UnifiedQA	May 02, 2020	AI2 & UW	89.4	-	-

Abstractive QA

- MS MARCO (Bajaj et al., 2016)
- CoQA (Reddy et al., 2018)
- MOCHA (Chen et al., 2020a)
- extractive and multi-choice datasets easily converted to freeform



MS MARCO: example (Bajaj et al., 2016)

Q: what is a corporation?

A: "A corporation is a company or group of people authorized to act as a single entity and recognized as such in law."

Context 1: "A company is incorporated in a specific nation, often within the bounds of a smaller subset of that nation, such as a state or province. The corporation is then governed by the laws of incorporation in that state. A corporation may issue stock, either private or public, or may be classified as a non-stock corporation. If stock is issued, the corporation will usually be governed by its shareholders, either directly or indirectly."

...

MS MARCO: data (Bajaj et al., 2016)

- Collection methodology: crowd workers wrote answers to real BING queries based on provided evidence (snippets)
- Challenges: real queries, noisy queries and text, answers not necessarily explicit in the evidence
- Dataset size: 100K (1M version released)

MC MARCO: status (Bajaj et al., 2016)

Human performance: ?

MS MARCO V1:RETIRED(12/01/2016-03/31/2018)

Rank	Model	Submission Date	Rouge-L	Bleu-1
1	**MARS** YUANFUDAO research NLP	March 26th, 2018	0.497	0.480
2	**Human Performance**	December 2016	0.470	0.460
3	**V-Net** Baidu NLP [Wang et al '18]	February 15th, 2018	0.462	0.445
4	**S-Net** Microsoft AI and Research [Tan et al. '17]	June 2017	0.452	0.438
5	**R-Net** Microsoft AI and Research [Wei et al. '16]	May 2017	0.429	0.422
6	**HieAttnNet** Akaitsuki	March 26th, 2018	0.423	0.448
7	**BiAttentionFlow+** ShanghaiTech University GeekPie_HPC team	March 11th, 2018	0.415	0.381
8	**ReasoNet** Microsoft AI and Research [Shen et al. '16]	April 28th, 2017	0.388	0.399

Still not done with types of QA data!

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)
- Dialogue: (Reddy et al., 2019; Choi et al., 2018)

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)
- Dialogue: (Reddy et al., 2019; Choi et al., 2018)
- Academic: (Clark et al., 2018b; Mihaylov et al., 2018)

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)
- Dialogue: (Reddy et al., 2019; Choi et al., 2018)
- Academic: (Clark et al., 2018b; Mihaylov et al., 2018)
- News: (Trischler et al., 2016; Hermann et al., 2015)

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)
- Dialogue: (Reddy et al., 2019; Choi et al., 2018)
- Academic: (Clark et al., 2018b; Mihaylov et al., 2018)
- News: (Trischler et al., 2016; Hermann et al., 2015)
- Biomedical: (Jin et al., 2019; Tsatsaronis et al., 2015)

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)
- Dialogue: (Reddy et al., 2019; Choi et al., 2018)
- Academic: (Clark et al., 2018b; Mihaylov et al., 2018)
- News: (Trischler et al., 2016; Hermann et al., 2015)
- Biomedical: (Jin et al., 2019; Tsatsaronis et al., 2015)
- Health: (Vilares and Gómez-Rodríguez, 2019; Suster and Daelemans, 2018)

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)
- Dialogue: (Reddy et al., 2019; Choi et al., 2018)
- Academic: (Clark et al., 2018b; Mihaylov et al., 2018)
- News: (Trischler et al., 2016; Hermann et al., 2015)
- Biomedical: (Jin et al., 2019; Tsatsaronis et al., 2015)
- Health: (Vilares and Gómez-Rodríguez, 2019; Suster and Daelemans, 2018)
- Fiction: (Kocisky et al., 2018a; Hill et al., 2015b)

QA: types of text

- Encyclopedia: (Rajpurkar et al., 2016; Yang et al., 2018b)
- Dialogue: (Reddy et al., 2019; Choi et al., 2018)
- Academic: (Clark et al., 2018b; Mihaylov et al., 2018)
- News: (Trischler et al., 2016; Hermann et al., 2015)
- Biomedical: (Jin et al., 2019; Tsatsaronis et al., 2015)
- Health: (Vilares and Gómez-Rodríguez, 2019; Suster and Daelemans, 2018)
- Fiction: (Kocisky et al., 2018a; Hill et al., 2015b)
- Multi-domain: QuAIL (Rogers et al., 2020), MRQA (Fisch et al., 2019), ORB (Dua et al., 2019a)

Types of reasoning

Unspecified

(most datasets)

Types of reasoning

Specialized

- **Coreference:** Quoref (Dasigi et al., 2019)

Unspecified

(most datasets)

Types of reasoning

Specialized

- **Coreference:** Quoref (Dasigi et al., 2019)
- **Temporal:** (Ning et al., 2020; Jia et al., 2018b,a)

Unspecified

(most datasets)

Types of reasoning

Unspecified

(most datasets)

Specialized

- Coreference: Quoref (Dasigi et al., 2019)
- Temporal: (Ning et al., 2020; Jia et al., 2018b,a)
- Numerical reasoning: (Dua et al., 2019b; Upadhyay and Chang, 2017; Miao et al., 2020)

Types of reasoning

Unspecified

(most datasets)

Specialized

- Coreference: Quoref (Dasigi et al., 2019)
- Temporal: (Ning et al., 2020; Jia et al., 2018b,a)
- Numerical reasoning: (Dua et al., 2019b; Upadhyay and Chang, 2017; Miao et al., 2020)
- Causality: (Lin et al., 2019)

Types of reasoning

Unspecified
(most datasets)

Specialized

- **Coreference:** Quoref (Dasigi et al., 2019)
- **Temporal:** (Ning et al., 2020; Jia et al., 2018b,a)
- **Numerical reasoning:** (Dua et al., 2019b; Upadhyay and Chang, 2017; Miao et al., 2020)
- **Causality:** (Lin et al., 2019)
- **Properties:** (Tafjord et al., 2019)

Types of reasoning

Is this still text-based QA?

Coreference

Temporal

Numerical reasoning

Causality

Properties

What counts as “commonsense reasoning"?

Working definition

Commonsense information: information that is commonly known and is thus not expected to be explicitly stated in the text

Fuzzy border between commonsense reasoning and RC

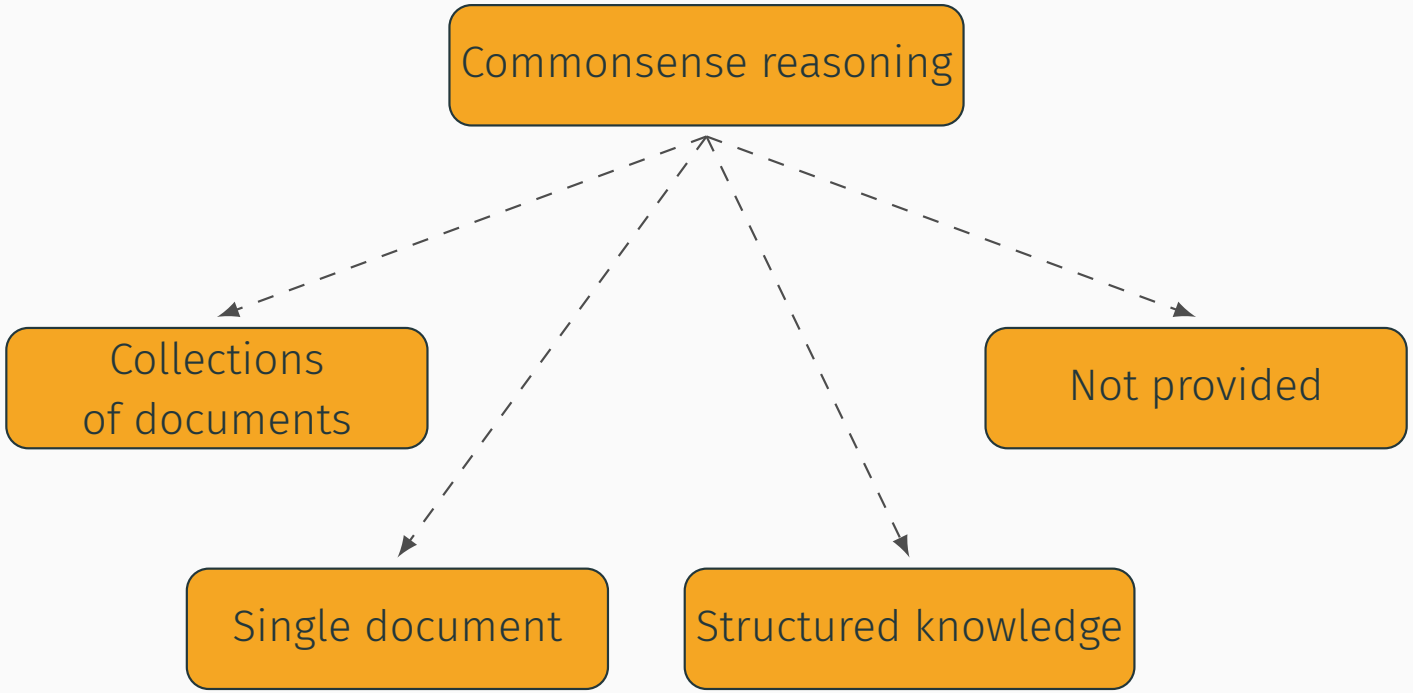
- cannot be defined in terms of types of information, because none are consistently stated or not
- e.g. temporal or causal information may or may not be stated

Fuzzy border between commonsense reasoning and RC

- cannot be defined in terms of types of information, because none are consistently stated or not
- e.g. temporal or causal information may or may not be stated

Context 1: John watched news.
Context 2: John watched news for half an hour.
Q: For how long did John watch the news?

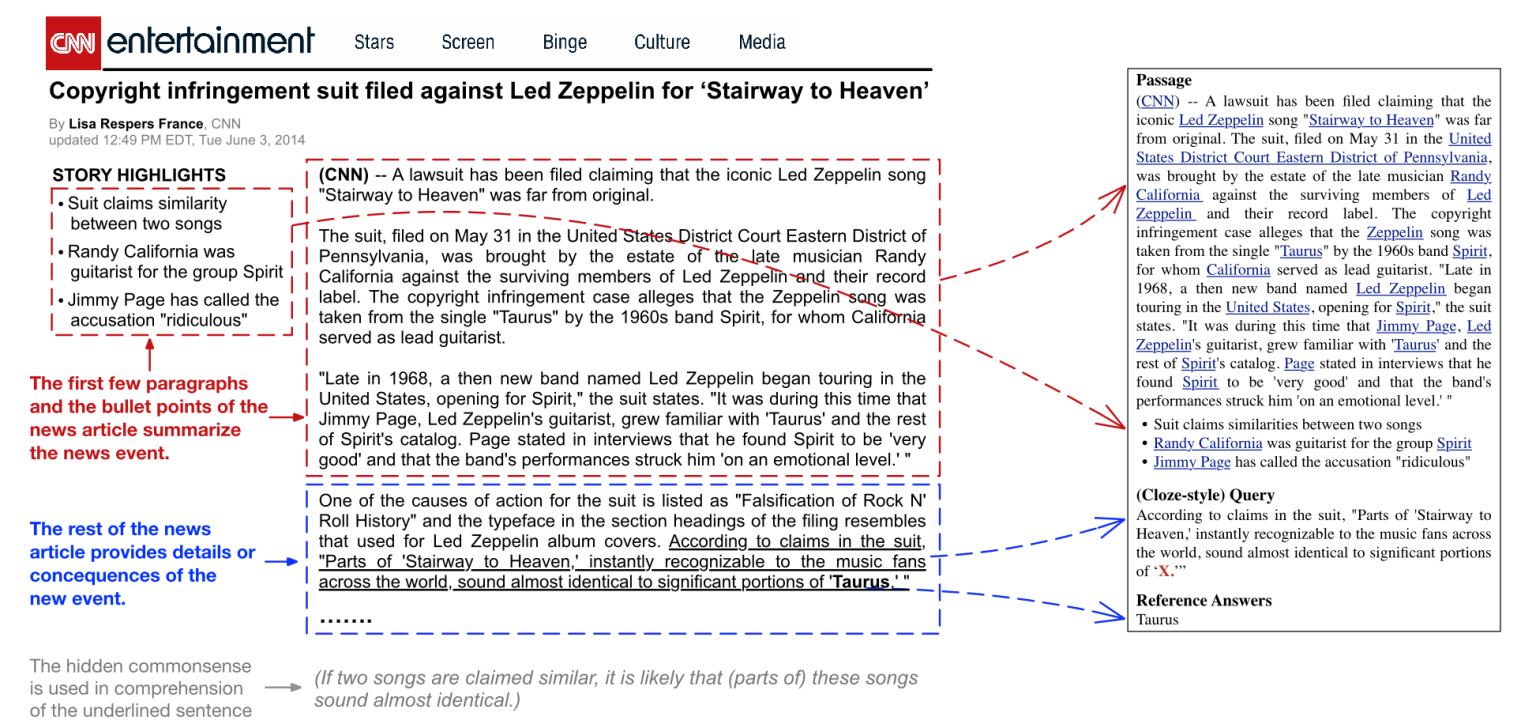
Sources of information in QA: also apply to commonsense reasoning



Commonsense knowledge sources

- ConceptNet (Speer et al., 2017)
- BabelNet (Navigli and Ponzetto, 2010)
- FrameNet (Baker et al., 1998)
- DeScript (?)
- ... pre-trained language models? (Cui et al., 2020)

Format 1: cloze (ReCoRD)



(Zhang et al., 2018)

ReCoRD: data

- Collection methodology: auto-generated from CNN/Daily mail dataset with manual filtering
- Challenges: completing cloze task only partially supported by the text
- Dataset size: 120K

ReCoRD: status

Human performance: 79.7% on the Wikipedia domain, and 75.4% on the web domain

Leaderboard

Rank	Model	EM	F1
	Human Performance *Johns Hopkins University* (Zhang et al. '18)	91.31	91.69
1 Mar 26, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP*	**90.64**	**91.21**
2 Jul 20, 2019	XLNet + MTL + Verifier (ensemble) *PingAn Smart Health & SJTU*	83.09	83.74
3 Jul 20, 2019	XLNet + MTL + Verifier (single model) *PingAn Smart Health & SJTU*	81.46	82.66
3 Jul 09, 2019	CSRLM (single model) *Anonymous*	81.78	82.58
4 Jul 24, 2019	{SKG-NET} (single model) *Anonymous*	79.48	80.04
5 Jan 11, 2019	KT-NET (single model) *Baidu NLP*	71.60	73.62
5 May 16, 2019	SKG-BERT (single model) *Anonymous*	72.24	72.78

Format 2: long context + question (MCScript)

T It was a long day at work and I decided to stop at the gym before going home. I ran on the treadmill and lifted some weights. I decided I would also swim a few laps in the pool. Once I was done working out, I went in the locker room and stripped down and wrapped myself in a towel. I went into the sauna and turned on the heat. I let it get nice and steamy. I sat down and relaxed. I let my mind think about nothing but peaceful, happy thoughts. I stayed in there for only about ten minutes because it was so hot and steamy. When I got out, I turned the sauna off to save energy and took a cool shower. I got out of the shower and dried off. After that, I put on my extra set of clean clothes I brought with me, and got in my car and drove home.

Q1 Where did they sit inside the sauna?
a. on the floor b. on a bench

Q2 How long did they stay in the sauna?
a. about ten minutes b. over thirty minutes

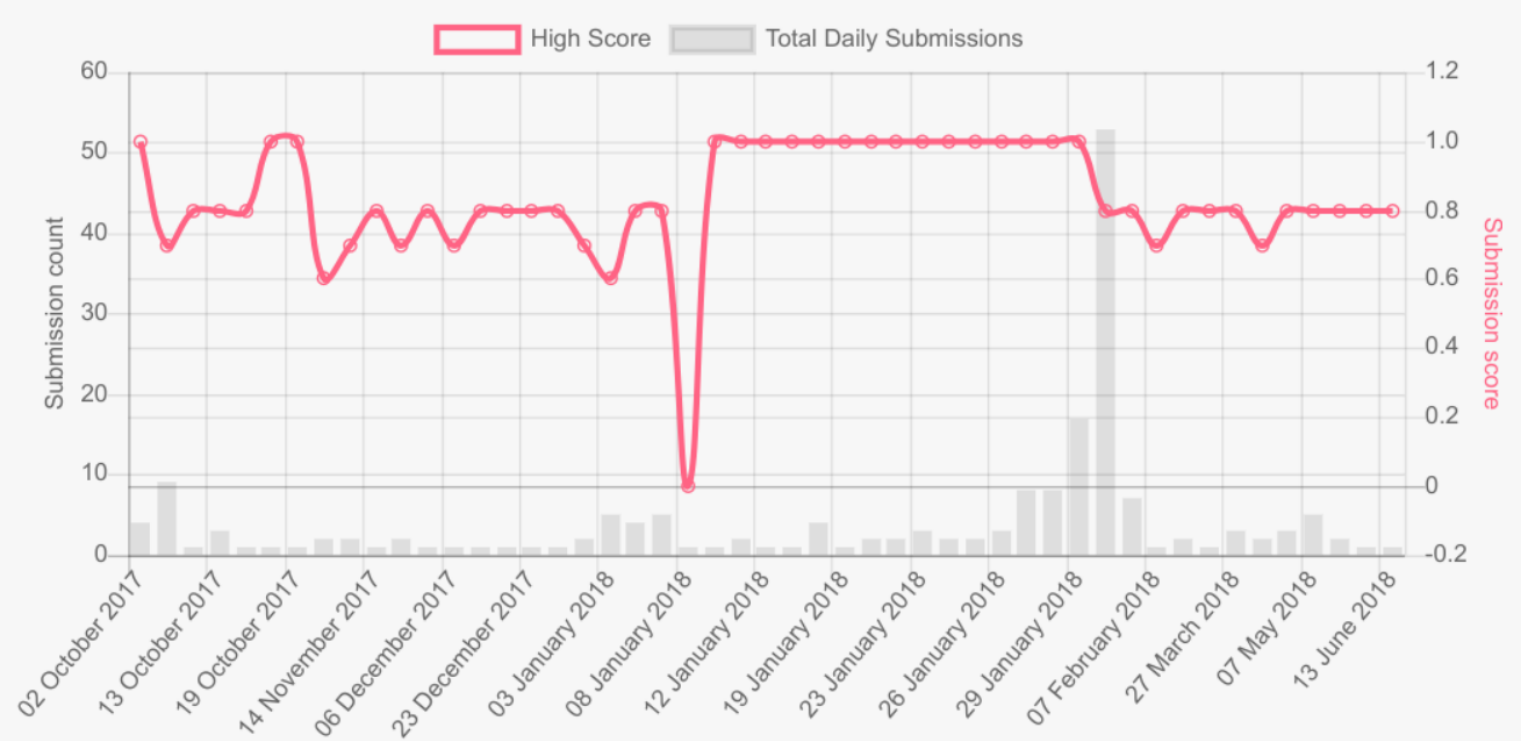
(Ostermann et al., 2018)

MCScript: data

- Collection methodology: (a) crowdsourced narrative, (b) separately collected questions about certain scripts, (c) writing correct and distractor answers
- Challenges: reasoning about everyday activities with script knowledge
- Dataset size: 14K questions

MCScript: status

Human performance: 98.%



Format 2: story completion (RocStories)

Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.

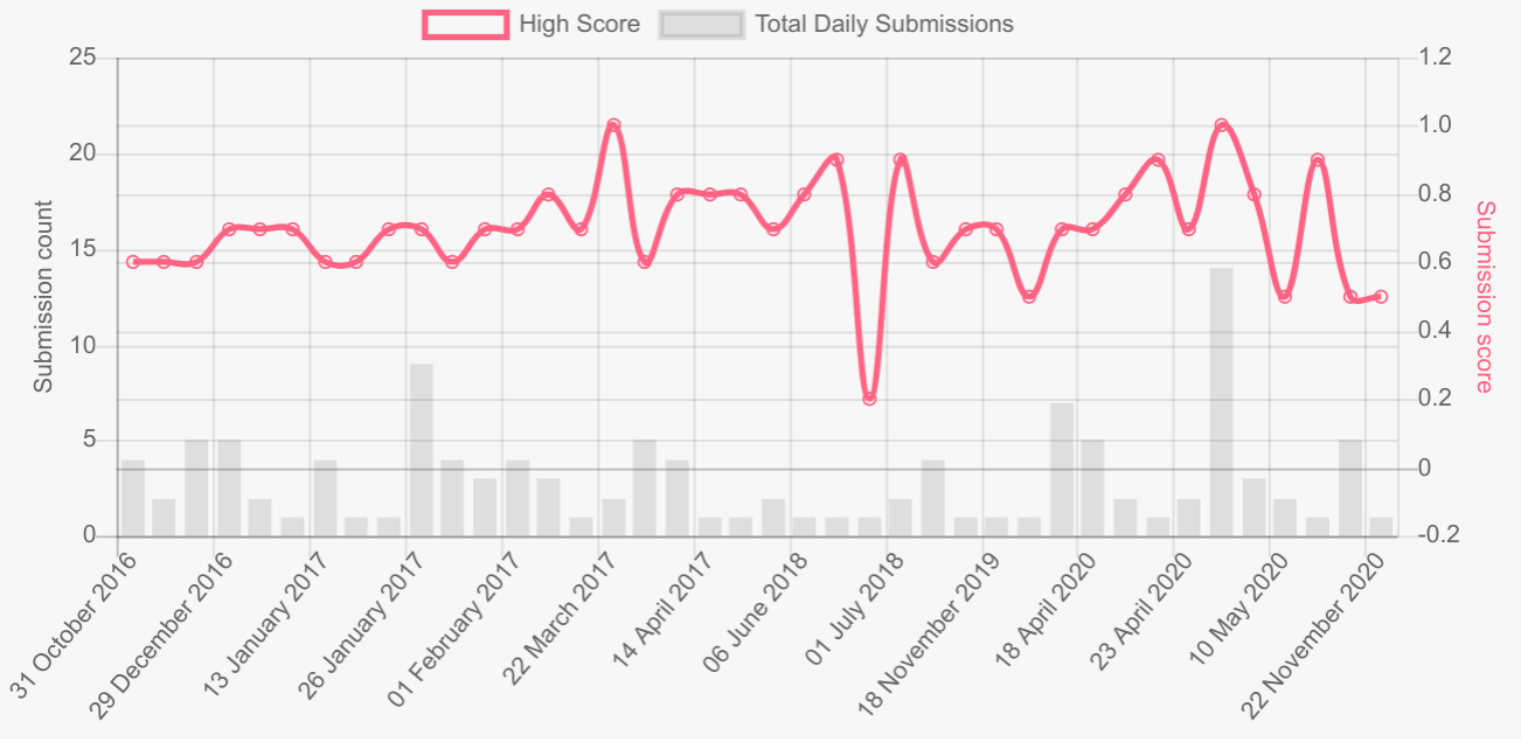
(Mostafazadeh et al., 2017; Sharma et al., 2018)

RocStories: data

- Collection methodology: (a) crowdsourced narratives, (b) asked crowdworkers to write a non-sensible ending with about the same number of words
- Challenges: reasoning about possible story endings
- Dataset size: 3K (release v.1.5)

RocStories: status

Human performance: 100%



Format 3: short context + ending, story completion-style (SWAG, Zellers et al. (2018)

A girl is going across a set of monkey bars. She a) jumps up across the monkey bars. b) struggles onto the monkey bars to grab her head. **c) gets to the end and stands on a wooden plank.** d) jumps up and does a back flip.
The woman is now blow drying the dog. The dog **a) is placed in the kennel next to a woman's feet.** b) washes her face with the shampoo. c) walks into frame and walks towards the dog. d) tried to cut her face, so she is trying to do something very close to her face.

Table 1: Examples from SWAG; the correct answer is **bolded**. Adversarial Filtering ensures that stylistic models find all options equally appealing.

(Zellers et al., 2018)

SWAG: data (Zellers et al., 2018)

- Collection methodology: sequential video captions are used as ground truth, subject of the second sentence is used to generate false endings with a language model, which are human-filtered
- Challenges: adversarially filtered distractor endings
- Dataset size: 113K

SWAG: status (Zellers et al., 2018)

	Human Performance		Accuracy: 0.8800

Download

Rank	Submission	Created	Accuracy
1	**DeBERTa** *Microsoft Dynamics 365 AI*	10/26/2020	**0.9171**
2	**ALUM** *Xiaodong Liu*	03/09/2020	0.9100
3	**RoBERTa** *Facebook AI*	07/19/2019	0.8992
4	**BigBird** *Pengcheng He, Weizhu Chen fro...*	05/16/2019	0.8706
5	**BERT (Bidirectional Encoder R...** *Jacob Devlin, Ming-Wei Chang,...*	10/12/2018	0.8628
6	**BERT-Large-Cased**	12/30/2019	0.8434

HellaSWAG: status (Zellers et al., 2019)

HellaSWAG is an update on SWAG with a more advanced adversarial filtering

Rank	Model	Overall accuracy	In-domain accuracy	Zero-shot accuracy	ActivityNet accuracy	WikiHow accuracy
	Human Performance *University of Washington* (Zellers et al. '19)	95.6	95.6	95.7	94.0	96.5
🥇 March 23, 2020	ALUM *MSR* https://github.com/namisan/mt-dnn	**85.6**	86.5	**84.6**	**77.1**	90.1
2 July 25, 2019	RoBERTa *Facebook AI*	85.2	**87.3**	83.1	74.6	**90.9**
3 February 7, 2020	G-DAug-inf *Anonymous*	83.7	85.6	81.8	73.0	89.6
4 January 19, 2020	HighOrderGN + RoBERTa *USC MOWGLI/INK Lab*	82.2	84.3	80.2	71.5	88.1
5 July 25, 2019	Grover-Mega *Univerity of Washington* https://rowanzellers.com/grover	75.4	79.1	71.7	64.8	81.2
6 July 25, 2019	Grover-Large *Univerity of Washington* https://rowanzellers.com/grover	57.2	60.7	53.6	53.3	59.2

Format 3: short context + question: twin sentences (?)

- The trophy doesn't fit in the brown suitcase because it's too big. What is too big?

  Answer 0: the trophy
  Answer 1: the suitcase

- The trophy doesn't fit in the brown suitcase because it's too small. What is too small?

  Answer 0: the trophy
  Answer 1: the suitcase

WinoGRANDE: data (Sakaguchi et al., 2019)

	Twin sentences	Options (**answer**)
✗	The monkey loved to play with the balls but ignored the blocks because he found **them** *exciting*.	**balls** / blocks
	The monkey loved to play with the balls but ignored the blocks because he found **them** *dull*.	balls / **blocks**
✗	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *weak*.	**William** / Jason
	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *strong*.	William / **Jason**
✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *less* time to get ready for school.	**Robert** / Samuel
	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *more* time to get ready for school.	Robert / **Samuel**
✓	The child was screaming after the baby bottle and toy fell. Since the child was *hungry*, **it** stopped his crying.	**baby bottle** / toy
	The child was screaming after the baby bottle and toy fell. Since the child was *full*, **it** stopped his crying.	baby bottle / **toy**

Table 2: Examples that have *dataset-specific* bias detected by AFLITE (marked with ✗). The words that include (dataset-specific) polarity bias (§3) are highlighted (positive and negative). For comparison, we show examples selected from WINOGRANDE$_{debiased}$ (marked with ✓).

WinoGRANDE: data

- Collection methodology: crowdsourcing + adversarial filtering based on embedding associations
- Challenges: systematic bias reduction with AFLite
- Dataset size: 44K

WinoGRANDE: status

	Human Performance						AUC: 0.9400

Download

Rank	Submission	Created	AUC	Acc (XS)	Acc (S)	Acc (M)	Acc (L)
1	**UNICORN** *Anonymous*	07/28/2020	**0.8664**	**0.7923**	**0.8359**	**0.8732**	**0.9038**
2	**UnifiedQA (T5,11B) - finetuned** *AI2*	05/31/2020	0.8571	0.7878	0.8336	0.8687	0.8851
3	**TTTTT** *University of Waterloo and Ac...*	03/13/2020	0.7673	0.6825	0.7051	0.7759	0.8240
4	**Roberta-large + G-DAug-Combo** *anonymous*	02/24/2020	0.7146	0.6106	0.6712	0.7119	0.7736

Further confusion with "inference"

- ”grounded commonsense inference” of SWAG (Zellers et al., 2018)

  On stage, a woman takes a seat at the piano. She
  a) sits on a bench as her sister plays with the doll.
  b) smiles with someone as the music plays.
  c) is in the crowd, watching the dancers.
  d) nervously sets her fingers on the keys.

Further confusion with "inference"

- ”grounded commonsense inference” of SWAG (Zellers et al., 2018)

  On stage, a woman takes a seat at the piano. She
  a) sits on a bench as her sister plays with the doll.
  b) smiles with someone as the music plays.
  c) is in the crowd, watching the dancers.
  d) nervously sets her fingers on the keys.

- commonsense reasoning datasets recast as NLI, e.g. WNLI (??)

  Premise: *I couldn’t put the pot on the shelf because it was too tall.*
  Hypothesis: *The pot was too tall.*
  Label: *entailed*

CHALLENGE: bridging RC and commonsense

(Rogers et al., 2020)

QuAIL: data (Rogers et al., 2020)

The Bear (Michael E. Shea)

The air exploded in a flash of bone and steel and blood. The clash of metal rang through the forest. An arrow pierced through the darkness, its barbed head tearing through flesh and muscle. A roar echoed off of the mountains far to the west. A cry broke through soon after. Then silence.
Char stood over a pile of black fur and red blood. He held a curved sword, jagged half way down the wide blade and hilted in bone. He held a large thick bow in the other. Lorfel and Ranur stood behind him, panting. Lorfel, a short man of twenty six held a large axe in both hands and still prepared to swing it hard. Ranur, the largest of the three held a pike in one hand, its tip hanging low towards the ground. He buried his other hand in his gray tunic.
"Did it get either of you?" Char's voice rasped low in the silence of the night.
"No" Lorfel said. He planted his axe head on the ground with a thud and leaned on the tall handle. There was a pause. Char turned towards Ranur.
"Are you hurt?"
"Mm...My hand." Ranur took his hand out of his tunic. Moonlight gleamed red off of the ragged wound. Char thought he saw a glimmer of bone.
"Did he claw you or bite you?" Char's voice held an urgency that set both Lorfel and Ranur on edge.
Ranur paused and then spoke low. "He bit me."
Char picked Lorfel and Ranur as his hunting partners for their speed and sharpness in battle. They had hunted beasts of the deep woods all of their lives. They hunted the beasts that hunted men. They all knew the risks of battling such creatures. The old man dropped his curved sword, drew his bow, and fired. The arrow hammered into Ranur's chest, burying itself in his heart. Lorfel saw the gleaming arrow head sticking almost a foot out of his companion's back. Ranur fell face first to the ground.

Text-based questions

Q: When did the roar happen? — Temporal order

A. before the cry
B. after the silence
C. NEI*
D. when Char was speaking

* Not enough information to answer this question

Q: Who bit Ranur? — Coreference

A. the beast
B. Lorfel
C. Char
D. NEI

Q: What color was the beast's fur? — Factual questions

A. brown
B. NEI
C. black
D. red

Unanswerable questions

Q: What was done with Ranur's body?

A. burned to avoid spreading disease
B. left abandoned along with the beasts' corpse
C. buried in the ground
D. NEI

World knowledge questions

Q: Why was there blood? — Causality

A. because Char shot something
B. NEI
C. because Lorfel had an axe
D. because Char had a sword

Q: After the end of this text, Ranur is: — Subsequent state

A. standing up
B. NEI
C. on the ground
D. in the sky

Q: Ranur probably died: — Event duration

A. a month later
B. instantly
C. NEI
D. a year later

Q: What is probably true about the beast's bite? — Properties

A. it is harmless
B. it is extremely dangerous
C. NEI
D. it helps people

Q: Who was concerned about his companions' injuries? — Belief states

A. NEI
B. Char
C. Lorfel
D. Ranur

QuAIL: data (Rogers et al., 2020)

- Collection methodology: crowdsourcing specific reasoning types with keyword-based checks
- Challenges: balanced across 9 reasoning types and 4 domains, full range of uncertainty
- Dataset size: 14K

QuAIL: status (Rogers et al., 2020)

Leaderboard (Last updated: 13 Dec 20 08:10 EST)

Rank	All	Temp.	Caus.	Fact.	Char.	Ent.	Belief	Sub.	Dur.	Unans.	Team	Submitted	Model
1	**53.4**	53.3	61.2	62.1	42.9	55.4	58.8	53.3	62.9	30.8	matt.downey18	31 Oct 20 09:02 EDT	TML BERT Baseline
2	**31.1**	34.2	29.2	35.4	33.3	26.2	25.8	25.0	25.8	45.0	matt.downey18	30 Oct 20 20:16 EDT	TML PMI Baseline

Full range of uncertainty: trouble with human evaluation

Question type	All questions	Text+ Unanswerable		World knowledge	
Temporal	0.66	0.67		--	--
Coreference	0.7	0.79		--	--
Factual	0.75	0.82		--	--
Causality	0.76	--	--	0.86	
Subsequent	0.53	--	--	0.62	
Duration	0.32	--	--	0.37	
Properties	0.67	--	--	0.78	
Beliefs	0.62	--	--	0.85	
Unanswerable	0.25	0.83		--	--
All questions	0.6	0.78		0.7	

(Rogers et al., 2020)

Natural Language Inference (Anna Rumshisky)

Outline

CoQA: new datasets get “solved" immediately!

Leaderboard

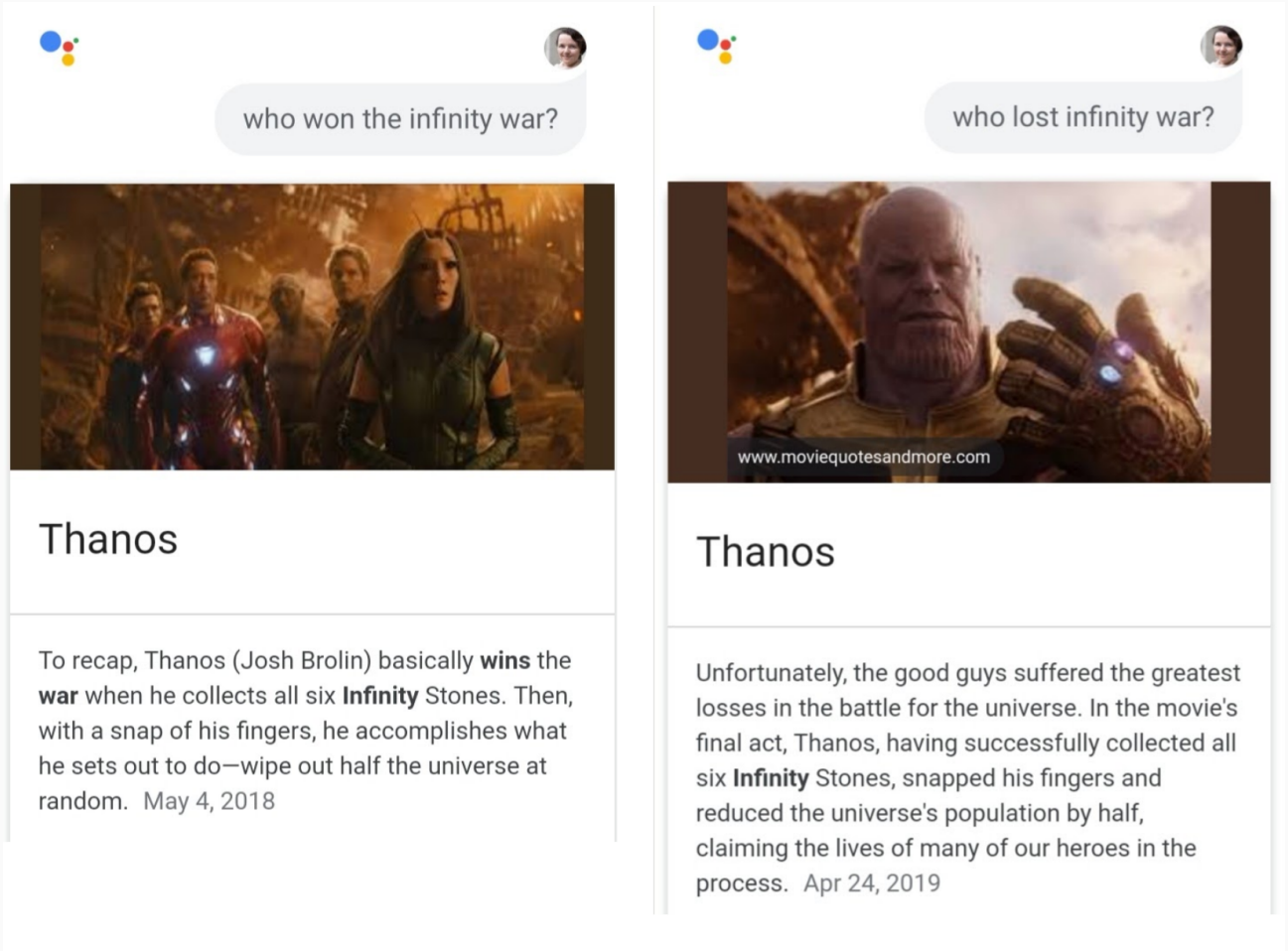
Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* (Reddy & Chen et al. TACL '19)	89.4	87.4	88.8
1 Sep 05, 2019	RoBERTa + AT + KD (ensemble) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	91.4	**89.2**	**90.7**
1 Apr 22, 2020	TR-MT (ensemble) *WeChatAI*	**91.5**	88.8	**90.7**
2 Sep 05, 2019	RoBERTa + AT + KD (single model) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	90.9	**89.2**	90.4
3 Jan 01, 2020	TR-MT (ensemble) *WeChatAI*	91.1	87.9	90.2
4 Mar 29, 2019	Google SQuAD 2.0 + MMFT (ensemble) *MSRA + SDRG*	89.9	88.0	89.4
5 Dec 18, 2019	TR-MT (single model) *WeChatAI*	90.4	86.8	89.3
6 Sep 13, 2019	XLNet + Augmentation (single model)	89.9	86.9	89.0

SuperGLUE: new datasets get “solved" immediately!

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
✚	2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
✚	3	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
✚	4	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
✚	5	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	6	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	7	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
✚	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		84.0	86.1	93.2/95.2	91.2	80.1/45.9	89.9/89.3	87.9	71.3	89.0	35.2	93.8/68.8
✚	9	Timo Schick	iPET (ALBERT) - Few-Shot (32 Examples)		75.4	81.2	79.9/88.8	90.8	74.1/31.7	85.9/85.4	70.8	49.3	88.4	36.2	97.8/57.9
	10	Adrian de Wynter	Bort (Alexa AI)		74.1	83.7	81.9/86.4	89.6	83.7/54.1	49.8/49.0	81.2	70.1	65.8	48.0	96.1/61.5

(Wang et al., 2019)

However...


who won the infinity war?
Thanos
To recap, Thanos (Josh Brolin) basically **wins** the **war** when he collects all six **Infinity** Stones. Then, with a snap of his fingers, he accomplishes what he sets out to do—wipe out half the universe at random. May 4, 2018
who lost infinity war?
www.moviequotesandmore.com
Thanos
Unfortunately, the good guys suffered the greatest losses in the battle for the universe. In the movie's final act, Thanos, having successfully collected all six **Infinity** Stones, snapped his fingers and reduced the universe's population by half, claiming the lives of many of our heroes in the process. Apr 24, 2019

Are our models that good, or our data that bad?

(Heinzerling, 2019)

Performance issues

EVIDENCE: lack of basic linguistic capabilities

Rychalska et al. (2018) swapped verbs in SQuAD questions with their antonyms:

Original question	Question with verb antonym
Q: How many teams **participate** in the Notre Dame Bookstore Basketball tournament?	**Q:** How many teams **drop out** in the Notre Dame Bookstore Basketball tournament?
Q: Which art museum **does** Notre Dame administer?	**Q:** Which art museum **doesn't** Notre Dame administer?

In 90.5% cases DrQA model prediction didn't change!

EVIDENCE: lack of basic linguistic capabilities

Rychalska et al. (2018) swapped verbs in SQuAD questions with their antonyms:

Original question	Question with verb antonym
Q: How many teams **participate** in the Notre Dame Bookstore Basketball tournament?	**Q:** How many teams **drop out** in the Notre Dame Bookstore Basketball tournament?
Q: Which art museum **does** Notre Dame administer?	**Q:** Which art museum **doesn't** Notre Dame administer?

In 90.5% cases DrQA model prediction didn't change!

Spoiler alert:

BERT doesn't "understand" negation either (Ettinger, 2020).

EVIDENCE: easily distracted

Jia and Liang (2017) added adversarial distractor sentences to SQuAD texts

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

The accuracy of 16 published models drops from an avg of 75% F1 to 36% (down to 7% with ungrammatical adversarial sentences).

EVIDENCE: position bias in extractive QA

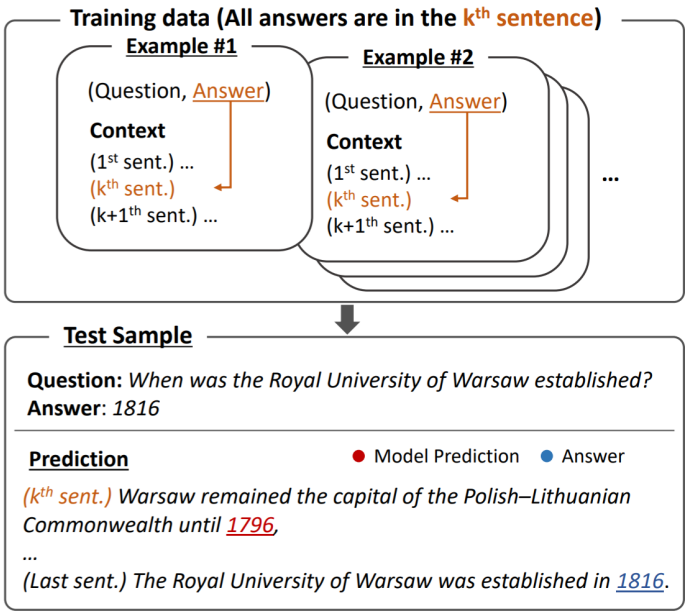


Figure 1: Example of position bias. BERT trained on the dataset with a skewed answer position distribution, provides wrong predictions, biased to the specific sentence position.

BERT trained on a biased training set (every answer in the first sentence) achieves 37.48% F1 on dev, and the same amount of randomly sampled examples achieves 85.06% F1 (Ko et al., 2020).

EVIDENCE: insensitivity to corrupted inputs

Original context

[...] By now you have probably heard about Chris Ulmer, the 26-year-old teacher in Jacksonville, Florida, who starts his special education class by calling up each student individually to give them much admiration and a high-five. I couldn't help but be reminded of Syona's teacher and how she supports each kid in a very similar way. Ulmer recently shared a video of his teaching experience. All I could think was: how lucky these students are to have such inspirational teachers. [...]

Context with shuffled context words

[...] their with and to kids combined , t always of (has) mean problems the palsy five cerebral that communication , her standard " assess (. teacher a a now gesture Florida admiration and , much calling Ulmer to individually (of class his heard Jacksonville year special you up Chris greeting five) congratulation by give education who , them or about probably the in by each - student high , old - - have starts 26 . I s she similar reminded be ' each t and in help ' kid teacher [...]

Question

What can we learn about Chris Ulmer?

Options (the answer is in bold)

(A) He praises his students one by one. (B) He is Syona's favorite teacher. (C) He use videos to teach his students. (D) He asks his students to help each other.

Figure 2: Example of questions with shuffled context words from RACE. Although the question appears unsolvable for humans, the baseline model predicts the correct answer.

10 tested datasets can be "solved" with jumbled texts, and without numbers, pronouns, logical and causal terms (Sugawara et al., 2020).

EVIDENCE: model-independent attacks Wallace et al. (2019)

Question: Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a ..... **why how because to kill american people.**	crime and poverty → to kill american people

72% of "why" questions in SQuAD: the prediction changes to "to kill american people"!

EVIDENCE: don't really need NLU!

Simple context/type matching heuristic: F1 of 78.1 on SQuAD!

***When** did building activity occur on St. Kazimierz Church?*

Building activity occurred in numerous noble palaces and churches [...]. One of the best examples [..] are Krasinski Palace (1677-1683), Wilanow Palace (1677-1696) and St. Kazimierz Church (**1688-1692**)

(Weissenborn et al., 2017)

EVIDENCE: don't really need NLU!

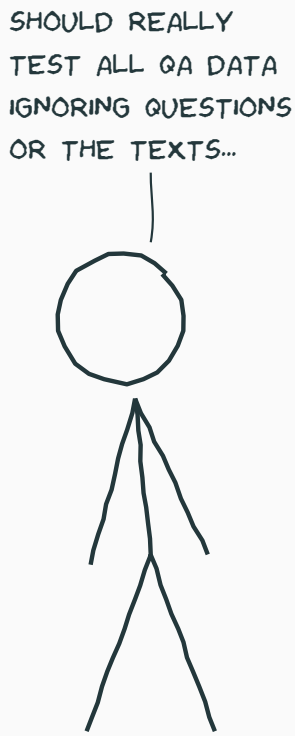
Much of NLI data can be solved without even looking at premises! (Poliak et al., 2018)

Dataset	DEV Hyp-Only	MAJ	|Δ|
DPR	50.21	50.21	0.00
SPR	86.21	65.27	+20.94
FN+	62.43	56.79	+5.64
ADD-1	75.10	75.10	0.00
SciTail	66.56	50.38	+16.18
SICK	56.76	56.76	0.00
MPE	40.20	40.20	0.00
JOCI	61.64	57.74	+3.90
SNLI	69.17	33.82	+35.35
MNLI-1	55.52	35.45	+20.07
MNLI-2	55.18	35.22	+19.96

EVIDENCE: don't really need NLU!

Much of NLI data can be solved without even looking at premises! (Poliak et al., 2018)

Dataset	DEV Hyp-Only	MAJ	|Δ|
DPR	50.21	50.21	0.00
SPR	86.21	65.27	+20.94
FN+	62.43	56.79	+5.64
ADD-1	75.10	75.10	0.00
SciTail	66.56	50.38	+16.18
SICK	56.76	56.76	0.00
MPE	40.20	40.20	0.00
JOCI	61.64	57.74	+3.90
SNLI	69.17	33.82	+35.35
MNLI-1	55.52	35.45	+20.07
MNLI-2	55.18	35.22	+19.96



see Sugawara et al. (2020)

EVIDENCE: multi-hop questions not multi-hop Min et al. (2019)

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?
Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of the Bonobo apes.
Paragraph 1: The bonobo ("Pan paniscus"), formerly called the **pygmy chimpanzee** and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus "Pan".

The only paragraph about an animal

HotpotQA: a single-hop BERT-based RC model achieves F1 of 67 , comparable to SOTA multi-hop models!

Outline

What's wrong with our data?

The models, unable to discern
the intentions of the data sets designers,
happily recapitulate any statistical patterns
they find in the training data.
(Linzen, 2020)

Annotation artifacts

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%

Table 4: Top 5 words by PMI($word, class$), along with the proportion of *class* training samples containing *word*. MultiNLI is abbreviated to MNLI.

(Gururangan et al., 2018)

Syntactic homogeneity

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

4 neural systems including BERT drop to under 15% accuracy (McCoy et al., 2019b)

Testing IR or reasoning?

Set	Question	Answer	Rationale
Training	Name this sociological phenomenon, the *taking of one's own life*.	Suicide	Paraphrase
Challenge	Name this *self-inflicted method of death*.	Arthur Miller	
Training	Clinton played the *saxophone on The Arsenio Hall Show*	Bill Clinton	Entity Type Distractor
Challenge	He was edited to appear in the film "Contact"... For ten points, name this American president who played the *saxophone on an appearance on the Arsenio Hall Show*.	Don Cheadle	

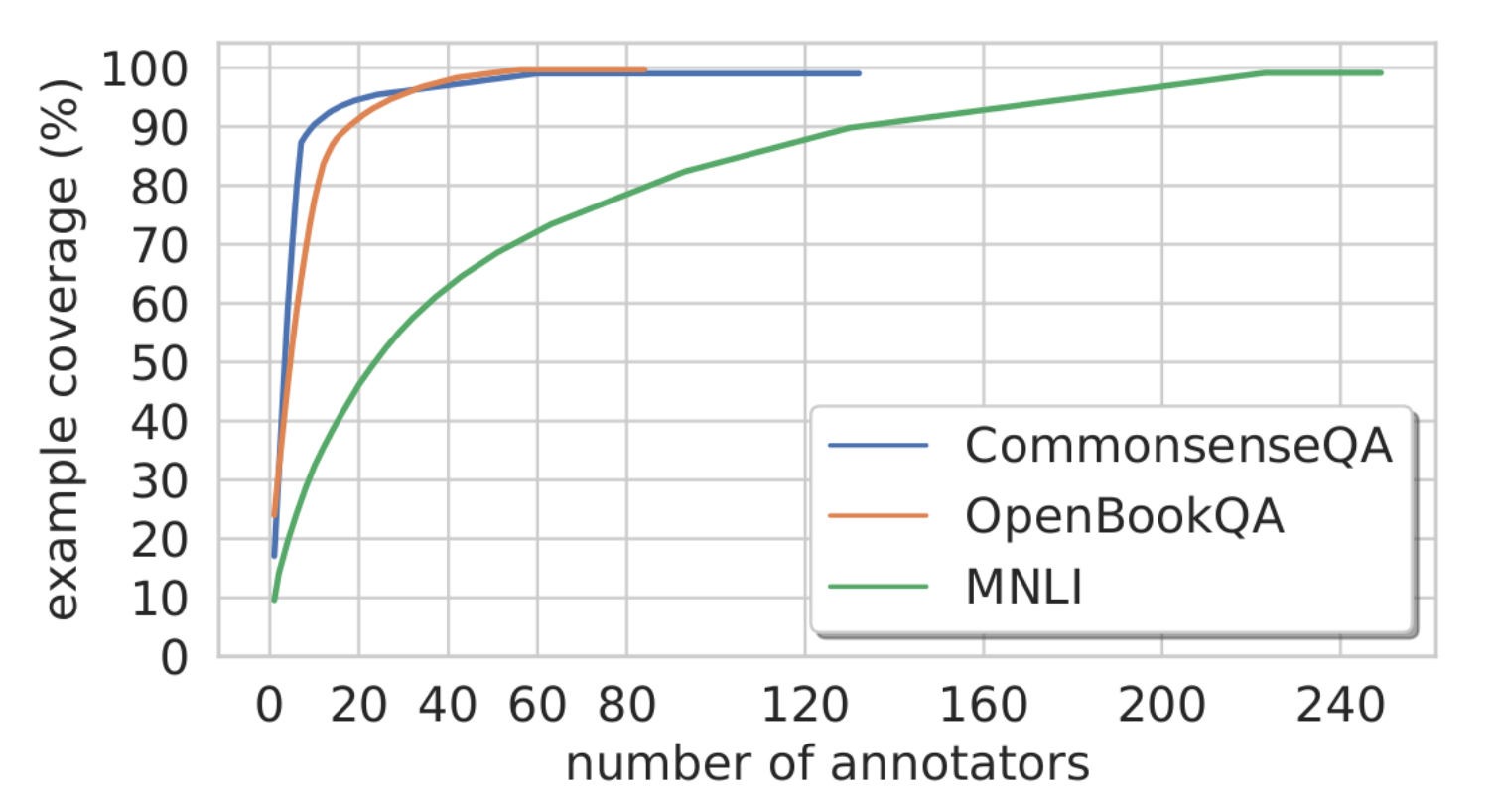
(Wallace and Boyd-Graber, 2018)

Testing IR or reasoning?

Question	Prediction	Answer	Rationale
This man, who died at the Battle of the Thames, experienced a setback when his brother Tenskwatawa's influence over their tribe began to fade	Battle of Tippecanoe	Tecumseh	Triangulation
This number is one hundred and fifty more than the number of Spartans at Thermopylae.	Battle of Thermopylae	450	Operator
A building dedicated to this man was the site of the "I Have A Dream" speech	Martin Luther King Jr.	Abraham Lincoln	Multi-Step Reasoning

(Wallace and Boyd-Graber, 2018)

Annotator biases



BERT does not generalize to examples generated by unseen annotators! 23 point drop on multi-annotator vs random split on OpenBookQA, 10 on CommonsenseQA, 5 on MNLI (Geva et al., 2019)

Social biases

- like other NLP data, QA/NLI data may contain statistical patterns with undesirable social implications

[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

Social biases

- like other NLP data, QA/NLI data may contain statistical patterns with undesirable social implications
- the source data is already biased: e.g. most factoid datasets are based on Wikipedia, which contains fewer and shorter pages about women[1]

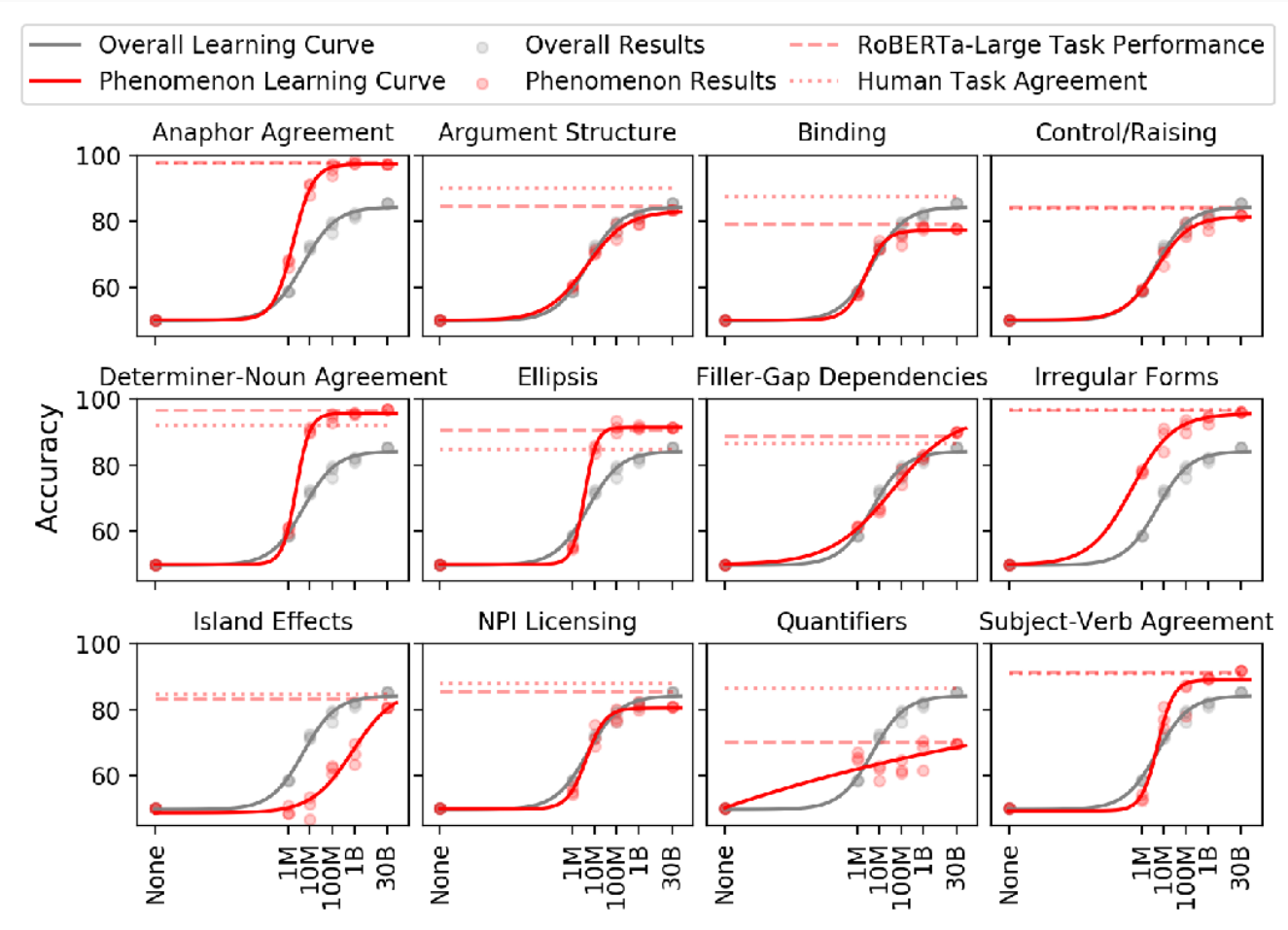
[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

Social biases

- like other NLP data, QA/NLI data may contain statistical patterns with undesirable social implications
- the source data is already biased: e.g. most factoid datasets are based on Wikipedia, which contains fewer and shorter pages about women[1]
- research on bias reduction is budding (Sun et al., 2019), but current techniques are limited

[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

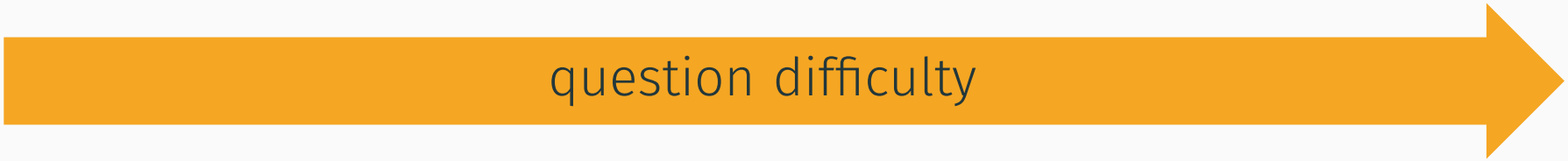
Long tail phenomena



It takes RoBERTa a lot longer to learn some linguistic phenomena (Zhang et al., 2020b)

Solution 1: Quality over quantity!

Discriminative questions



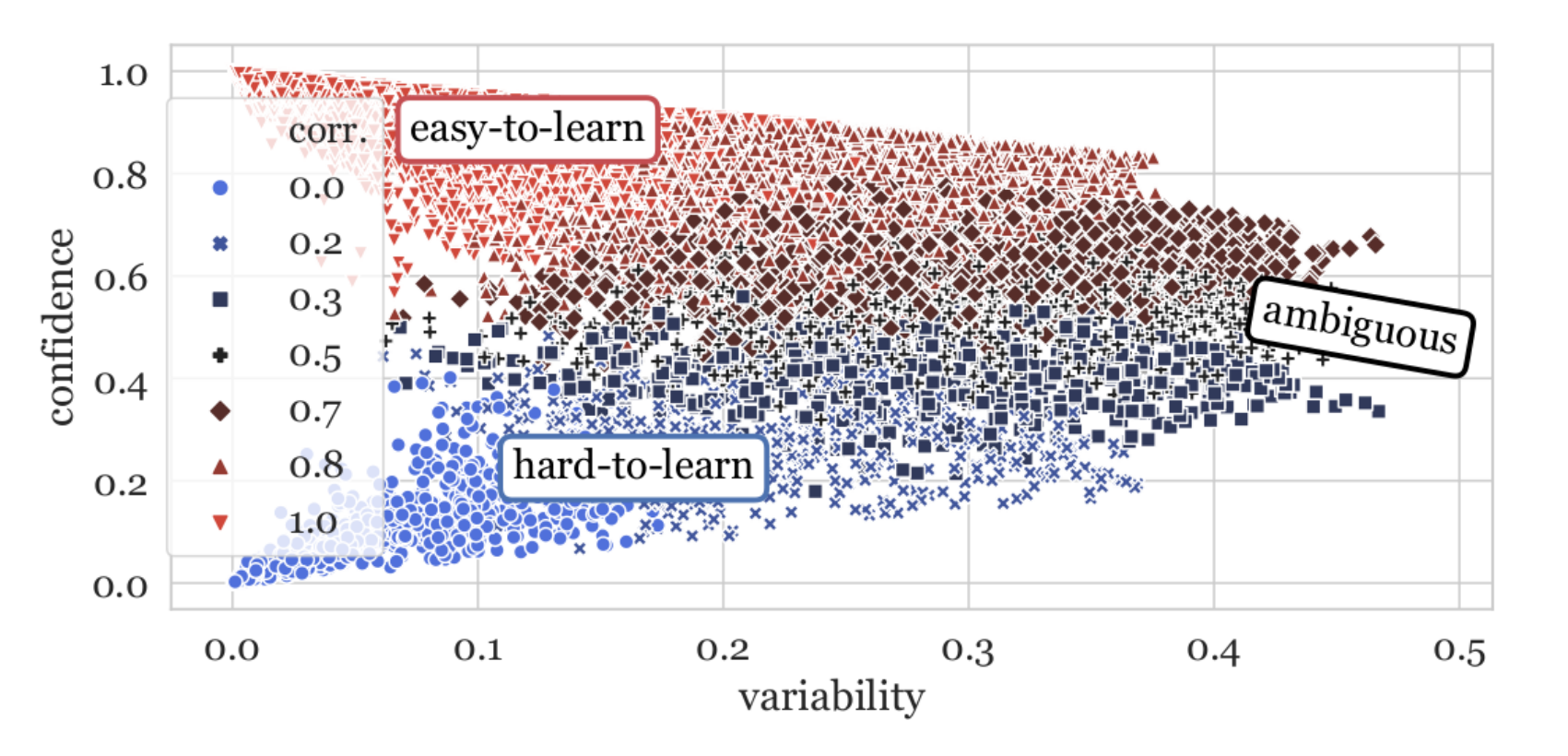
Too easy:
What is the capital of Poland?

Discriminative:
Who lives in 221B and uses Vicodin?

Too difficult:
What was the cause of the US civil war?

The discriminative questions should be as error-free as possible! (Boyd-Graber, 2019)

Do we need datasets to be bigger or better?

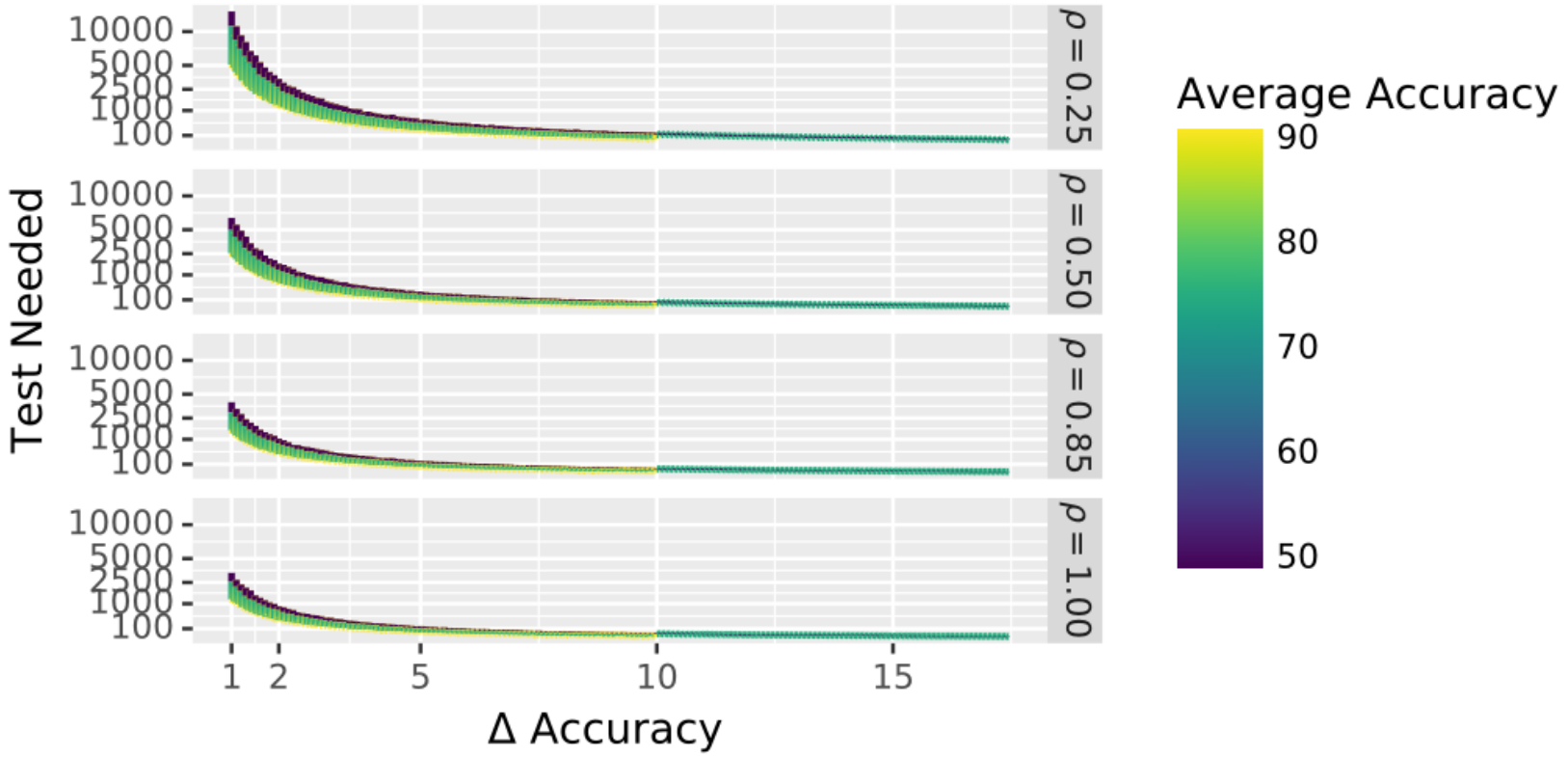


SNLI dataset, based on RoBERTa-large classifier (Swayamdipta et al., 2020)

Confidence: confidence in the true class

Variability: changes in confidence during training

How big should the test datasets be?



Test data size depends on (a) the difference in accuracy between the systems, (b) their avg accuracy (closer to 50% is harder), and (c) the amount of discriminative questions. If 100% questions are disciminative, 2.5K is enough even at 1% Δ with 50% average accuracy. If only 25% is discriminative, we need 15K. (Boyd-Graber, 2019)

Solution 2: Diversifying the data

Paraphrasing

- question-first question collection (Kwiatkowski et al., 2019)

Paraphrasing

- question-first question collection (Kwiatkowski et al., 2019)
- writing questions without seeing the target text
  text (Kocisky et al., 2018b)

Paraphrasing

- question-first question collection (Kwiatkowski et al., 2019)
- writing questions without seeing the target text
  text (Kocisky et al., 2018b)
- paraphrasing existing questions (Rogers et al., 2020)

Adversarial question authoring: requiring questions the model can't answer

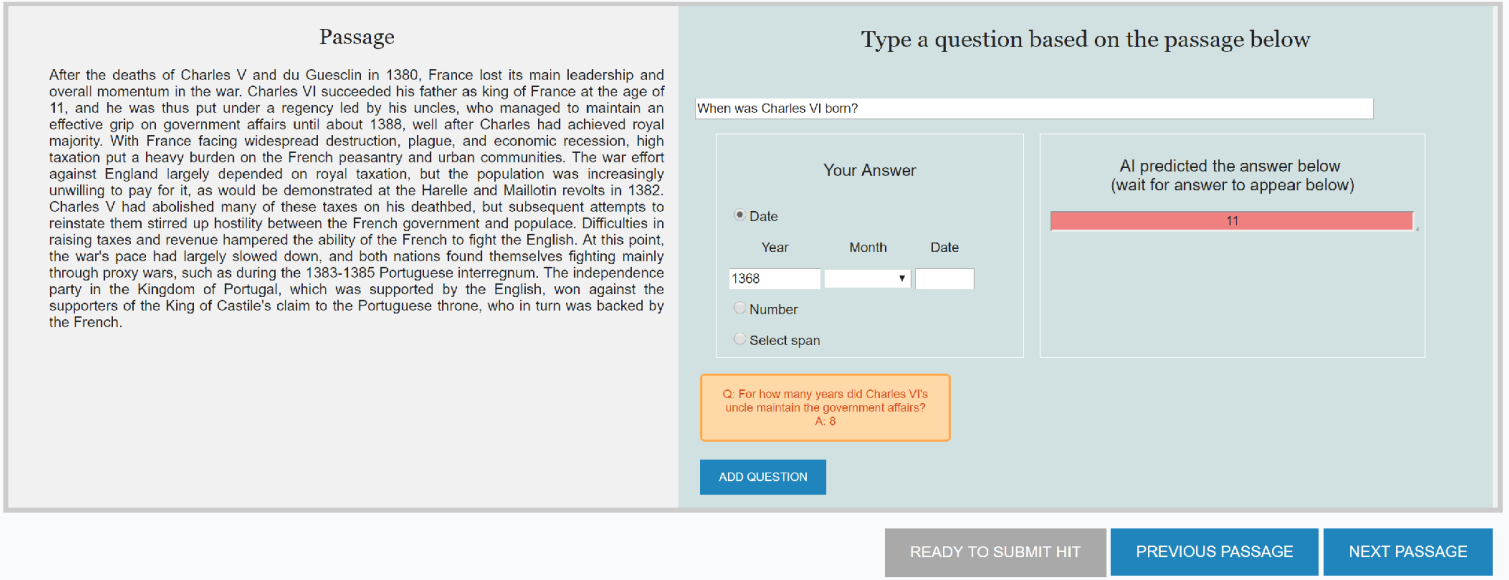
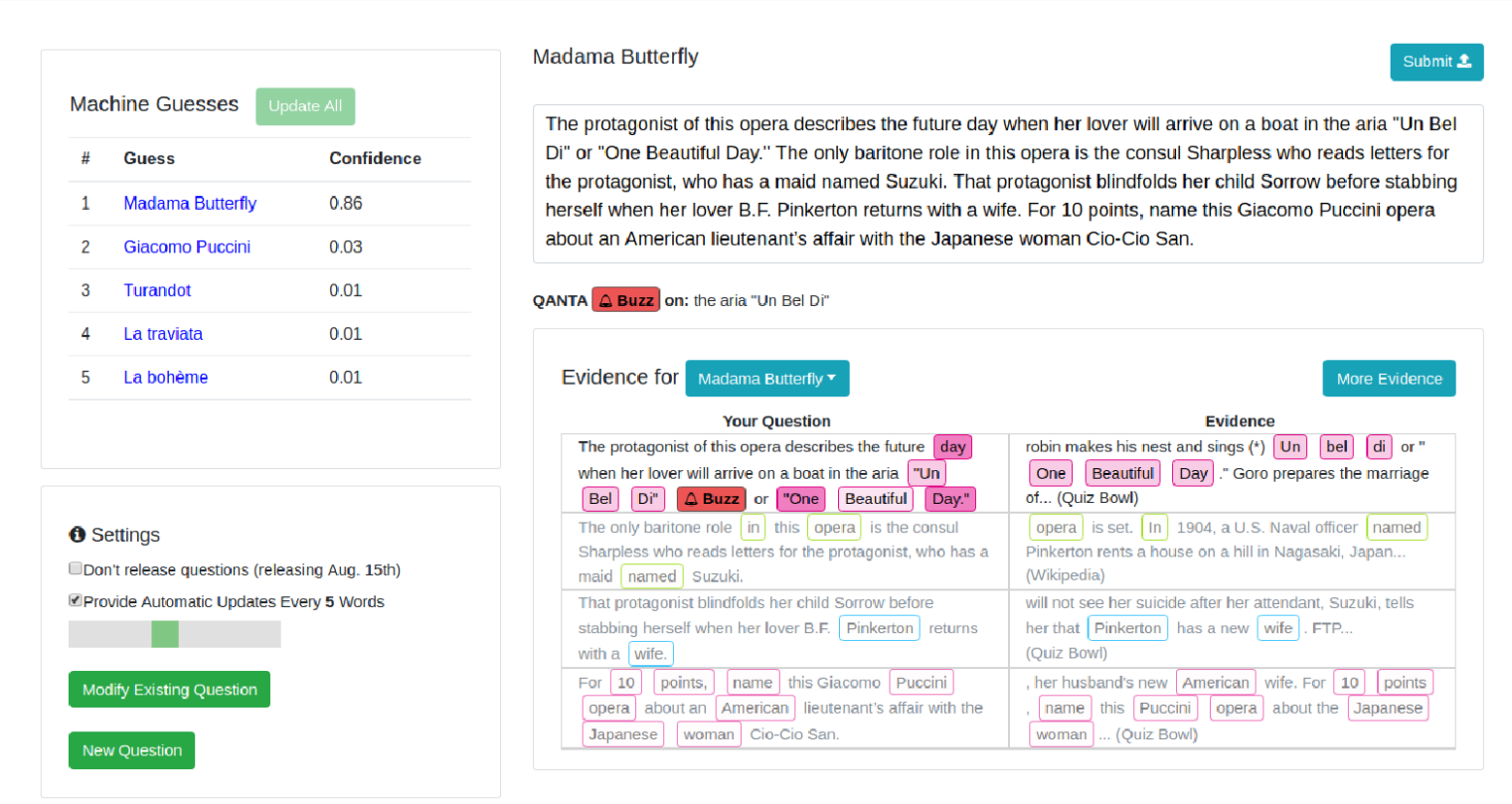


Figure 2: Question Answering HIT sample with passage on the left and input fields for answer on the right

(Dua et al., 2019b)

Adversarial question authoring: answer + explanation



(Wallace and Boyd-Graber, 2018)

Diversity by design: synthetic (Weston et al., 2015)

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions
John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

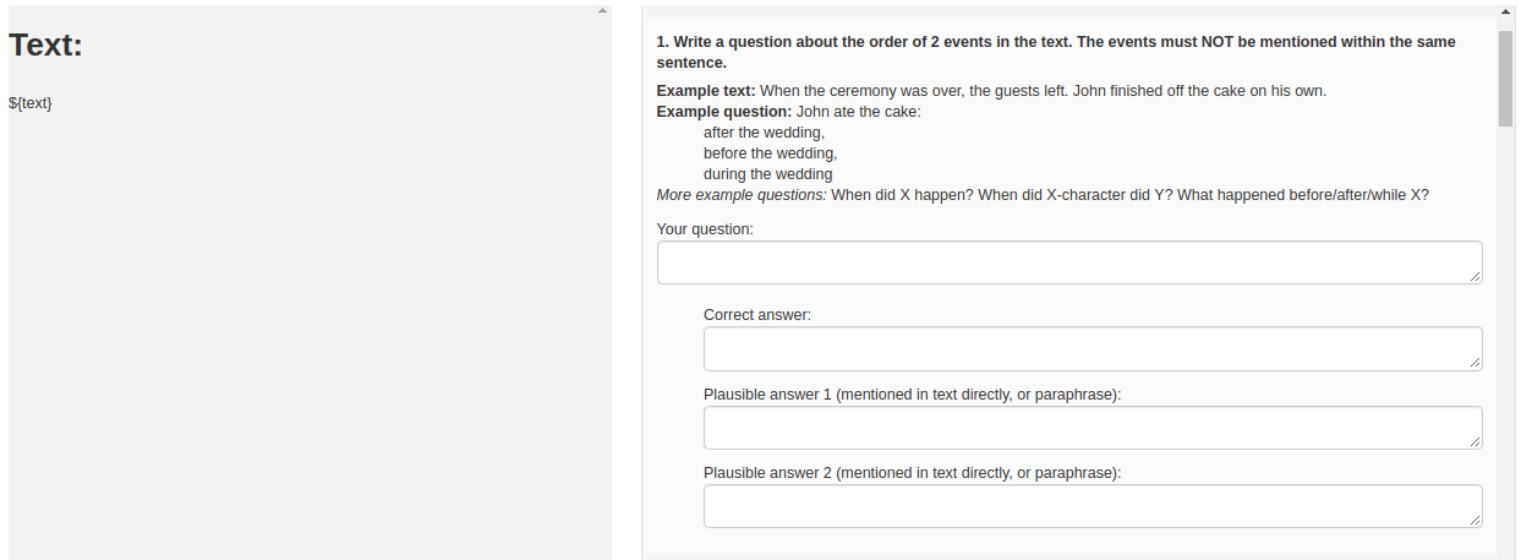
Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets
Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge
John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

Diversity by design: crowdsourced (Rogers et al., 2020)



Solution 3: More difficult types of reasoning

Combining information from several sources

S1: Most young mammals, including humans, play. S2: Play is how they learn the skills that they will need as adults. S6: Big cats also play. S8: At the same time, they also practice their hunting skills. S11: Human children learn by playing as well. S12: For example, playing games and sports can help them learn to follow rules. S13: They also learn to work together.
What do human children learn by playing games and sports? A)* They learn to follow rules and work together B) hunting skills C)* skills that they will need as adult

Figure 1: Examples from our MultiRCcorpus. Each example shows relevant excerpts from a paragraph; multi-sentence question that can be answered by combining information from multiple sentences of the paragraph; and corresponding answer-options. The correct answer(s) is indicated by a *. Note that there can be multiple correct answers per question.

(Khashabi et al., 2018)

Queries that require logical or numerical operations

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (31.2%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (20.4%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (18.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker
Addition (12%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992	2 March 1992
Count (16%) and Sort (8.8%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay	Matt Prater
Coreference Resolution (4%)	**James Douglas** was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before **1543 he married Elizabeth**, daughter of James Douglas, 3rd Earl of Morton. **In 1553 James Douglas succeeded to the title and estates of his father-in-law**.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10	1553

(Dua et al., 2019b)

Unanswerable questions: know when you don't know

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*

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.

(Rajpurkar et al., 2018)

Full range of uncertainty: text-based + guessable + unanswerable

(Rogers et al., 2020)

Solution 4: different levels of difficulty

Multiple "views" of the same benchmark

We cannot control for the kinds of reasoning that the model employs, but we can control what data it has access to.

Multiple "views" of the same benchmark

We cannot control for the kinds of reasoning that the model employs, but we can control what data it has access to.

- adding/removing metadata for coreference, semantic parses, disambiguation et. (Boyd-Graber, 2019)

Multiple "views" of the same benchmark

We cannot control for the kinds of reasoning that the model employs, but we can control what data it has access to.

- adding/removing metadata for coreference, semantic parses, disambiguation et. (Boyd-Graber, 2019)
- “ablate” parts or structure of the input (Sugawara et al., 2020)

Multiple "views" of the same benchmark

We cannot control for the kinds of reasoning that the model employs, but we can control what data it has access to.

- adding/removing metadata for coreference, semantic parses, disambiguation et. (Boyd-Graber, 2019)
- “ablate” parts or structure of the input (Sugawara et al., 2020)
- settings with/without adversarial distractors (Yang et al., 2018a)

Solution 5: Multi-step quality control

Multi-step quality control (Quizbowl)

NLP style:

- write the questions
- check for answerability

QuizBowl style: (Boyd-Graber, 2019)

- the question is written
- subject editor: removing ambiguity, clarifying acceptable answers, making the question more discriminative
- head editor: diversity of the question set, uniform difficulty, repeats
- post-mortem error analysis

Solution 6: fixing the incentives

Crowdworker incentives

- optimizing for reward is known to be detrimental to performance and creativity (Englmaier et al., 2018)

Crowdworker incentives

- optimizing for reward is known to be detrimental to performance and creativity (Englmaier et al., 2018)
- gamifying the crowdsourcing, making the tasks enjoyable (Boyd-Graber et al., 2012)

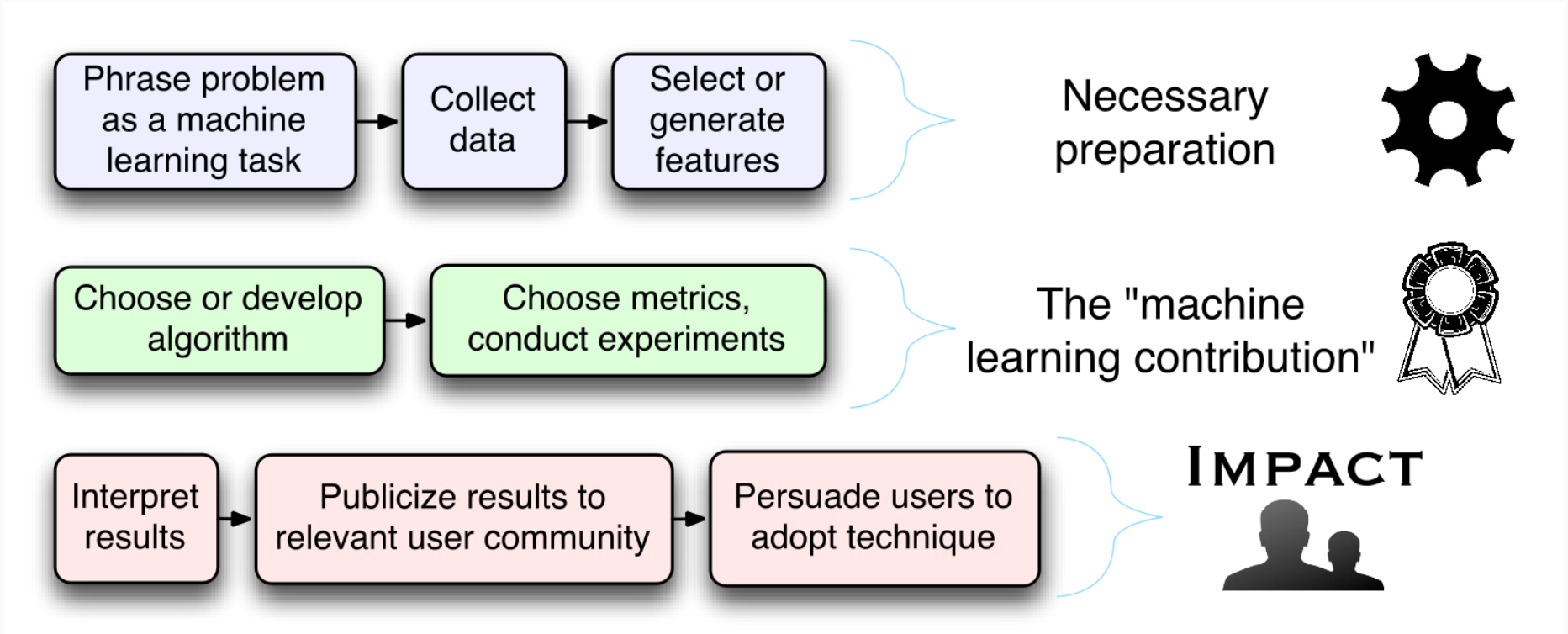
Crowdworker incentives

- optimizing for reward is known to be detrimental to performance and creativity (Englmaier et al., 2018)
- gamifying the crowdsourcing, making the tasks enjoyable (Boyd-Graber et al., 2012)
- leveraging communities of enthusiasts (Boyd-Graber, 2019)

Crowdworker incentives

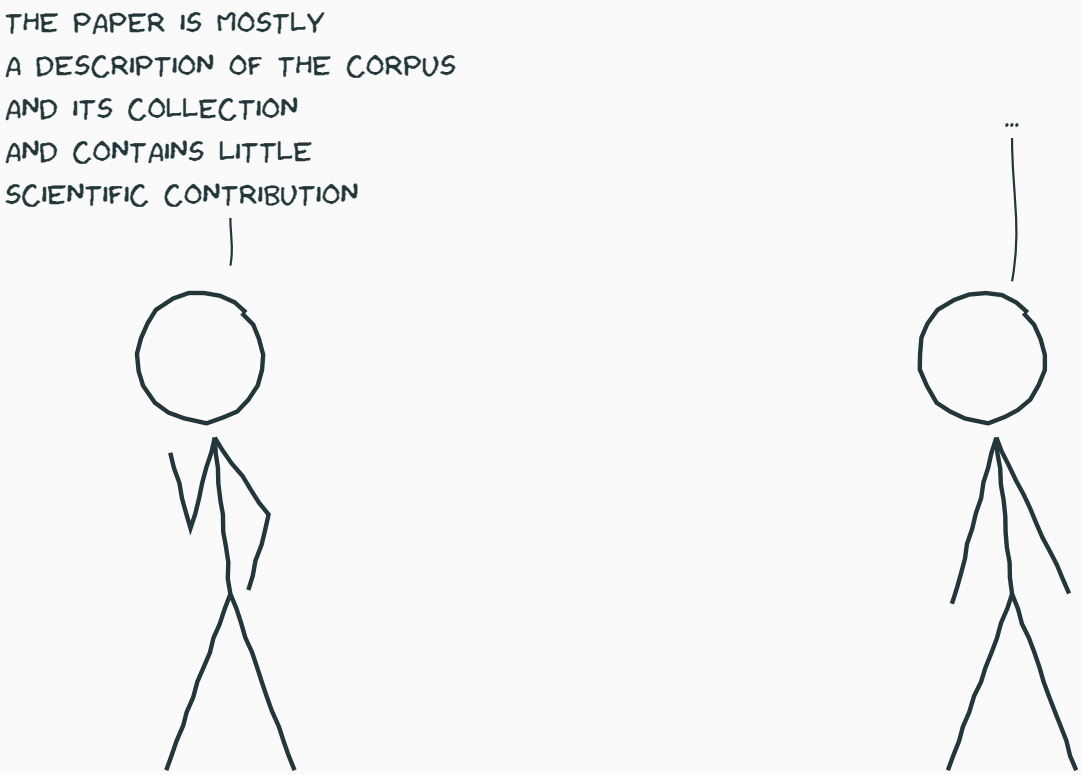
- optimizing for reward is known to be detrimental to performance and creativity (Englmaier et al., 2018)
- gamifying the crowdsourcing, making the tasks enjoyable (Boyd-Graber et al., 2012)
- leveraging communities of enthusiasts (Boyd-Graber, 2019)
- some evidence of Hawthorn effect (Rogers et al., 2020)

Dataset author incentives



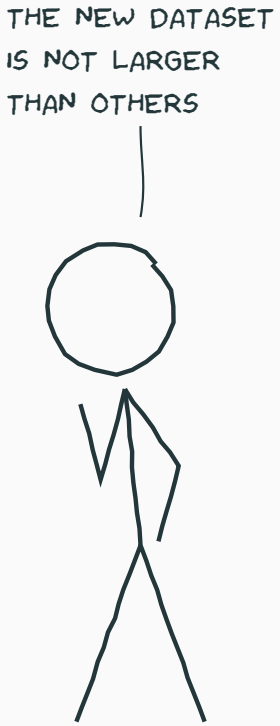
(Wagstaff, 2012)

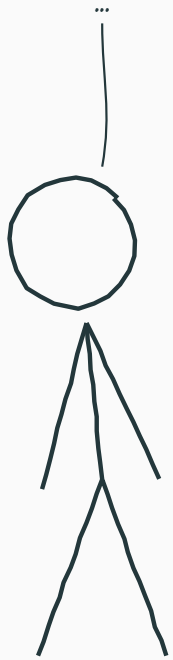
Reviewer 2 and resource papers



(Bawden, 2019; Rogers, 2020)

Reviewer 2 and resource papers



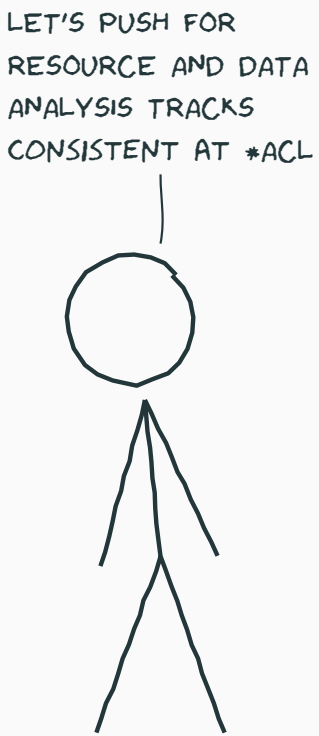


(Bawden, 2019; Rogers, 2020)

Emerging trend: data analysis!

- Zhang et al. (2020a): "WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge"
- Boratko et al. (2018): "A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset"
- Yatskar (2019): "A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC"
- Yue et al. (2020): "Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset"
- Chen et al. (2016): "A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task"
- ...

Reviewer 2 and resource papers



Solution 7: diagnostic data

Type 1: challenge datasets for diagnosing heuristics

- adversarial SQuAD (Jia and Liang, 2017)
- HANS (McCoy et al., 2019b)
- ...

Type 1: challenge datasets for diagnosing heuristics

- adversarial SQuAD (Jia and Liang, 2017)
- HANS (McCoy et al., 2019b)
- ...

Problem: there may be other heuristics that can go unnoticed

Type 2: testing specific linguistic capabilities

- adversarial SNLI with lexical knowledge (Glockner et al., 2018)
- RC requiring coreference (Dasigi et al., 2019)
- ...

Type 2: testing specific linguistic capabilities

- adversarial SNLI with lexical knowledge (Glockner et al., 2018)
- RC requiring coreference (Dasigi et al., 2019)
- ...

Problem: other linguistic capabilities may be missing

Type 3: batteries of tests

- multi-task (?Wang et al., 2019)
- multi-domain (Rogers et al., 2020; Dua et al., 2019a)
- multiple types of reasoning (Rogers et al., 2020; Dua et al., 2019a)
- multiple linguistic capabilities (Ribeiro et al., 2020)
- ...

Type 3: batteries of tests

- multi-task (?Wang et al., 2019)
- multi-domain (Rogers et al., 2020; Dua et al., 2019a)
- multiple types of reasoning (Rogers et al., 2020; Dua et al., 2019a)
- multiple linguistic capabilities (Ribeiro et al., 2020)
- ...

Problem: when do we have enough?

When is the benchmark enough?

Spatial *(sample entries)*:

- Rover is in the yard from when he runs out the door until he runs inside.
- Rover is in the house from when he runs inside until the end of the story.

Temporal *(sample entries)*:

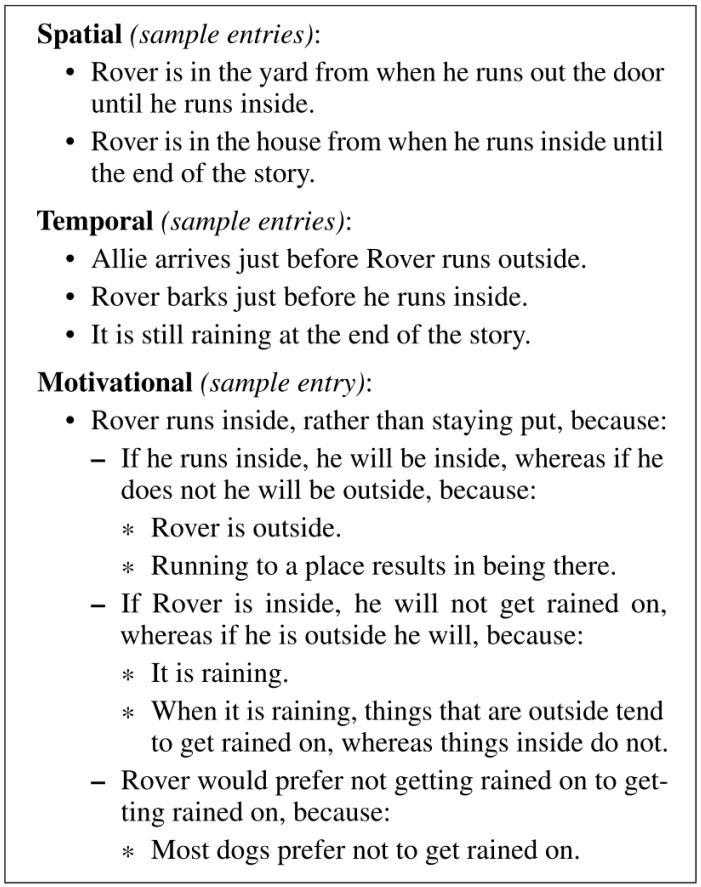
- Allie arrives just before Rover runs outside.
- Rover barks just before he runs inside.
- It is still raining at the end of the story.

Motivational *(sample entry)*:

- Rover runs inside, rather than staying put, because:
  - If he runs inside, he will be inside, whereas if he does not he will be outside, because:
    - Rover is outside.
    - Running to a place results in being there.
  - If Rover is inside, he will not get rained on, whereas if he is outside he will, because:
    - It is raining.
    - When it is raining, things that are outside tend to get rained on, whereas things inside do not.
  - Rover would prefer not getting rained on to getting rained on, because:
    - Most dogs prefer not to get rained on.

Figure 1: A partial RoU for the following simple story fragment: *. . . One day, it was raining. When Allie arrived, Rover ran out the door. He barked when he felt the rain. He ran right back inside.*

When is the benchmark enough?



Spatial *(sample entries)*:

- Rover is in the yard from when he runs out the door until he runs inside.
- Rover is in the house from when he runs inside until the end of the story.

Temporal *(sample entries)*:

- Allie arrives just before Rover runs outside.
- Rover barks just before he runs inside.
- It is still raining at the end of the story.

Motivational *(sample entry)*:

- Rover runs inside, rather than staying put, because:
  - If he runs inside, he will be inside, whereas if he does not he will be outside, because:
    - Rover is outside.
    - Running to a place results in being there.
  - If Rover is inside, he will not get rained on, whereas if he is outside he will, because:
    - It is raining.
    - When it is raining, things that are outside tend to get rained on, whereas things inside do not.
  - Rover would prefer not getting rained on to getting rained on, because:
    - Most dogs prefer not to get rained on.

Figure 1: A partial RoU for the following simple story fragment: *. . . One day, it was raining. When Allie arrived, Rover ran out the door. He barked when he felt the rain. He ran right back inside.*

Dunietz et al. (2020): "templates of understanding" based on : the machine needs to understand spatial, temporal, causal and motivational aspects of stories (Schank and Abelson, 1977; Zwaan et al., 1995).

Data for reasoning type diagnostics is scarce

most datasets only provide manual analysis of a small sample in the paper, e.g. (Yang et al., 2018a)

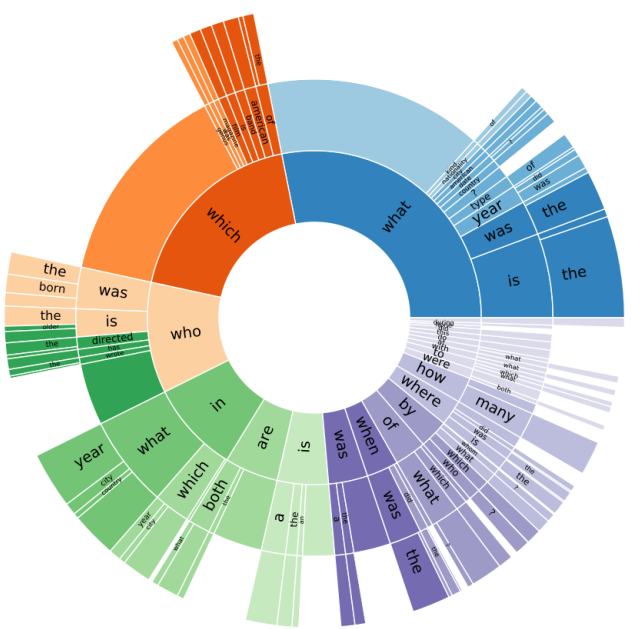


Figure 2: Types of questions covered in HOTPOTQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

Data for reasoning type diagnostics is scarce

most datasets only provide manual analysis of a small sample in the paper, e.g. (Yang et al., 2018a)

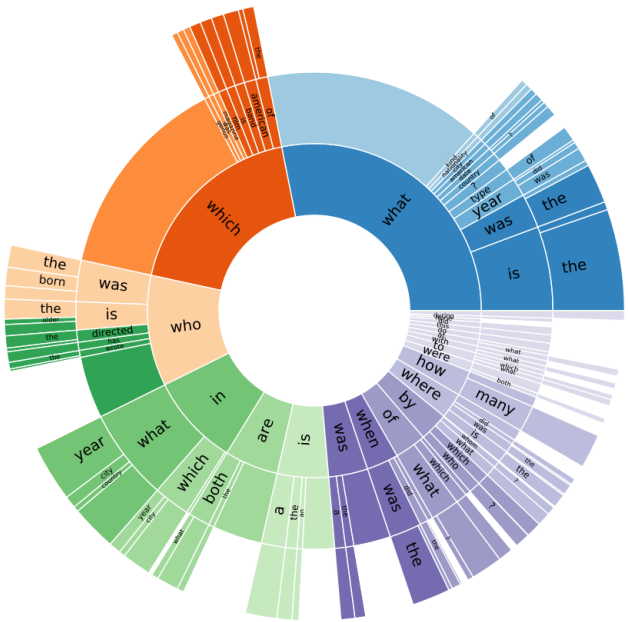


Figure 2: Types of questions covered in HOTPOTQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

alternatives:

- synthetic data (Weston et al., 2015; Labutov et al., 2018)

Data for reasoning type diagnostics is scarce

most datasets only provide manual analysis of a small sample in the paper, e.g. (Yang et al., 2018a)

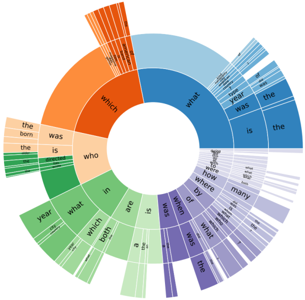


Figure 2: Types of questions covered in HotpotQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

alternatives:

- synthetic data (Weston et al., 2015; Labutov et al., 2018)
- pseudo-labeling, e.g. MS MARCO (Nguyen et al.)

Data for reasoning type diagnostics is scarce

most datasets only provide manual analysis of a small sample in the paper, e.g. (Yang et al., 2018a)

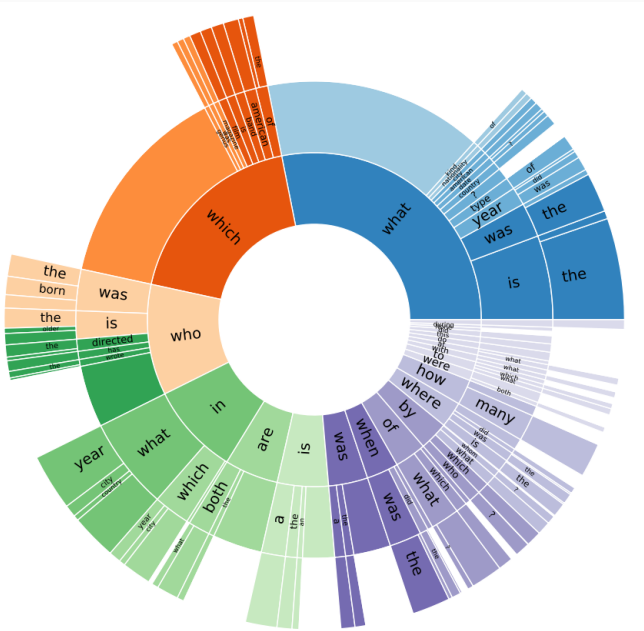


Figure 2: Types of questions covered in HOTPOTQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

alternatives:

- synthetic data (Weston et al., 2015; Labutov et al., 2018)
- pseudo-labeling, e.g. MS MARCO (Nguyen et al.)
- new balanced datasets (Rogers et al., 2020)

Data for reasoning type diagnostics is scarce

most datasets only provide manual analysis of a small sample in the paper, e.g. (Yang et al., 2018a)

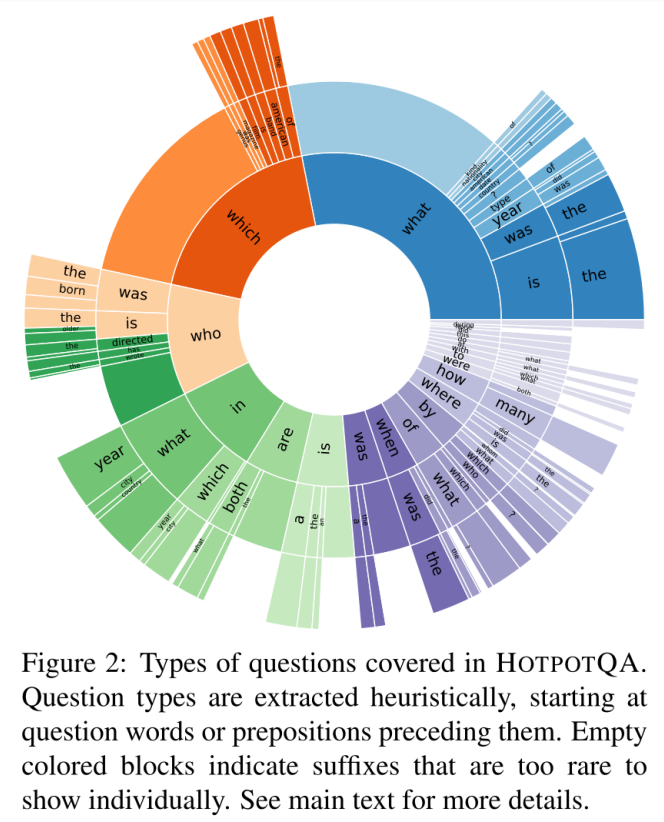


Figure 2: Types of questions covered in HOTPOTQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

alternatives:

- synthetic data (Weston et al., 2015; Labutov et al., 2018)
- pseudo-labeling, e.g. MS MARCO (Nguyen et al.)
- new balanced datasets (Rogers et al., 2020)
- recasting data from other tasks, e.g. semantic resources for inference (White et al., 2017)

Data for reasoning type diagnostics is scarce

most datasets only provide manual analysis of a small sample in the paper, e.g. (Yang et al., 2018a)

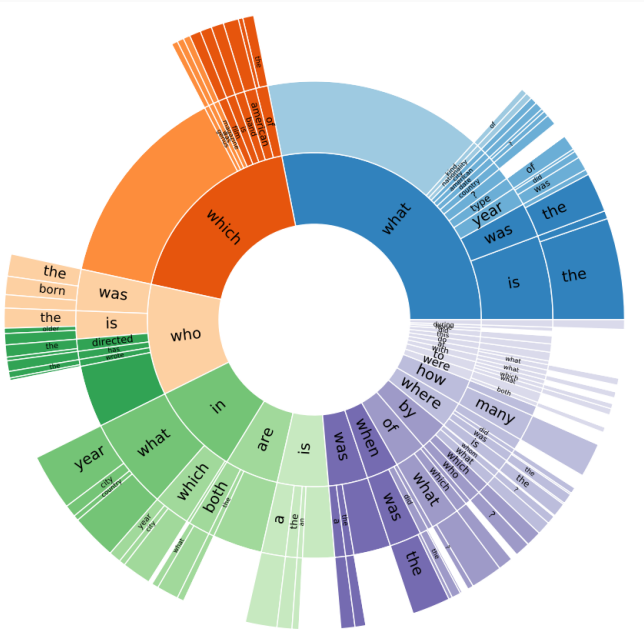


Figure 2: Types of questions covered in HOTPOTQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

alternatives:

- synthetic data (Weston et al., 2015; Labutov et al., 2018)
- pseudo-labeling, e.g. MS MARCO (Nguyen et al.)
- new balanced datasets (Rogers et al., 2020)
- recasting data from other tasks, e.g. semantic resources for inference (White et al., 2017)
- collections of datasets, e.g. ORB (Dua et al., 2019a)

Data for reasoning type diagnostics is scarce

most datasets only provide manual analysis of a small sample in the paper, e.g. (Yang et al., 2018a)

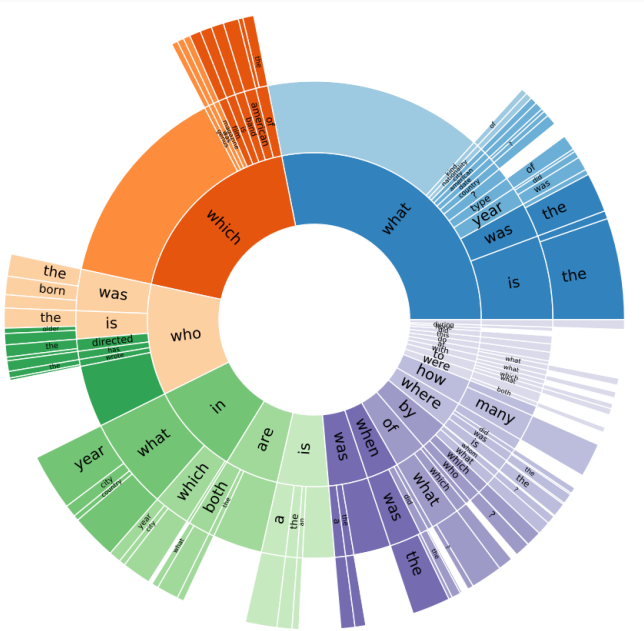


Figure 2: Types of questions covered in HOTPOTQA. Question types are extracted heuristically, starting at question words or prepositions preceding them. Empty colored blocks indicate suffixes that are too rare to show individually. See main text for more details.

alternatives:

- synthetic data (Weston et al., 2015; Labutov et al., 2018)
- pseudo-labeling, e.g. MS MARCO (Nguyen et al.)
- new balanced datasets (Rogers et al., 2020)
- recasting data from other tasks, e.g. semantic resources for inference (White et al., 2017)
- collections of datasets, e.g. ORB (Dua et al., 2019a)
- re-annotating existing datasets, e.g. ARC (Boratko et al., 2018)

Checklisting SQuAD-tuned BERT (Ribeiro et al., 2020)

	Test *TYPE* and Description	Failure Rate (🤖)	Example Test cases (with expected behavior and 🤖 prediction)
Vocab	***MFT:*** comparisons	20.0	**C:** Victoria is younger than Dylan. **Q:** Who is less young? **A:** Dylan 🤖: Victoria
	MFT: intensifiers to superlative: most/least	91.3	**C:** Anna is worried about the project. Matthew is extremely worried about the project. **Q:** Who is least worried about the project? **A:** Anna 🤖: Matthew
Taxonomy	***MFT:*** match properties to categories	82.4	**C:** There is a tiny purple box in the room. **Q:** What size is the box? **A:** tiny 🤖: purple
	MFT: nationality vs job	49.4	**C:** Stephanie is an Indian accountant. **Q:** What is Stephanie's job? **A:** accountant 🤖: Indian accountant
	MFT: animal vs vehicles	26.2	**C:** Jonathan bought a truck. Isabella bought a hamster. **Q:** Who bought an animal? **A:** Isabella 🤖: Jonathan
	MFT: comparison to antonym	67.3	**C:** Jacob is shorter than Kimberly. **Q:** Who is taller? **A:** Kimberly 🤖: Jacob
	MFT: more/less in context, more/less antonym in question	100.0	**C:** Jeremy is more optimistic than Taylor. **Q:** Who is more pessimistic? **A:** Taylor 🤖: Jeremy
Robust.	***INV:*** Swap adjacent characters in **Q** (typo)	11.6	**C:** ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... **Q:** What was the ideal duty → udty of a Newcomen engine? **A:** INV 🤖: 7 million → 5 million
	INV: add irrelevant sentence to **C**	9.8	(no example)
Temporal	***MFT:*** change in one person only	41.5	**C:** Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail 🤖: Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle 🤖: Logan
Neg.	***MFT:*** Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca 🤖: Aaron
	MFT: **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron 🤖: Mark
Coref.	***MFT:*** Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio 🤖: Melissa
	MFT: Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria 🤖: Alex
	MFT: former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly 🤖: Jennifer
SRL	***MFT:*** subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth 🤖: Richard
	MFT: subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa 🤖: Jose

Table 3: A selection of tests for Machine Comprehension.

Type 4: reasoning support

- crowdworkers identify sentences with supporting facts (Yang et al., 2018a)
- annotation of relevant evidence spans (Dua et al., 2020)

Type 4: reasoning support

- crowdworkers identify sentences with supporting facts (Yang et al., 2018a)
- annotation of relevant evidence spans (Dua et al., 2020)

Question:
How many touchdown passes did Cutler throw in the second half?
Answer: 3

.....In the third quarter, the Vikes started to rally with running back Adrian Peterson's 1-yard touchdown run (with the extra point attempt blocked). The Bears increased their lead over the Vikings with Cutler's 3-yard TD pass to tight end Desmond Clark. The Vikings then closed out the quarter with quarterback Brett Favre firing a 6-yard TD pass to tight end Visanthe Shiancoe. An exciting .... with kicker Ryan Longwell's 41-yard field goal, along with Adrian Peterson's second 1-yard TD run. The Bears then responded with Cutler firing a 20-yard TD pass to wide receiver Earl Bennett. The Vikings then completed the remarkable comeback with Favre finding wide receiver Sidney Rice on a 6-yard TD pass on 4th-and-goal with 15 seconds left in regulation. The Bears then took a knee to force overtime.... The Bears then won on Jay Cutler's game-winning 39-yard TD pass to wide receiver Devin Aromashodu. With the loss, not only did the Vikings fall to 11-4, they also surrendered homefield advantage to the Saints.

Figure 1: Example from DROP, showing the intermediate annotations that we collected via crowd-sourcing.

(Dua et al., 2020)

Outline

Methodology issues

The SOTA chase - statistical testing = trouble

			Δ	
Model	AP	RR	AP	RR
Yu et al. (2014)	0.6190	0.6281		
Yang et al. (2015)	0.6520	0.6652	0.0330	0.0371
dos Santos et al. (2016)	0.6886	0.6957	0.0366	0.0305
Miao et al. (2016)	0.6886	0.7069	0.0000	0.0112
Yin et al. (2016)	0.6921	0.7108	0.0035	0.0039
Rao et al. (2016)	0.701	0.718	0.0080	0.0072
Wang et al. (2016b)	0.7058	0.7226	0.0048	0.0046
He and Lin (2016)	0.7090	0.7234	0.0032	0.0008
Yin and Schütze (2017)	0.7124	0.7237	0.0034	0.0003
Chen et al. (2017a)	0.7212	0.7312	0.0088	0.0075
Wang et al. (2016a)	0.7341	0.7418	0.0129	0.0106
Wang and Jiang (2016)	0.7433	0.7545	0.0092	0.0127

Table 3: State-of-the-art (gathered by manual inspection) results on the WikiQA dataset, annotated with improvement over prior state-of-the-art results.

(Crane, 2018)

Much of reported improvements are unreproducible and within variability due to unrelated factors

Much of reported improvements are unreproducible and within variability due to unrelated factors (Crane, 2018)

- versions of the model, underlying framework and low-level libraries;
- threading
- GPU computation
- random seed (see also (Dodge et al., 2020))
- interaction between the above
- reporting roundin

Some random seeds are MUCH better! (McCoy et al., 2019a)

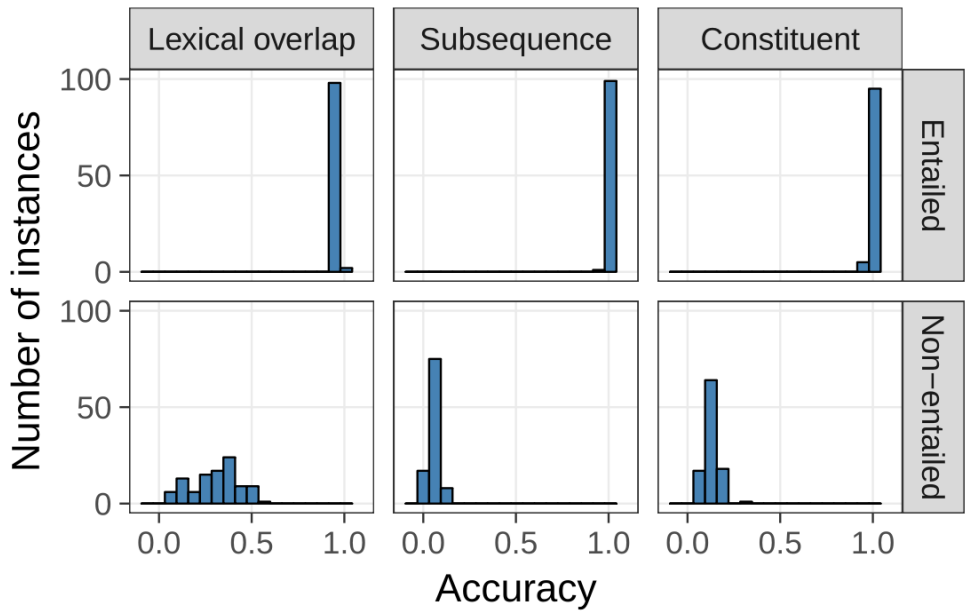
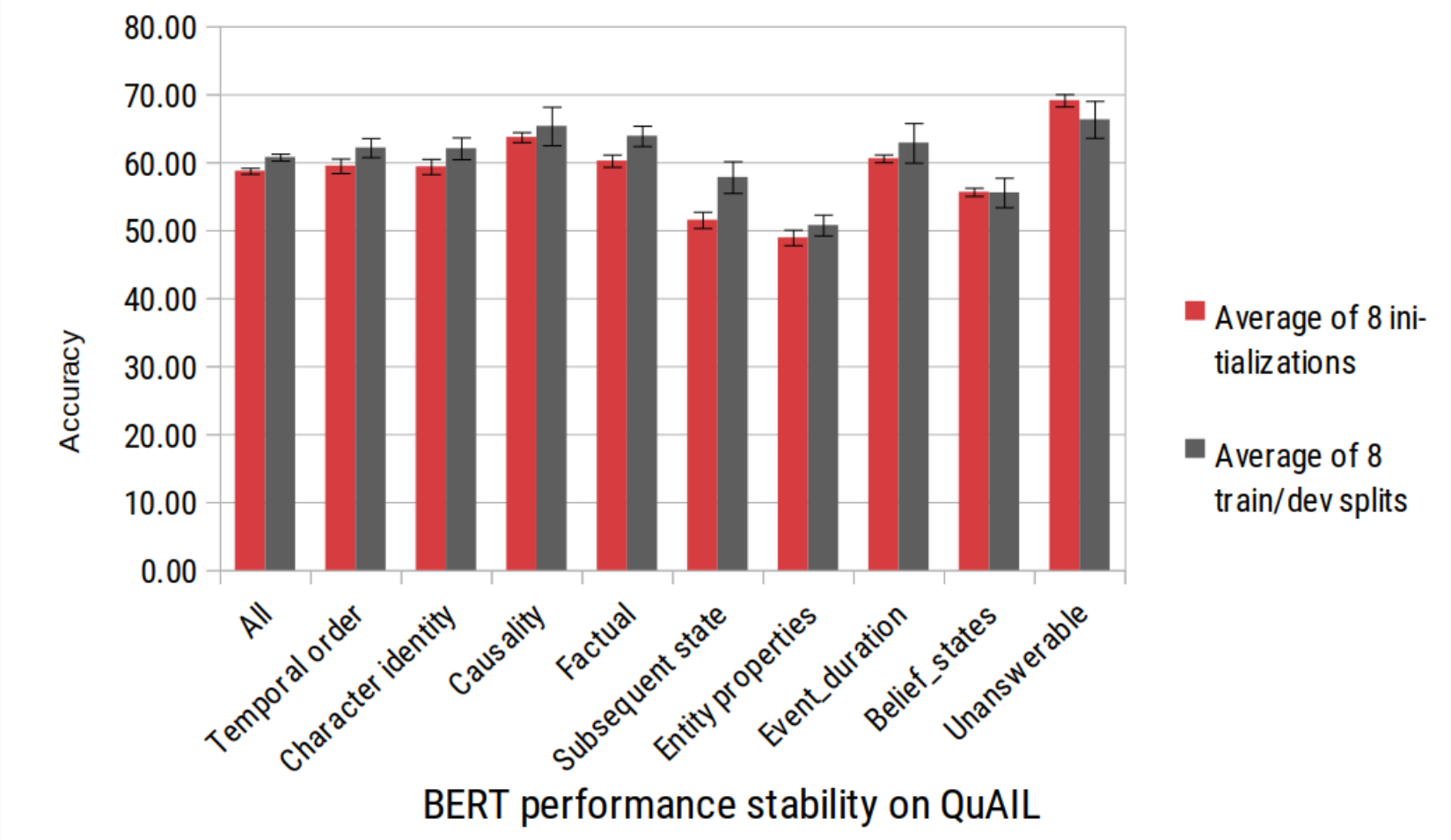


Figure 3: Out-of-distribution generalization: Performance on the HANS evaluation set, broken down into six categories of examples based on which syntactic heuristic each example targets and whether the correct label is *entailment* or *non-entailment*. The non-entailed lexical overlap cases (lower left plot) display a large degree of variability across instances.

Data order matters just as much!

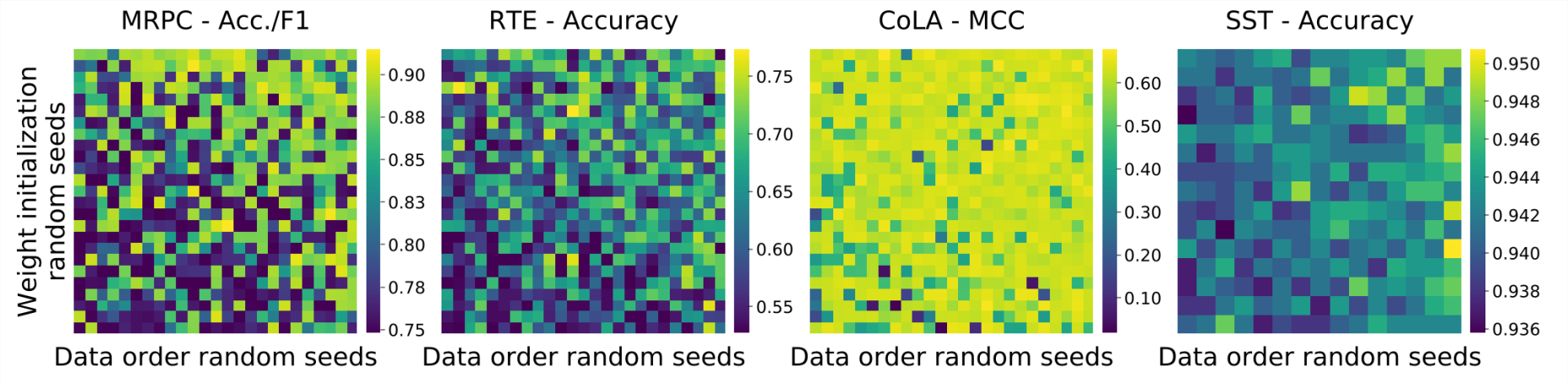


(Rogers et al., 2020)

Data order matters just as much!

- SOTA on the ”standard split” may not reproduce on a random split (Gorman and Bedrick, 2019)
- both random and standard splits overestimate performance on new samples (Søgaard et al., 2020)

Interaction between data order and model inits



(Dodge et al., 2020)

Data collection ethics

NAACL 2021: https://2021.naacl.org/ethics/faq/

> "Detail the dataset collection process and conditions. If manual work was involved, describe measures taken to ensure that crowd workers or other annotators were fairly compensated and how fair compensation was determined."

NAACL 2021: https://2021.naacl.org/ethics/faq/

- Describe the characteristics of the dataset in enough detail for a reader to understand which speaker populations the technology could be expected to work for. (For suggestions of what kind of information to include, see (Bender and Friedman, 2018; Mitchell et al., 2019; Gebru et al., 2020).
- Finally, describe the steps taken to ensure that potential problems with the quality of the dataset do not create additional risks.

Data documentation is important not only from bias & fairness standpoint!

Data leaks vs fair evaluation on new data

GPT3-3 (Brown et al., 2020): filtering pre-training data to avoid direct overlaps with specific benchmark datasets (based on 13-gram overlap criterion)

Name	Split	Metric	N	Acc/F1/BLEU	Total Count	Dirty Acc/F1/BLEU	Dirty Count	Clean Acc/F1/BLEU	Clean Count	Clean Percentage	Relative Difference Clean vs All
Quac	dev	f1	13	44.3	7353	44.3	7315	54.1	38	1%	20%
SQuADv2	dev	f1	13	69.8	11873	69.9	11136	68.4	737	6%	-2%
DROP	dev	f1	13	36.5	9536	37.0	8898	29.5	638	7%	-21%
Symbol Insertion	dev	acc	7	66.9	10000	66.8	8565	67.1	1435	14%	0%
CoQa	dev	f1	13	86.0	7983	85.3	5107	87.1	2876	36%	1%
ReCoRD	dev	acc	13	89.5	10000	90.3	6110	88.2	3890	39%	-1%
Winograd	test	acc	9	88.6	273	90.2	164	86.2	109	40%	-3%
BoolQ	dev	acc	13	76.0	3270	75.8	1955	76.3	1315	40%	0%
MultiRC	dev	acc	13	74.2	953	73.4	558	75.3	395	41%	1%
RACE-h	test	acc	13	46.8	3498	47.0	1580	46.7	1918	55%	0%
LAMBADA	test	acc	13	86.4	5153	86.9	2209	86.0	2944	57%	0%
LAMBADA (No Blanks)	test	acc	13	77.8	5153	78.5	2209	77.2	2944	57%	-1%
WSC	dev	acc	13	76.9	104	73.8	42	79.0	62	60%	3%
PIQA	dev	acc	8	82.3	1838	89.9	526	79.3	1312	71%	-4%

Data leaks vs fair evaluation on new data

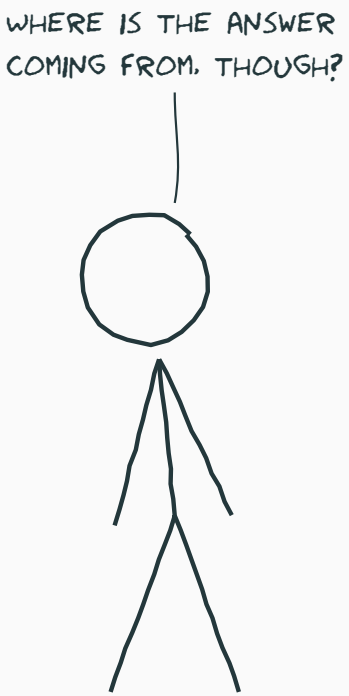
GPT3-3 (Brown et al., 2020):

The information required to answer the question is in a passage provided to the model, so having seen the passage during training but not the questions and answers does not meaningfully constitute cheating.

Data leaks vs fair evaluation on new data

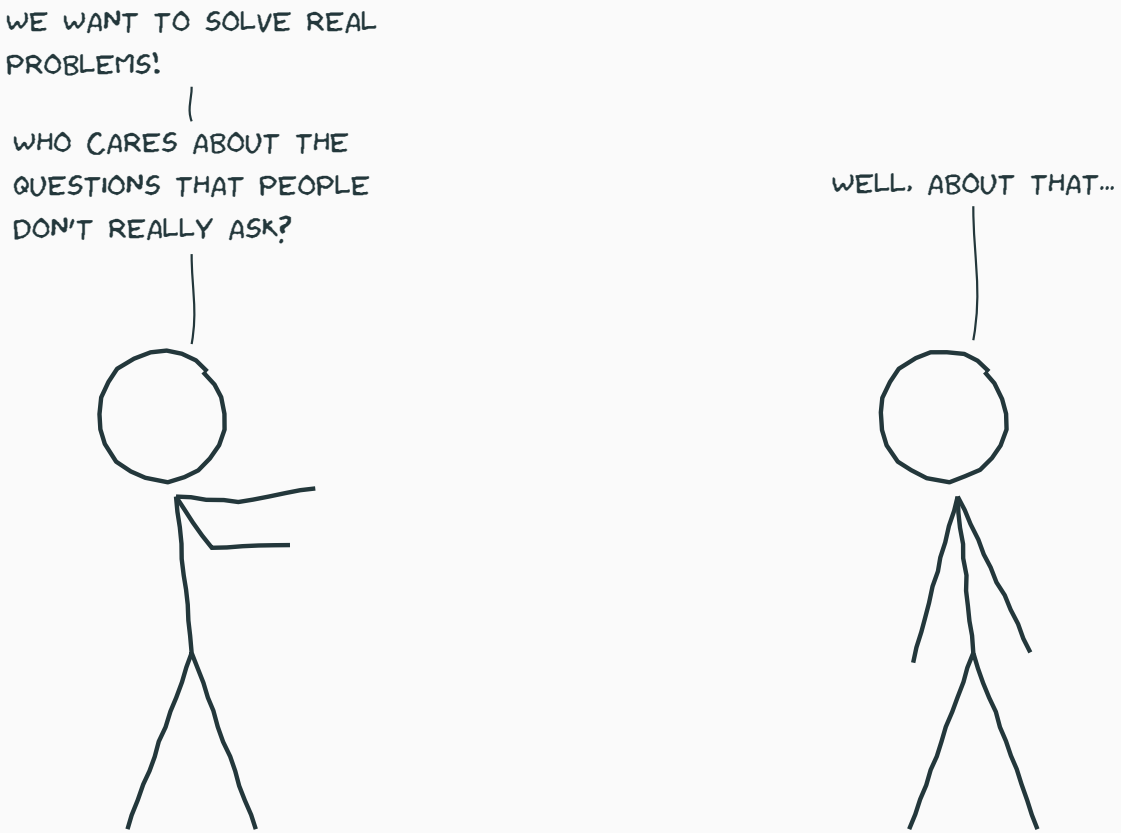
GPT3-3 (Brown et al., 2020):

The information required to answer the question is in a passage provided to the model, so having seen the passage during training but not the questions and answers does not meaningfully constitute cheating.



The “Natural questions" dilemma

The "Natural questions" dilemma



The "Natural questions" dilemma

- the users are already used to the limitations of search engines and voice assistants, and formulate the questions that they think more likely to get answered;

The "Natural questions" dilemma

- the users are already used to the limitations of search engines and voice assistants, and formulate the questions that they think more likely to get answered;
- it's often queries rather than questions;

The "Natural questions" dilemma

- the users are already used to the limitations of search engines and voice assistants, and formulate the questions that they think more likely to get answered;
- it's often queries rather than questions;
- the distribution of questions we can obtain from real queries is limited;

The "Natural questions" dilemma

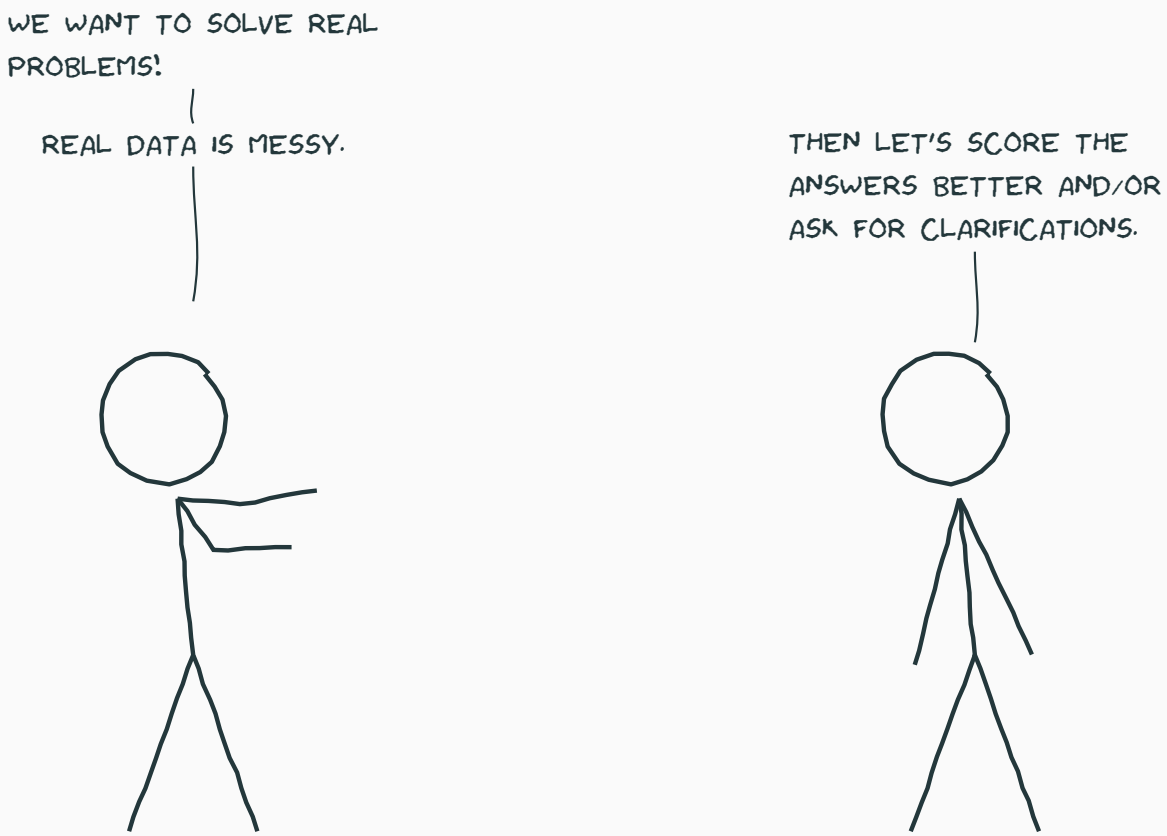
- the users are already used to the limitations of search engines and voice assistants, and formulate the questions that they think more likely to get answered;
- it's often queries rather than questions;
- the distribution of questions we can obtain from real queries is limited;
- the questions may be ambiguous and/or having implicit assumptions (Boyd-Graber, 2019);

The "Natural questions" dilemma

The questions "can i buy wine in kentucky on sunday", "where am i on the steelers waiting list", "when is the real housewives on", and "who has majority in the house and senate" are all answerable, but depend on which county of Kentucky you're in, when you paid for your season pass, and the local network syndicating Real Housewives. However, Natural Questions calls these unanswerable, while the previous questions are answerable with implicit assumptions.

(Boyd-Graber, 2019)

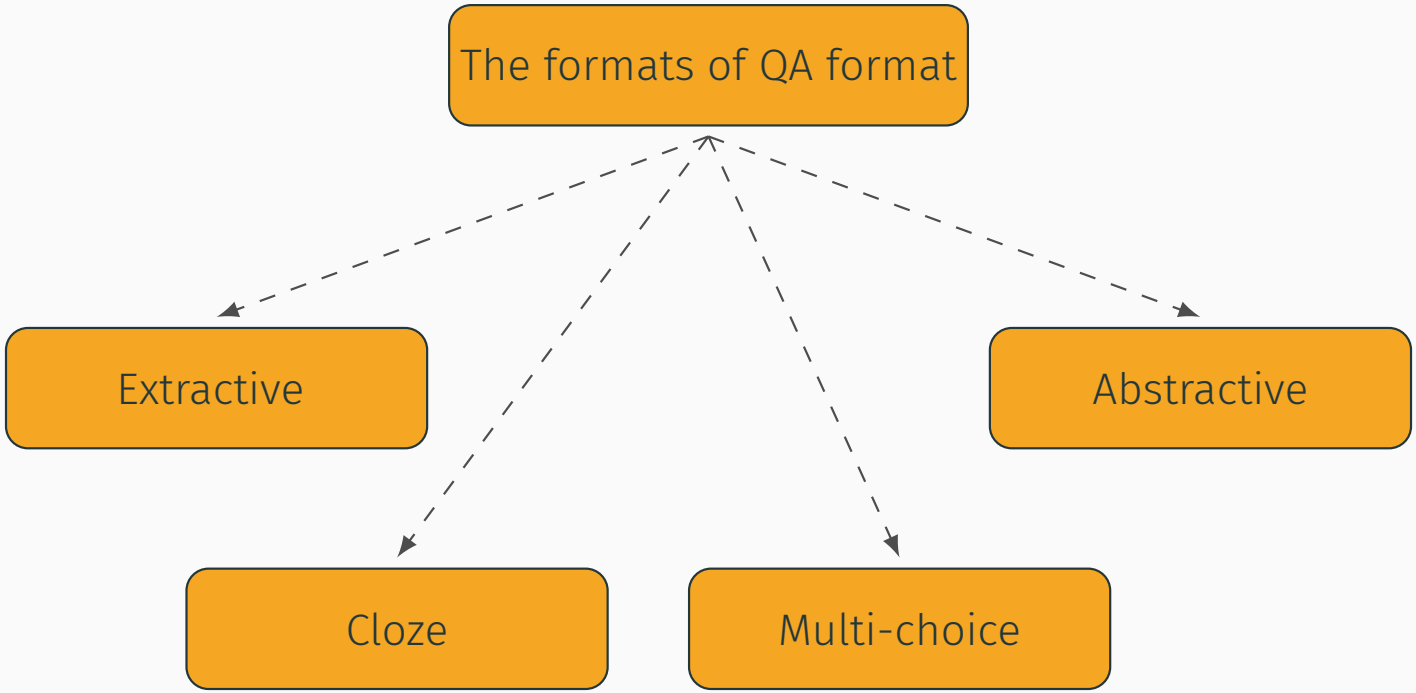
The "Natural questions" dilemma



See (Boyd-Graber, 2019; Elgohary et al., 2019)

Seriously, what format should it be?

The war of formats



Extractive or multi-choice?

Extractive:

- easier to create
- limited to information that is explicitly stated
- IR-leaning

Multi-choice:

- harder to create (Berzak et al., 2020)
- any information could be queried
- more like reasoning

CBT: the thin border between extractive and multi-choice QA

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

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 Mr. Baxter had exaggerated matters a little.

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 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 .
20 Esther felt relieved .

q: She thought that Mr. _____ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

(Hill et al., 2015a)

Ideally - abstractive QA, but...

Have to solve evaluation for text generation first!

Current automated metrics are not great

Context: . . . After Peter returns, they eventually figure out her proper care, right down to diaper changes, baths, and feedings. The next day, **two men** (**who are drug dealers**) arrive at the apartment to pick up the package. . .

Question: Who comes to pick up the package the next day?
Gold Answers: **drug dealers, the drug dealer**
Prediction: **two men**

Human Judgement: 5 out of 5
ROUGE-L: 0
METEOR: 0

(a) Example from the generative **NarrativeQA** dataset.

existing metrics don't use the context, and fail to capture coreferences (Chen et al., 2019)

Current automated metrics are not great

Context: ...David got five exercise tips from his personal trainer, **tip A**, **tip B** ... **Tip A** involves weight lifting, but **tip B** does not involve weight lifting ...
Question: In which tip the skeletal muscle would not be bigger, **tip A** or **tip B**?
Gold Answers: **tip B**
Prediction: **tip A**

Human Judgement: 1 out of 5
F1: 0.66

(b) Example from the span-based **ROPES** dataset.

changing a single token can make a prediction incorrect, but F1 will be non-zero (Chen et al., 2019)

What should the data do?

Option 1: data for training + testing

Have to give the model a fair chance to learn! (Geiger et al., 2019)

Not learning the deeper patterns!

Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

Adversarial entailment: replacing a word in SNLI premises with its synonyms or hypernyms

Contradiction: replacing words with mutually exclusive co-hyponyms and antonyms

Result: 10-30% performance drop for 3 neural NLI systems

Not learning the deeper patterns!

Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

Adversarial entailment: replacing a word in SNLI premises with its synonyms or hypernyms

Contradiction: replacing words with mutually exclusive co-hyponyms and antonyms

Result: 10-30% performance drop for 3 neural NLI systems

"What mostly affects the systems' ability to correctly predict a test example is the amount of similar examples found in the training set. Given that training data will always be limited, this is a rather inefficient way to learn lexical inference." (Glockner et al., 2018)

Despite our best efforts, we may never be able to create a benchmark that does not have unintended statistical regularities. (Linzen, 2020)

Option 1: data for training + testing

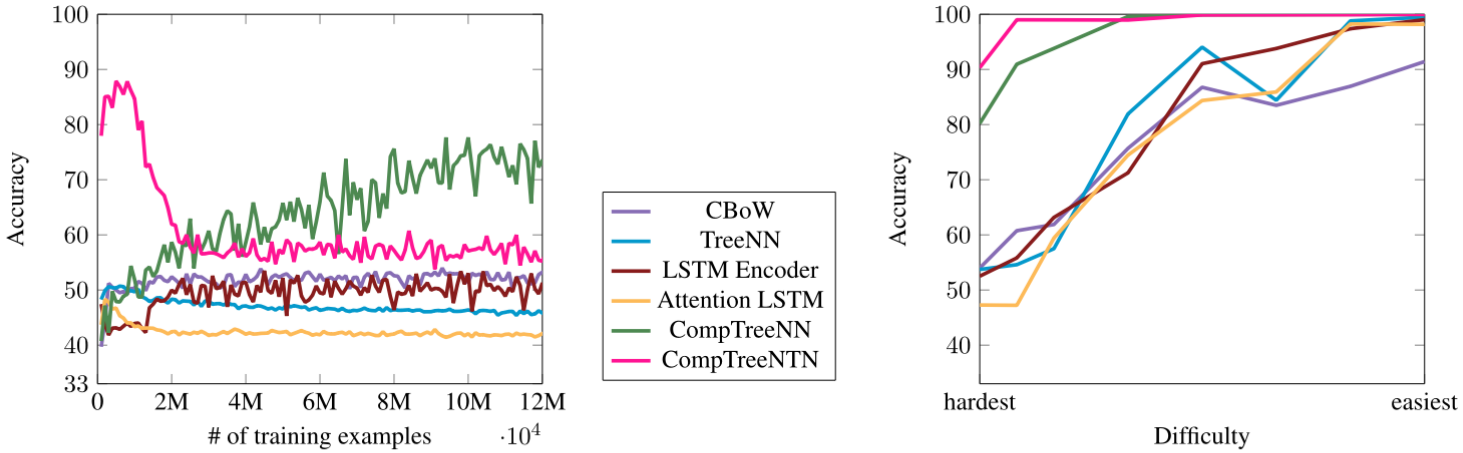


Figure 5: **Left**: Model performance on our difficult but fair generalization task throughout training. **Right**: Mean accuracy of 5 runs as we move from true generalization tasks ('hardest') to problems in which the training set contains so much redundant encoding of the test set that the task is essentially one of memorization ('easiest'). Only the task-specific CompTreeNN and CompTreeNTN are able to do well on true generalization tasks. The other neural models succeed only where memorization suffices, and the CBoW model never succeeds because it does not encode word order.

Classic baselines don't learn even when they have a chance to! (Geiger et al., 2019)

Option 2: test-only benchmarks

- "Generalization leaderboard": train on separate data (Linzen, 2020);

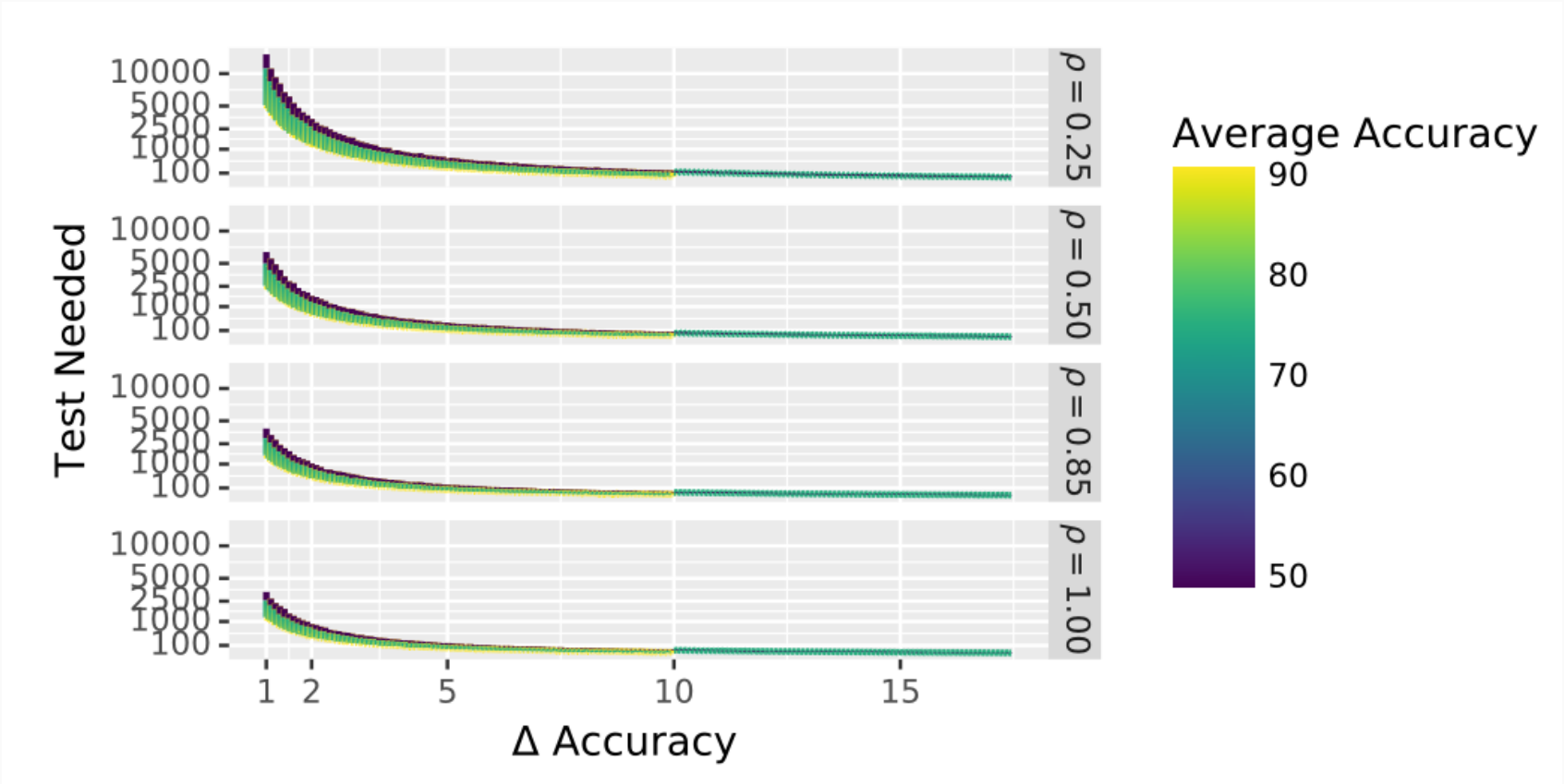
Option 2: test-only benchmarks

- "Generalization leaderboard": train on separate data (Linzen, 2020);
- rigorously test various capabilities (Ribeiro et al., 2020)

Option 2: test-only benchmarks

- "Generalization leaderboard": train on separate data (Linzen, 2020);
- rigorously test various capabilities (Ribeiro et al., 2020)
- consider not only accuracy, but also compute and data efficiency etc. (Rogers, 2019; Ethayarajh and Jurafsky, 2020; Boyd-Graber, 2019)

How big should the test datasets be?



Test data size depends on (a) the difference in accuracy between the systems, (b) their avg accuracy (closer to 50% is harder), and (c) the amount of discriminative questions. If 100% questions are disciminative, 2.5K is enough even at 1% Δ with 50% average accuracy. If only 25% is discriminative, we need 15K. (Boyd-Graber, 2019)

Thank You!

Tutorial page:

`https://annargrs.github.io/dataset-explosion`

Anna Rogers
University of Copenhagen
arogers@sodas.ku.dk
@annargrs

Anna Rumshisky
University of Massachusetts Lowell
arum@cs.uml.edu
@arumshisky

Acknowledgements

Many thanks to:

- XKCD comics[2] for inspiration;
- Michael Ciuffo for the Humor-Sans font[3]
- Percusse on Tex StackExchange for the xkcd comic code[4]

[2] https://xkcd.com/
[3] http://antiyawn.com/uploads/humorsans.html
[4] https://tex.stackexchange.com/a/74881

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the Cross-lingual Transferability of Monolingual Representations. *arXiv:1910.11856 [cs]*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268 [cs]*.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2017. Embracing data abundance: BookTest Dataset for Reading Comprehension. In *ICLR*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. Association for Computational Linguistics.

Rachel Bawden. 2019. One paper, nine reviews.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and

Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured Annotations for Reading Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 60–70.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale Simple Question Answering with Memory Networks. *arXiv:1506.02075 [cs]*.

Jordan Boyd-Graber. 2019. What Question Answering can Learn from Trivia Nerds. *arXiv:1910.14464 [cs]*.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the Quiz Master: Crowdsourcing Incremental Classification Games. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1290–1301, Jeju Island, Korea. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating Question Answering Evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020a. MOCHA: A Dataset for Training and Evaluating Generative Reading Comprehension Metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Minseok Cho, Reinald Kim Amplayo, Seung-won Hwang, and Jonghyuck Park. 2018. Adversarial TableQA: Attention Supervision for Question Answering on Tables. In *Proceedings of Machine Learning Research*, pages 391–406.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2924–2936.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TYDI QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. page 17.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018a. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457 [cs]*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018b. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457 [cs]*.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs. *arXiv:2004.11861 [cs]*.

Matt Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2020. Does BERT Solve Commonsense Task via Commonsense Knowledge? *arXiv:2008.03945 [cs]*.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. 2018. Building Dynamic Knowledge Graphs from Text using Machine Reading Comprehension. *arXiv:1810.05682 [cs]*.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5924–5931, Hong Kong, China. Association for Computational Linguistics.

Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. FQuAD: French Question Answering Dataset. *arXiv:2002.06071 [cs]*.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv:2002.06305 [cs]*.

Xinya Du and Claire Cardie. 2020. Event Extraction by Answering (Almost) Natural Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Matt Gardner, and Sameer Singh. 2019a. ORB: An Open Reading Benchmark for Comprehensive Evaluation of Machine Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 147–153, Hong Kong, China. Association for Computational Linguistics.

Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of Intermediate Annotations in Reading Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5627–5634.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019b. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To Test Machine Comprehension, Start by Defining Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q A Dataset Augmented with Context from a Search Engine. *arXiv:1704.05179 [cs]*.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. *arXiv:1912.09723 [cs]*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Florian Englmaier, Stefan Grimm, David Schindler, and Simeon Schudy. 2018. The Effect of Incentives in Non-Routine Analytical Teams Tasks-Evidence from a Field Experiment.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. *arXiv:2009.13888 [cs]*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-Scale QA-SRL Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question Answering is a Format; When is it Useful? *arXiv:1909.11291 [cs]*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]*.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing Fair Generalization Tasks for Natural Language Inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4475–4485, Hong Kong, China. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *EMNLP*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Kyle Gorman and Steven Bedrick. 2019. We Need to Talk about Standard Splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Shangmin Guo, Kang Liu, Shizhu He, Cao Liu, Jun Zhao, and Zhuoyu Wei. 2017. IJCNLP-2017 Task 5: Multi-choice Question Answering in Examinations. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 34–40, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. AmazonQA: A Review-Based Question Answering Task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4996–5002, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Namgi Han, Goran Topic, Hiroshi Noji, Hiroya Takamura, and Yusuke Miyao. 2020. An empirical analysis of existing systems and datasets toward general simple question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5321–5334.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Benjamin Heinzerling. 2019. NLP's Clever Hans Moment has Arrived.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015.

Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A Novel Large-scale Language Understanding Task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015a. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv:1511.02301 [cs]*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015b. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv:1511.02301 [cs]*.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. TempQuestions: A Benchmark for Temporal Question Answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 1057–1062, Lyon, France. ACM Press.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 1807–1810, Torino, Italy. Association for Computing Machinery.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A New Factoid QA Data Set Matching Trivia-Style Question-Answer Pairs with Freebase. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 318–323.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set

for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the First Sentence: Position Bias in Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018a. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018b. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Vladislav Korablinov and Pavel Braslavski. 2020. RuBQ: A Russian Dataset for Question Answering over Wikidata. *arXiv:2005.10659 [cs]*.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. In *International Conference on Machine Learning*, pages 1378–1387. PMLR.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria. 2018. Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in Simulated Worlds. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 833–844, Melbourne, Australia. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event Extraction as Multi-turn Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering. *arXiv:1607.06275 [cs]*.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-Relation Extraction as Multi-Turn Question Answering. In *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning Over Paragraph Effects in Situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? *arXiv:2005.00955 [cs]*.

Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2017. World Knowledge for Reading Comprehension: Rare Entity Prediction with Hierarchical LSTMs Using External Descriptions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 825–834. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The Natural Language Decathlon: Multitask Learning as Question Answering. *arXiv:1806.08730 [cs, stat]*.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2019a. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv:1911.02969 [cs]*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019b. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing Question-Answer Meaning Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short*

Papers), pages 560–568, New Orleans, Louisiana. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and*

Transparency, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Saif M. Mohammad. 2020. NLP Scholar: An Interactive Visual Explorer for Natural Language Processing Literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 232–255, Online. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Nathanael Chambers, and Annie Louis. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-Level Semantics*, pages 46–51. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. page 10.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. *arXiv:2005.00242 [cs]*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A Large-Scale Person-Centered Cloze Dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.

Cheoneum Park, Myungji Kim, Soyoon Park, Seungyoung Lim, Jooyoul Lee, and Changki Lee. Korean TableQA: Structured data question answering based on span prediction style with S3-NET. *ETRI Journal*, n/a(n/a).

Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Anselmo Pe

textasciitilde nas, Christina Unger, and Axel-Cyrille Ngonga Ngomo. 2014. Overview of CLEF Question Answering Track 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 300–306. Springer, Cham.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in

Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge. *arXiv:1808.07042 [cs]*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Matthew Richardson, Christopher J C Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, 18-21 October 2013.

Anna Rogers. 2019. How the Transformers broke NLP leaderboards.

Anna Rogers. 2020. Peer review in NLP: Resource papers.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8722–8731.

Barbara Rychalska, Dominika Basaj, Anna Wróblewska, and Przemyslaw Biecek. 2018. Does it care what you asked? Understanding Importance of Verbs in Deep Learning QA System. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 322–324. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. *arXiv:1907.10641 [cs]*.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. The Artificial Intelligence Series. L. Erlbaum Associates ;

distributed by the Halsted Press Division of John Wiley and Sons, Hillsdale, N.J. : New York.

Kim Schouten, Flavius Frasincar, and Franciska de Jong. 2017. Ontology-Enhanced Aspect-Based Sentiment Analysis. In Jordi Cabot, Roberto De Virgilio, and Riccardo Torlone, editors, *Web Engineering*, volume 10360, pages 302–320. Springer International Publishing, Cham.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2019. DRCD: A Chinese Machine Reading Comprehension Dataset. *arXiv:1806.00920 [cs]*.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the Story Ending Biases in The Story Cloze Test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Joost Bastings, and Katja Filippova. 2020. We Need to Talk About Random Splits. *arXiv:2005.00636 [cs]*.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In *AAAI*.

Ningyuan Sun, Xuefeng Yang, and Yunfeng Liu. 2020. TableQA: A Large-Scale Chinese Text-to-SQL Dataset for Table-Aware SQL Generation. *arXiv:2006.06434 [cs]*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Simon Suster and Walter Daelemans. 2018. CliCR: A Dataset of Clinical Case Reports for Machine Reading Comprehension. In *Proceedings*

of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1551–1563.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. *arXiv:2009.10795 [cs]*.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019. QuaRel: A Dataset and Models for Answering Questions about Qualitative Relationships. In *AAAI 2019*.

Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 641–651, New Orleans, Louisiana.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A Machine Comprehension Dataset. *arXiv:1611.09830 [cs]*.

Chen-Tse Tsai, Wen-tau Yih, Chris J.C. Burges, and Scott Wen-tau Yih. 2015. Web-based Question Answering: Revisiting AskMSR. Technical Report MSR-TR-2015-20.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Shyam Upadhyay and Ming-Wei Chang. 2017. Annotating Derivations: A New Evaluation Strategy and Dataset for Algebra Word Problems. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long*

Papers, pages 494–504, Valencia, Spain. Association for Computational Linguistics.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal Reasoning in Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. *arXiv:1906.04701 [cs]*.

Kiri Wagstaff. 2012. Machine Learning that Matters. In *ICML*.

Eric Wallace and Jordan Boyd-Graber. 2018. Trick Me If You Can: Adversarial Writing of Trivia Challenge Questions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 127–133. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. *EMNLP*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv:1905.00537 [cs]*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, and Xiaochuan Wang. 2020. ReCO: A Large Scale Chinese Reading Comprehension Dataset on Opinion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)*, page 8.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making Neural QA as Simple as Possible but not Simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale Cloze Test Dataset Created by Teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Mark Yatskar. 2019. A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long and Short Papers), pages 2318–2323.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL 2019*.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020a. WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for

Answering Winograd Schema Challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv:1810.12885 [cs]*.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020b. When Do You Need Billions of Words of Pretraining Data? *arXiv:2011.04946 [cs]*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *arXiv:1709.00103 [cs]*.

Rolf A. Zwaan, Mark C. Langston, and Arthur C. Graesser. 1995. The Construction of Situation Models in Narrative Comprehension: An Event-Indexing Model. *Psychological Science*, 6(5):292–297.