Advanced NLP

Lecture 2: Extrsinic Tasks

Dr. Gabriel Stanovsky

March 18, 2023

Suggested reading: Speech and Language Processing: an Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. Daniel Jurafsky & James H Martin

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM
الجامعة العبرية في اورشليم القدس

Last Week: Intrinsic Tasks

- **Intrinsic tasks** (aka *intermediate*)
  - Inherently required across extrinsic tasks
  - But are not directly useful on their own
  - Often correspond to much-studied linguistic phenomena

Recognizing Textual Entailment (Or NLI)

The task of deciding whether the meaning of one text (the **Hypothesis***) is entailed, or can be inferred, from another text (the* **Premise***)*[1]

- Typically consisting of **three labels**
  - **Premise:** "Yoko Ono unveiled a bronze statue for her late husband, John Lennon."
- **Entailment**
  "Yoko Ono is John Lennon's widow"
- **Contradiction**
  "John Lennon is Yoko Ono's widow"
- **Neutral**
  "John Lennon and Yoko Ono married in 1969"

Grounding

Mapping from text (or form) to a world (ontology, or meaning)

Coreference resolution

An important component of language processing is knowing who is being talked about in a text.

Victoria Chen, CFO of Megabucks Banking, saw *her* pay jump to $2.3 million, as *the 38-year-old* became the company's president. It is widely known that *she* came to Megabucks from rival Lotsabucks.

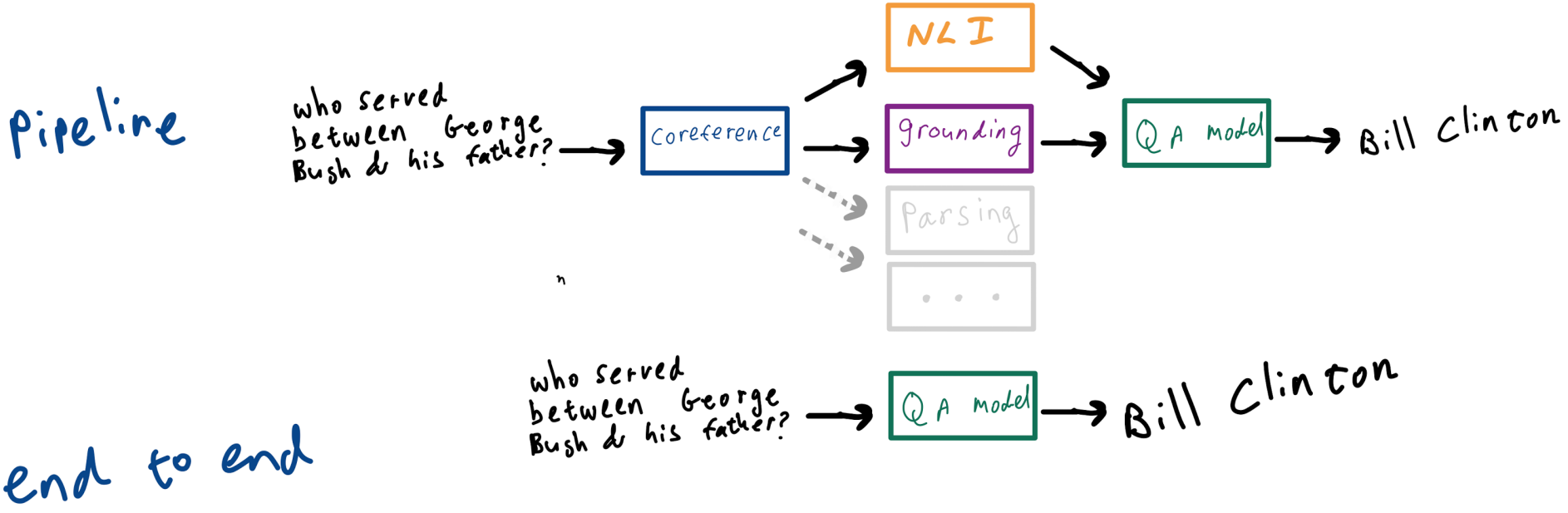
Entity Linking

The task of associating a mention in text with the representation of some real-world entity in an ontology.[2, 3] [Chap. 14.3]

George Bush reveals how he repeatedly turned to **his father** for advice as he contemplated following him into war against Saddam Hussein.

Pipeline vs. End-to-End Approaches

- **E2E models do not require intermediate task labels**
  - Trained on input and outputs w/o intermediate labels
- **But intrinsic tasks are still useful**
  - Resurgence in using pipeline architectures
  - More efficient models



Today: Extrinsic Tasks

- Tasks which have applicable value for external users
- We'll present prominent tasks, and discuss:
  - **Task Definition & Motivation**
  - **Annotation methodologies**
  - **Evaluation protocols**
  - **Popular Benchmarks**
- Focus on English today
  - We'll discuss multilingual NLP in a future lecture

- **Have clear user-facing value**
  - As opposed to intermediate tasks
- **Serve as motivation for much of NLP research**
- **Q: Which extrsinic tasks have you discussed?**
  - Machine translation (sometimes called NLP-complete)
  - Information Extraction

Today

1. Sentiment Analysis
2. Question Answering
3. Summarization
4. Information Extraction
5. Evaluation
6. Aggregated Benchmarks
7. Conclusion

Sentiment Analysis

Positive or negative orientation a writer expresses towards an object
[3][Chap. 4]

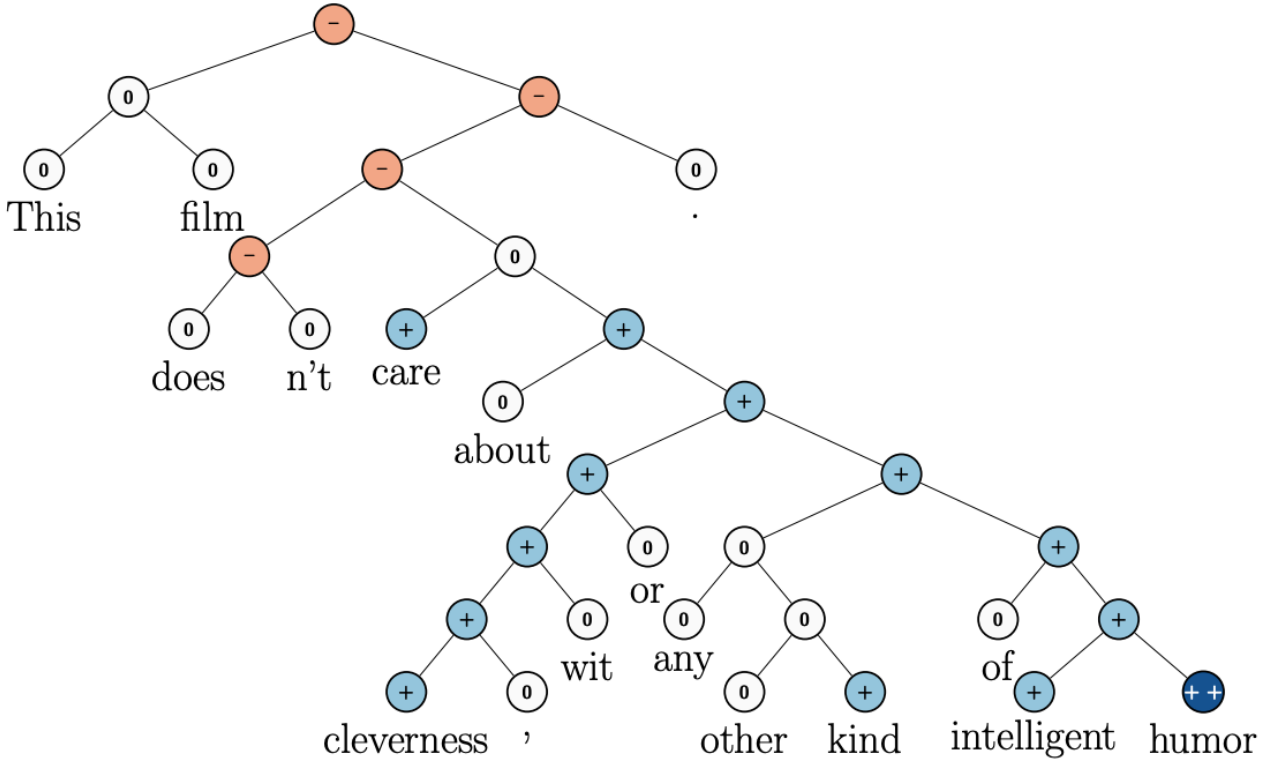
- **Positive**
  "Awesome caramel sauce and sweet toasty almonds. I it!"

- **Negative**
  "It was pathetic. The worst part about it was the boxing scenes."

- E.g., for reviews, political opinions, and more

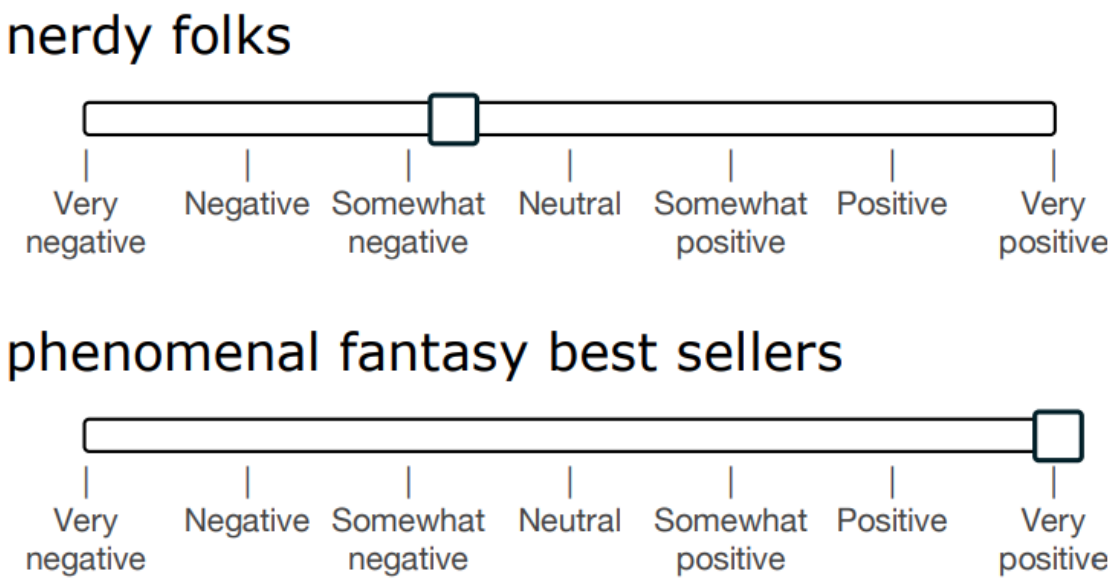
Benchmarks:SST

- Stanford Sentiment Treebank (SST) [4]
- 5-way classification
  - From very negative to very positive
- 11.8K movie reviews
- Also annotates subtrees



SST: Annotation

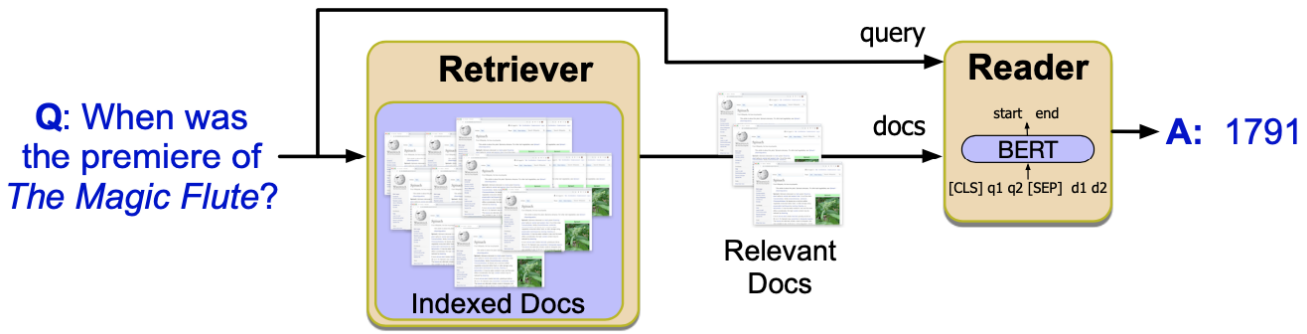
- Sentences broken down to 215K phrases
  - Using a syntactic parser
- Annotated on **Mechanical Turk**



- We'll revisit sentiment analysis as a testbed for text classification
- Other examples include **topic classification**, **spam detection**

Open Domain QA

Answer a user's question by finding short text segments from the web or some other large collection of documents[3][Chap. 14.2]



- NLP research mostly focuses on **reading comprehension**
  - Assuming relevant documents are given

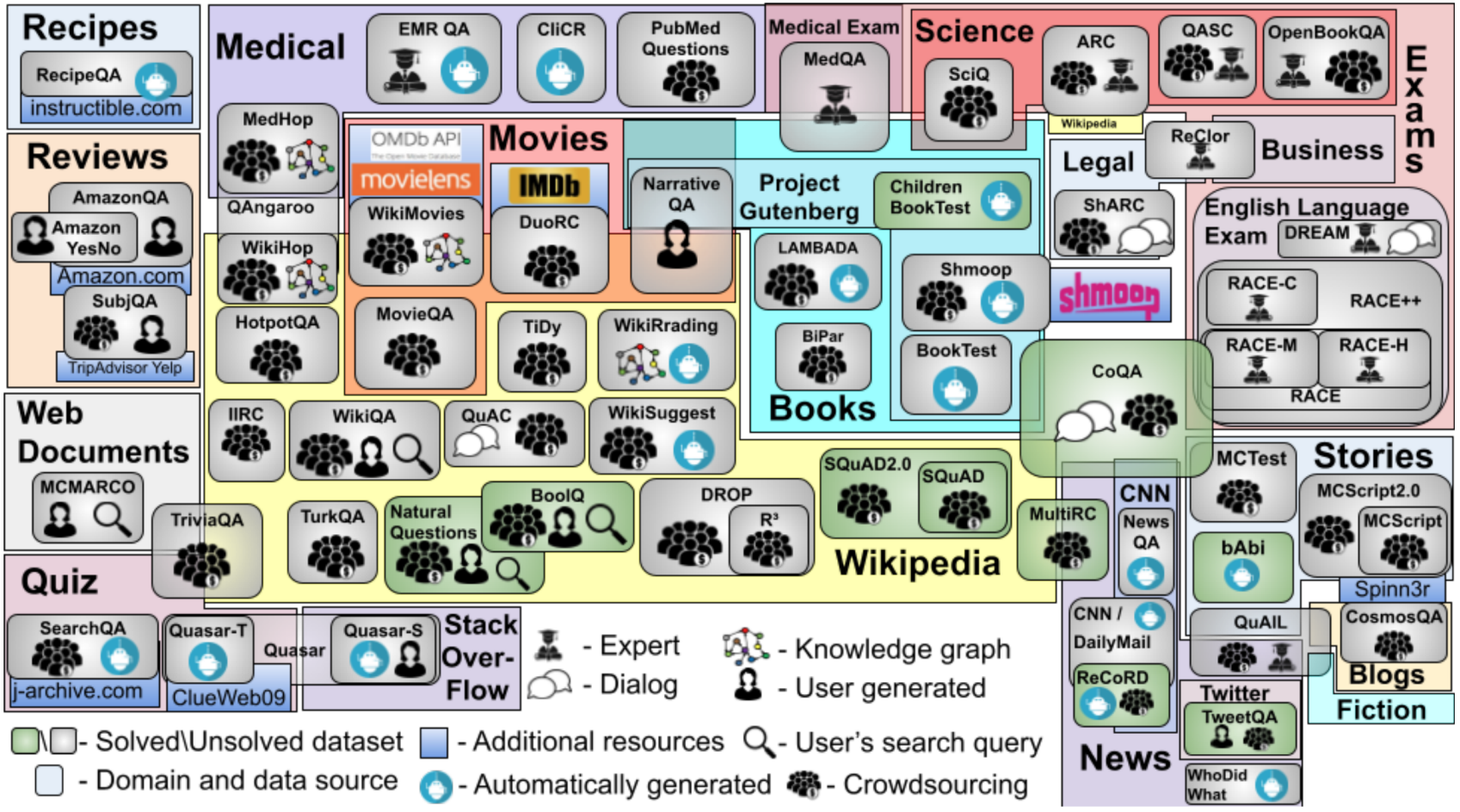

Recipes
RecipeQA
instructible.com
Reviews
AmazonQA
Amazon YesNo
Amazon.com
SubjQA
TripAdvisor Yelp
Web Documents
MCMARCO
Quiz
SearchQA
j-archive.com
Quasar-T
ClueWeb09
Quasar
Quasar-S
Stack Over-Flow
Medical
EMR QA
CliCR
PubMed Questions
MedHop
QAngaroo
WikiHop
HotpotQA
Movies
OMDb API
movielens
WikiMovies
IMDb
DuoRC
MovieQA
TiDy
WikiRrading
Narrative QA
Project Gutenberg
Children BookTest
LAMBADA
Shmoop
BiPar
BookTest
Books
IIRC
WikiQA
QuAC
WikiSuggest
TriviaQA
TurkQA
Natural Questions
BoolQ
DROP
R³
SQuAD2.0
SQuAD
Wikipedia
Medical Exam
MedQA
Science
SciQ
ARC
Wikipedia
QASC
OpenBookQA
Exams
Legal
ReClor
Business
ShARC
English Language Exam
DREAM
RACE-C
RACE++
RACE-M
RACE-H
RACE
shmoop
CoQA
MultiRC
CNN
News QA
CNN / DailyMail
ReCoRD
News
MCTest
Stories
MCScript2.0
MCScript
Spinn3r
bAbi
QuAIL
CosmosQA
Blogs
Fiction
Twitter
TweetQA
WhoDid What
- Expert
- Dialog
- Knowledge graph
- User generated
- Solved\Unsolved dataset
- Additional resources
- User's search query
- Domain and data source
- Automatically generated
- Crowdsourcing

Figure: [5]

SQuAD [7]

- 100K QA pairs over Wikipedia from Mechanical Turk
- Span selection format
- SQuAD2.0 added "Unanswerable" questions [6]

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".

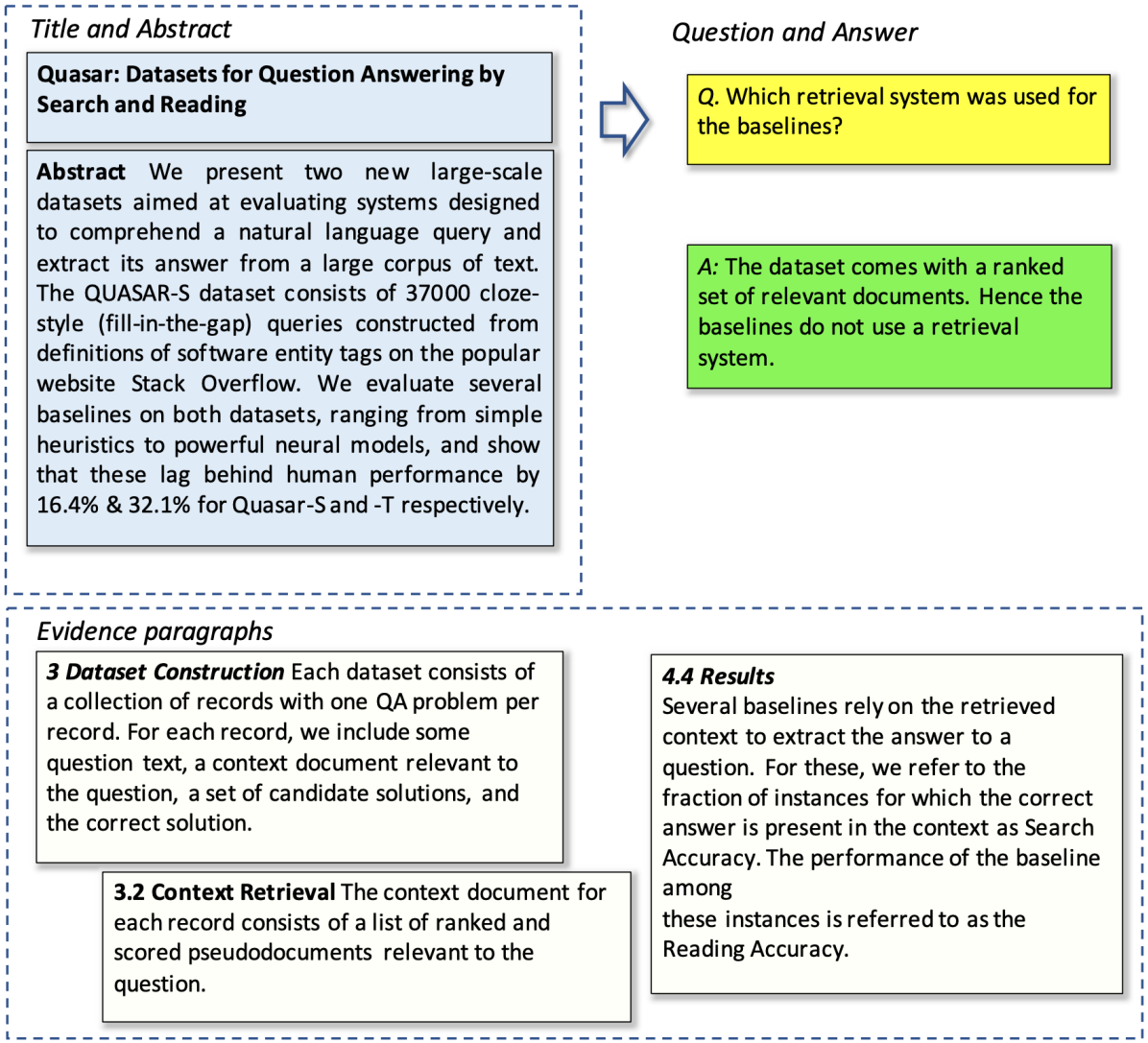
What causes precipitation to fall?
gravity

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

Qasper [8]

- 1.5K papers with 5K information-seeking questions
- Expert annotators, text generation format



DROP [10]

- QA over Wikipedia requiring **discrete reasoning**
- Crowdsourced on Mechnical Turk, span or number answers
- Employed **adversarial filtering** [9]

Reasoning	Passage (some parts shortened)	Question	Answer	BiDAF
Subtraction (28.8%)	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 (18.2%)	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 (19.4%)	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 (11.7%)	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 (16.5%) and Sort (11.7%)	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 (3.7%)	**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**	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his	10	1553

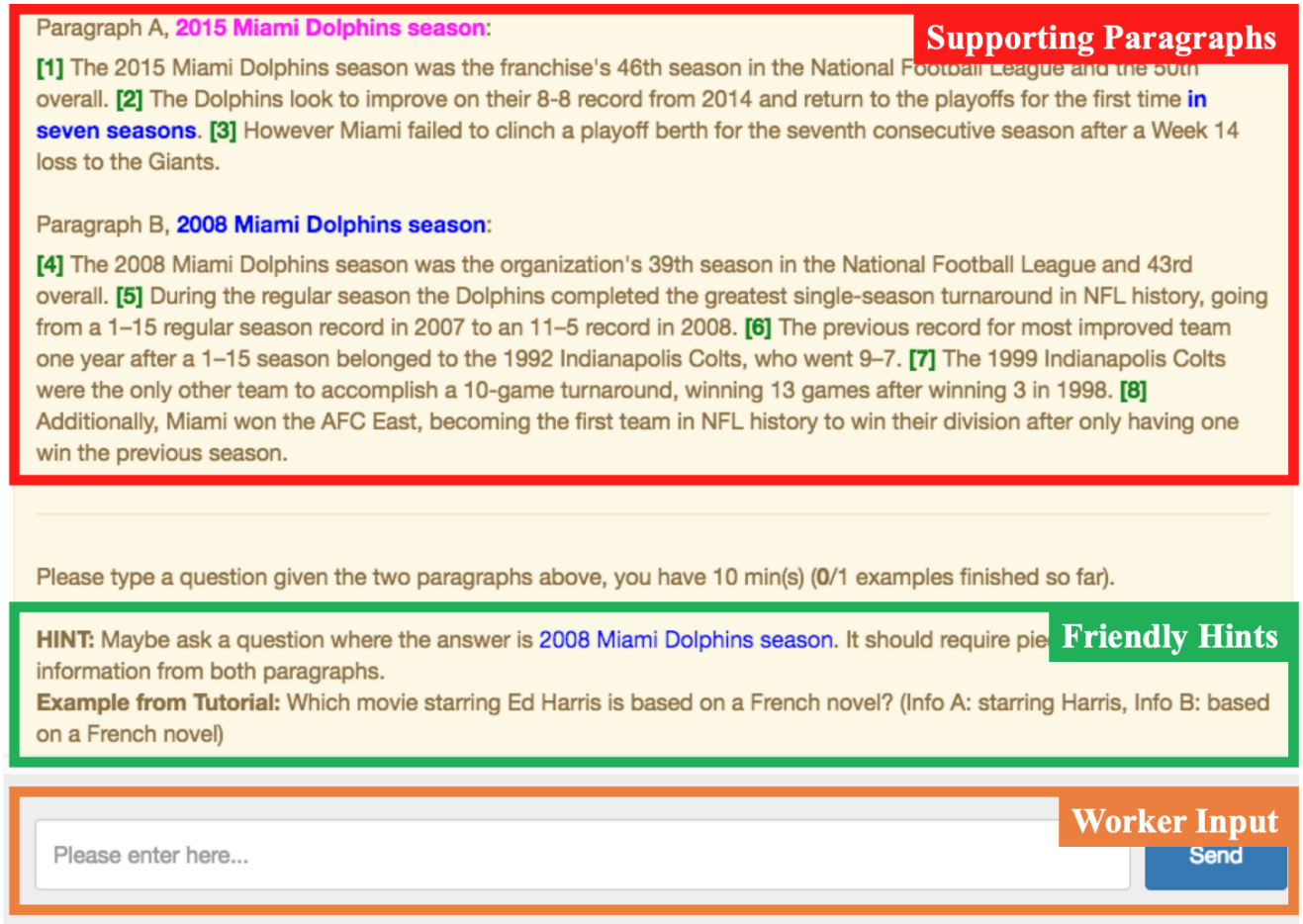
TriviaQA [11]

- 650K QA pairs from **trivia websites**
- Documents retrived **automatically**
  - **Not guarranteed to have the evidence**

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

HotpotQA [12]

- 113k Wikipedia-based QA pairs
- Require **multi-hop reasoning**



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

Natural Questions [13]

- 300K questions taken from actual Google queries
- Annotators choose wikipedia paragraphs from top results
- Answers are either spans, binary, or NULL

Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

Example 3
Question: why does queen elizabeth sign her name elizabeth r
Wikipedia Page: Royal_sign-manual
Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).

Short answer: NULL

Given a long text document(s)
produce a shorter version containing salient information

- E.g., summarizing a scientific paper or news article

Extractive Summarization

- The output summary is formed of contiguous spans
- I.e., **high-lighting** key phrases in the document
- **CNN/Daily Mail** [14] contains 287K news articles

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, ***muhammadu buhari*** told cnn's christiane amanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

Figure: [15]

Abstractive Summarization

- The output summary consists of free form text

Abstract While many approaches to make neural networks more fathomable have been proposed, they are restricted to interrogating the network with input data. [...] In this work, we propose neural persistence, a complexity measure for neural network architectures based on topological data analysis on weighted stratified graphs. [...]

Intro [...] In this work, we present the following contributions: We introduce neural persistence, a novel measure for characterizing the structural complexity of neural networks that can be efficiently computed. [...]

Conclusion [...] However, this did not yield an early stopping measure because it was never triggered, thereby suggesting that neural persistence captures salient information that would otherwise be hidden among all the weights of a network [...]

TLDR We develop a new topological complexity measure for deep neural networks and demonstrate that it captures their salient properties.

Figure: [16]

Other Variants

- **Multi-document summarization** [17]
- **Query-focused summarization** [18]
  - A user specifies a summarization query
- **Interactive summarization** [19]
  - Generalizes query to multi-turn dialog

Information Extraction

Turns unstructured information in texts into structured data

[3][Chap. 21]

Traditional (Closed) IE

- Classify the relation between two given entities [20]
  - **From a set of predefined relations**
- Useful for e.g., **Knowledge base population** [21]

'We are striving to have a strong renewed creative partnership with the CocaCola company,' Mr. Dooner says. However, odds of that happening are slim. since word from Coke headquarters in Atlanta is that CAA and other ad agencies will continue to handle Coke advertising.

Closed IE	(Coke headquarters; **Located-In**; Atlanta)
	(CAA; **Collaborate-With**; Coke)
	(CAA; **Is-A**; ad agency)

Open Information Extraction (Open IE)

- Extracts **stand-alone propositions** from text
  - *Barack Obama, a former U.S president, was **born in** Hawaii*
    (Barack Obama, **was born in**, Hawaii)
    (a former U.S president, **was born in**, Hawaii)
    (Barack Obama, **is**, a former U.S. president)

  - *Obama and Bush were **born in** America*
    (Obama, **born in**, America)
    (Bush, **born in**, America)

Open vs. Closed IE

'We are striving to have a strong renewed creative partnership with the CocaCola company,' Mr. Dooner says. However, odds of that happening are slim. since word from Coke headquarters in Atlanta is that CAA and other ad agencies will continue to handle Coke advertising.

Closed IE	(Coke headquarters; **Located-In**; Atlanta) (CAA; **Collaborate-With**; Coke) (CAA; **Is-A**; ad agency)
Open IE	(Mr. Dooner; **says**; We are striving to have a strong renewed creative partnership with the CocaCola company) (CAA; **continue to handle**; Coke advertising) (other ad companys; **continue to handle**; Coke advertising)

- No need for predefined schema
- Exhaustive vs. consolidated annotation
- Can be annotated effictively using QA [22]

- We've seen various tasks and formats
- Evaluation of model output is often non-trivial

Classification Metrics

- Often straight-forward
- Precision, recall, regression (if ordinal)

Span Selection Metrics

- Span selection often contains some variability
  - E.g., in inclusion / omission of determiners and prepositions
- Metrics include **Exact Match** (EM)
  or **Intersection over union** (IOU)

Text Generation Metrics (Rule-Based)

- Text generation tasks traditionally evaluated n-gram overlaps
  - E.g., abstractive summarization
- Seminal examples include BLEU [23] and ROUGE [24]
- **Useful for many years**,
- **Recently shown to diverge from human annotation**
  - "Stop Using BLEU – Neural Metrics Are Better" [25]

Neural Text Generation Metrics

- Given gold and reference, predict **task-based similarty**
  - Many differnt metrics are proposed
  - **Q: What are the pros & cons of learned metrics?**

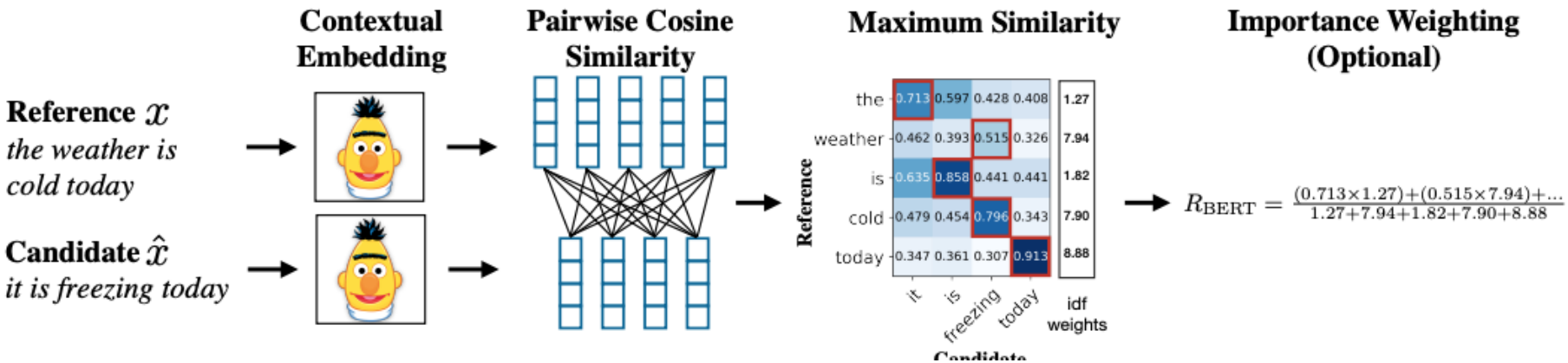


Figure: The BERTScore metric[26]

- Aggregated benchmarks group together different datasets
  - Facilitate testing a model on all of them
- Especially important for general purpose LLM
- Have gradually moved to seq2seq formulation

GLUE[27]

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

SuperGLUE[?]

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

ORB[28]

Dataset	Dev Size	Test Size	Context Length (Avg)	Answer Length (Avg)
SQuAD1.1	10,570	-	123.7	4.0
SQuAD2.0	10,570	-	127.5	4.2
DuoRC	12,233	13,449	1113.6	2.8
Quoref	2,418	2,537	348.2	2.7
DROP	9,536	9,622	195.1	1.5
ROPES	1,204	1,015	177.1	1.2
NewsQA	5,166	5,126	711.3	5.1
NarrativeQA	3,443	10,557	567.9	4.7

Scrolls[29]

Dataset	Task	Domain	Metric	Avg #Words		#Examples
				Input	Output	
GovReport (Huang et al., 2021)	Summ	Government	ROUGE	7,886	492.5	19,402
SummScreenFD (Chen et al., 2021)	Summ	TV	ROUGE	5,598	99.6	4,348
QMSum (Zhong et al., 2021)	QB-Summ	Meetings	ROUGE	9,497	69.7	1,810
Qasper (Dasigi et al., 2021)	QA	Science	F1	3,629	11.4	5,692
NarrativeQA (Kočiský et al., 2018)	QA	Literature, Film	F1	51,653	4.6	71,187
QuALITY (Pang et al., 2021)	MC-QA	Literature, Misc	EM	4,193	10.3	6,737
ContractNLI (Koreeda and Manning, 2021)	NLI	Legal	EM	1,706	1.4	10,319

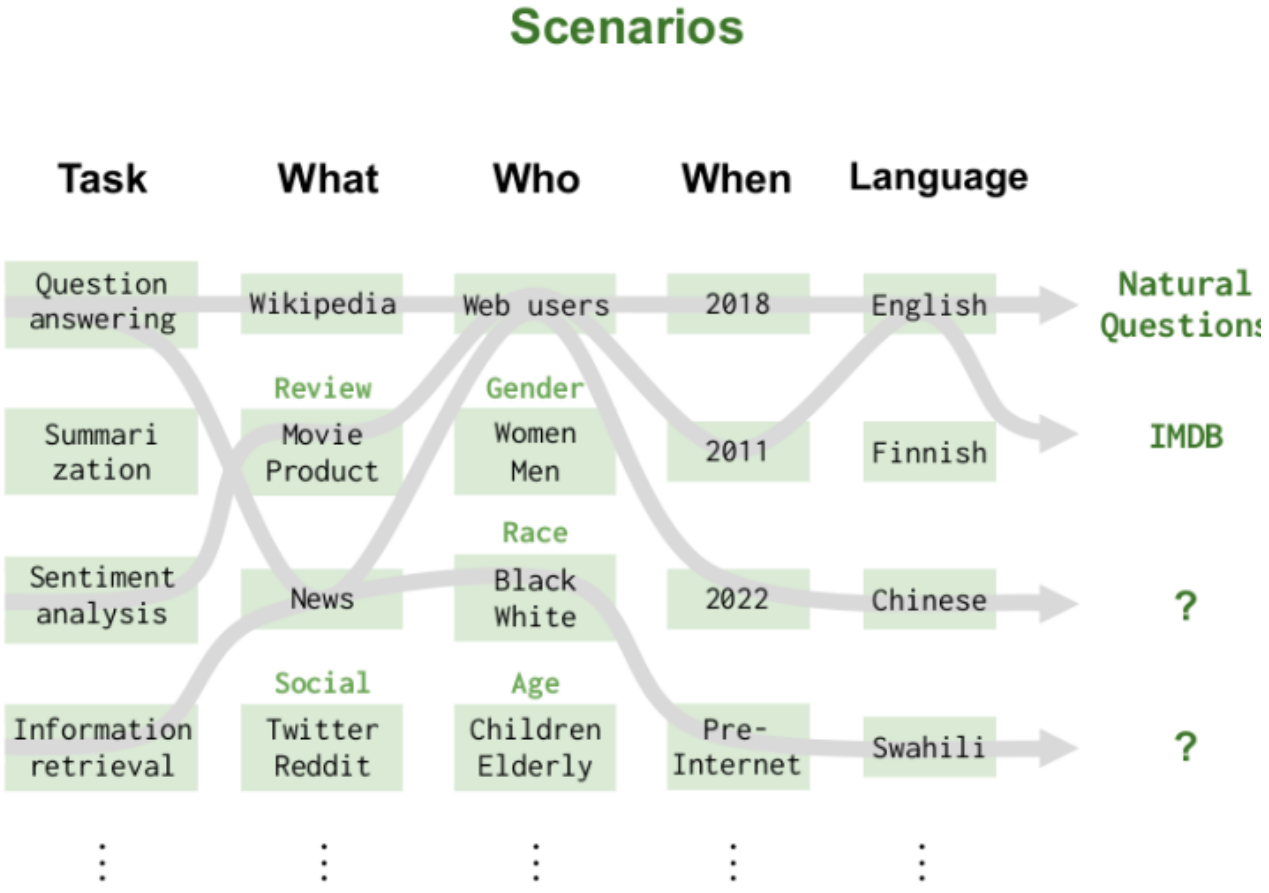
BIG-BENCH[30]

auto_debugging	known_unknowns	parsinlu_reading_comprehension
bbq_lite_json	language_identification	play_dialog_same_or_different
code_line_description	linguistics_puzzles	repeat_copy_logic
conceptual_combinations	logic_grid_puzzle	strange_stories
conlang_translation	logical_deduction	strategyqa
emoji_movie	misconceptions_russian	symbol_interpretation
formal_fallacies_...	novel_concepts	vitaminc_fact_verification
hindu_knowledge	operators	winowhy

Table 1: The 24 tasks included in BIG-bench Lite, a diverse subset of JSON tasks that can be evaluated cheaply.

HELM[31]

- Benchmark of LLMS over 30 different tasks for



References I

Ido Dagan, Oren Glickman, and Bernardo Magnini.
The pascal recognising textual entailment challenge.
In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer, 2006.

Heng Ji and Ralph Grishman.
Knowledge base population: Successful approaches and challenges.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

References II

James H Martin.
Speech and language processing: an introduction to speech recognition, computational linguistics and natural language processing.
daniel jurafsky & 4 n-grams.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts.
Recursive deep models for semantic compositionality over a sentiment treebank.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

References III

Daria Dzendzik, Jennifer Foster, and Carl Vogel.
English machine reading comprehension datasets: A survey.
In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang.
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, July 2018. Association for Computational Linguistics.

References IV

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang.
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, Austin, Texas, November 2016. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner.
A dataset of information-seeking questions and answers anchored in research papers.
In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June 2021. Association for Computational Linguistics.

References V

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi.
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, October-November 2018. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner.
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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

References VI

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer.
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, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning.
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, October-November 2018. Association for Computational Linguistics.

References VII


Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov.
Natural questions: A benchmark for question answering research.
Transactions of the Association for Computational Linguistics, 7:452–466, 2019.


Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom.
Teaching machines to read and comprehend.
Advances in neural information processing systems, 28, 2015.


Abigail See, Peter J. Liu, and Christopher D. Manning.
Get to the point: Summarization with pointer-generator networks, 2017.

References VIII


Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld.
TLDR: Extreme summarization of scientific documents.
In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online, November 2020. Association for Computational Linguistics.


Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz.
Multi-document summarization by sentence extraction.
In *NAACL-ANLP 2000 Workshop: Automatic Summarization*, 2000.


Tal Baumel, Matan Eyal, and Michael Elhadad.
Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models, 2018.

References IX

Ori Shapira, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan, and Ido Dagan.
Interactive abstractive summarization for event news tweets.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 109–114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Nancy Chinchor, Lynette Hirschman, and David D. Lewis.
Evaluating message understanding systems: An analysis of the third message understanding conference (muc-3).
Comput. Linguistics, 19:409–449, 1993.

References X

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning.
Leveraging linguistic structure for open domain information extraction.

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 344–354, Beijing, China, July 2015. Association for Computational Linguistics.

Gabriel Stanovsky and Ido Dagan.
Creating a large benchmark for open information extraction.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, Texas, November 2016. Association for Computational Linguistics.

References XI


Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.


Chin-Yew Lin.
ROUGE: A package for automatic evaluation of summaries.
In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.


Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins.
Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust.

References XII

In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi.
Bertscore: Evaluating text generation with bert.
arXiv preprint arXiv:1904.09675, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman.
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, November 2018. Association for Computational Linguistics.

References XIII

Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Sameer Singh, and Matt Gardner.
Orb: An open reading benchmark for comprehensive evaluation of machine reading comprehension, 2019.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy.
SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Aarohi Srivastava et al.
Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022.

References XIV

 Percy Liang et al.
Holistic evaluation of language models, 2022.