Part III

Datasets & Evaluation

Datasets & Evaluation

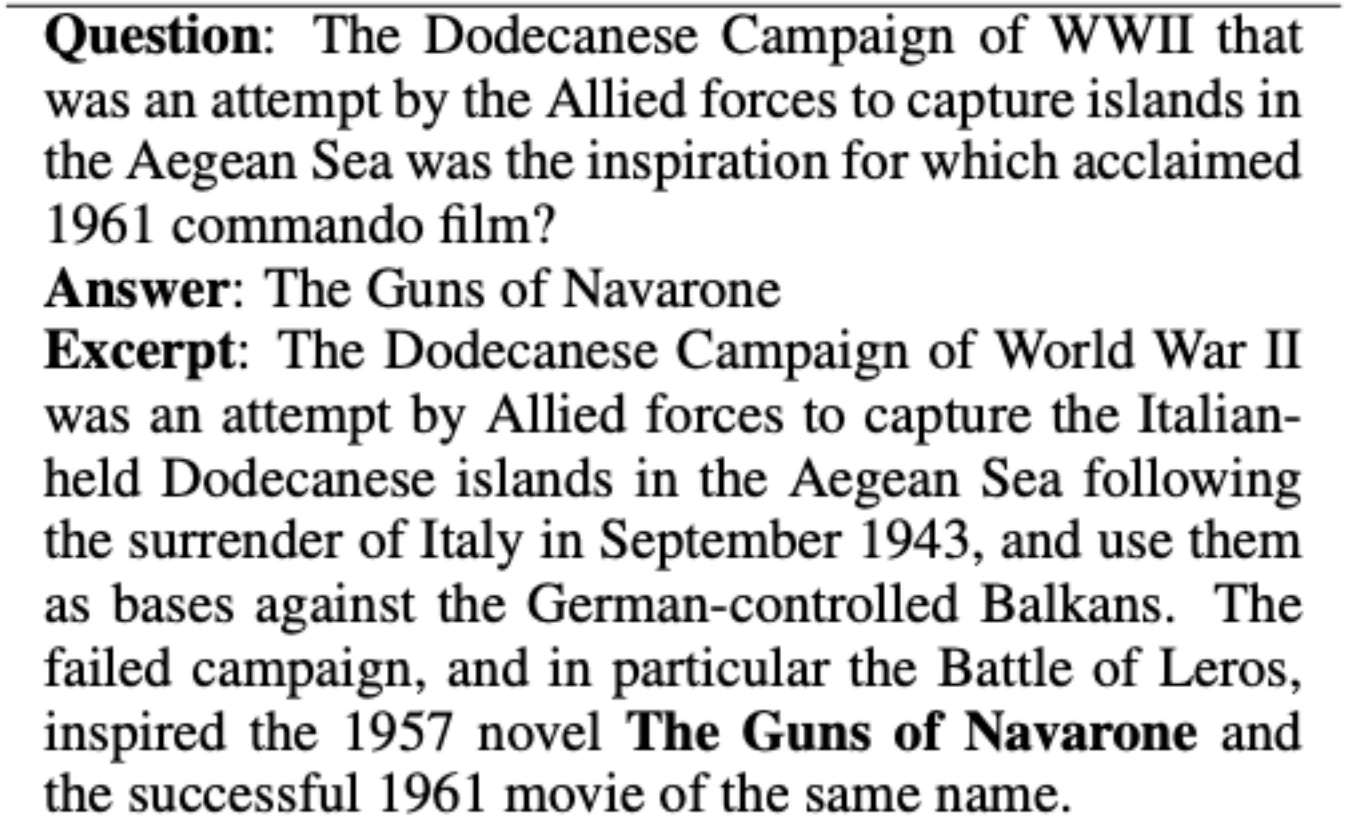
- Datasets popular for open-domain QA
  - TriviaQA [Joshi et al., 2017], SearchQA [Dunn et al., 2017], Quasar-T [Dhingra et al., 2017], Natural Questions [Kwiatkowski et al., 2019]
- Datasets repurposed for open-domain QA
  - SQuAD, CuratedTREC, WebQuestions

- Properties to check
  - Motivation: targeted "task" or "scenario"
  - Source of questions, answers and documents/passages
  - Evaluation metric & methods
  - Limitations when used for evaluating open-domain QA

TriviaQA [Joshi et al., 2017]

Motivation

- Large-scale reading comprehension dataset
- Complex, compositional questions



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

Joshi et al., 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

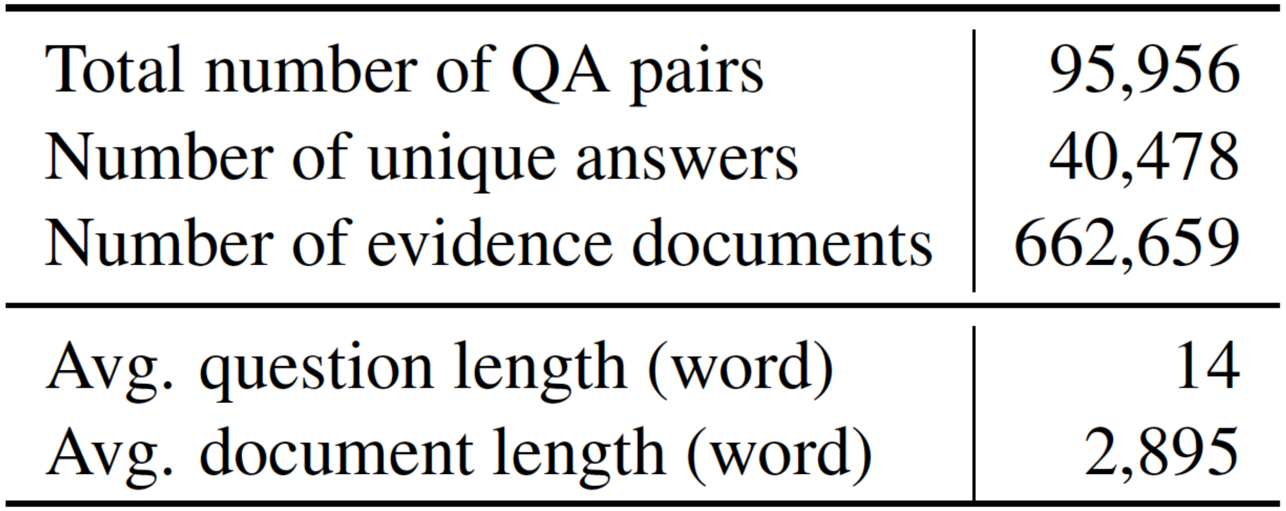
TriviaQA: Collection

- Question-answer pairs
    - 14 trivia and quiz-league Websites
- Textual evidence
    - Web search (Bing)
        - Search query: question
        - Top-10 Web pages (excl. trivia, question, answer, etc.)
    - Wikipedia
        - Identify entities in questions (via TAGME)
        - Add corresponding Wikipedia pages as evidence document
    - Filter documents that do **not** contain the correct answer string

Joshi et al., 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

TriviaQA: Statistics

- Filter documents that do **not** contain the correct answer string



Total number of QA pairs	95,956
Number of unique answers	40,478
Number of evidence documents	662,659
Avg. question length (word)	14
Avg. document length (word)	2,895

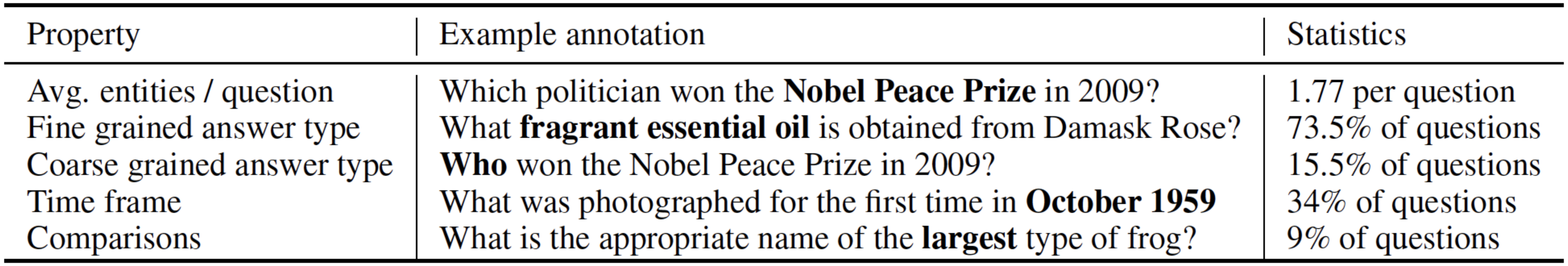
- Full unfiltered dataset
  - 110,495 QA pairs
  - 740k evidence documents

Open-domain Setting

Joshi et al., 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

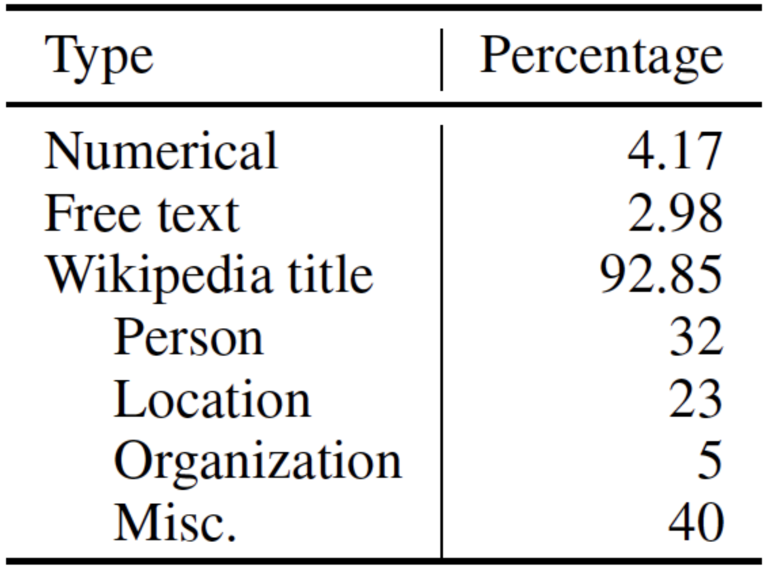
TriviaQA: Distribution

- Analysis based on 200 randomly sampled questions
- Questions



Property	Example annotation	Statistics
Avg. entities / question	Which politician won the **Nobel Peace Prize** in 2009?	1.77 per question
Fine grained answer type	What **fragrant essential oil** is obtained from Damask Rose?	73.5% of questions
Coarse grained answer type	**Who** won the Nobel Peace Prize in 2009?	15.5% of questions
Time frame	What was photographed for the first time in **October 1959**	34% of questions
Comparisons	What is the appropriate name of the **largest** type of frog?	9% of questions

- Answers
  - Wikipedia: Contains answers for 79.7% questions
  - Web: Contains answers for 75.4% questions



Type	Percentage
Numerical	4.17
Free text	2.98
Wikipedia title	92.85
Person	32
Location	23
Organization	5
Misc.	40

Joshi et al., 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

TriviaQA: Evaluation

- SQuAD metrics
  - Exact match (EM)
  - F1 over words in the answer(s).

- Questions that have numerical and free-form answers
  - The given answer

- Questions that have Wikipedia entities as answers
  - The given answer plus Wikipedia aliases

Joshi et al., 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

SearchQA [Dunn et al., 2017]

Motivation

- A general question-answering system should be open-domain
- Use search snippets as the context

Question: Guinness says that by number of users this language, devised by fictional language
Answer: Klingon
Snippet: The Klingons are a fictional extraterrestrial humanoid warriors … A dictionary, a book of sayings, and a cultural guide to the language have portrayed Montgomery Scott, devised the … of Guinness World Records, Klingon language by…

Dunn et al., 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine

SearchQA: Collection

Question-Answer Pairs

- Jeopardy! (J! Archive)

Textual evidence

- Web search (Google)
  - Search query: question-answer pair
- Snippets after some post-processing: removing Jeopardy! related
  - The air-date of the Jeopardy! episode
  - Exact copy of question
  - Terms "Jeopardy!", "quiz" or "trivia"

Dunn et al., 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine

SearchQA: Statistics

140,461 question-answer pairs

- Each pair is with 49.6 ± 2.10 snippets
- Each snippet is 37.3 ± 11.7 tokens

No learning from the future!

- Training, Validation, Test sets from non-overlapping years.
- The validation and test question-answer pairs are from years later than the training set's pairs.

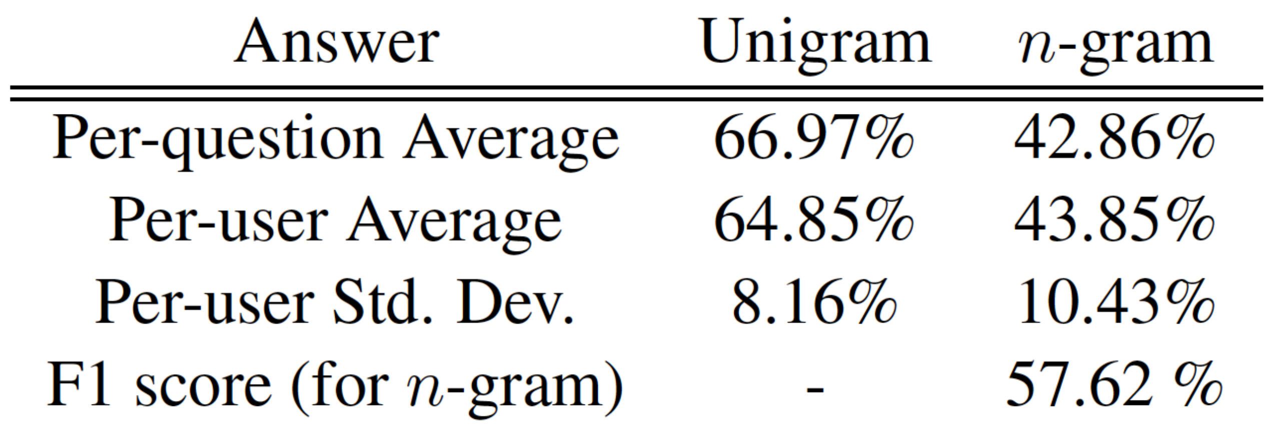
Split	# Examples
Training	99,820
Validation	13,393
Test	27,248

Dunn et al., 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine

SearchQA: Evaluation

- Single-word (unigram) answers
  - Top-1 & Top-5 accuracies
- Multi-word (*n*-gram) answers
  - F1 scores

- Human performance



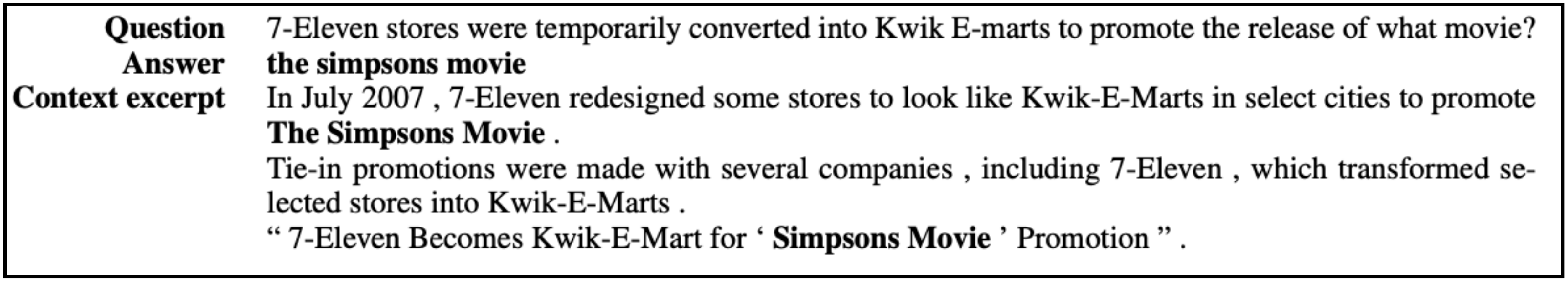
Answer	Unigram	n-gram
Per-question Average	66.97%	42.86%
Per-user Average	64.85%	43.85%
Per-user Std. Dev.	8.16%	10.43%
F1 score (for n-gram)	-	57.62 %

Dunn et al., 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine

Quasar-T [Dhingra et al., 2017]

Motivation

- Large-scale datasets for evaluating end-to-end QA systems
  - Search, aggregate information from multiple passages, extract answers
- Question answer by search and reading
- Quasar-T is based on trivia questions



Question	7-Eleven stores were temporarily converted into Kwik E-marts to promote the release of what movie?
Answer	**the simpsons movie**
Context excerpt	In July 2007 , 7-Eleven redesigned some stores to look like Kwik-E-Marts in select cities to promote **The Simpsons Movie** . Tie-in promotions were made with several companies , including 7-Eleven , which transformed selected stores into Kwik-E-Marts . “ 7-Eleven Becomes Kwik-E-Mart for ‘ **Simpsons Movie** ’ Promotion ” .

Dhingra et al., 2017. Quasar: Datasets for Question Answering by Search and Reading

Quasar-T: Collection

Question-Answer Pairs

- Collected by Reddit user 007craft and released in Dec 2015
- Remove True/False and multi-choice questions
- Most answers are noun phrases

Textual Evidence

- Source: ClueWeb09 [Callan et al., 2009]
  - 1 billion web pages collected between Jan. and Feb. 2009
- Phase 1: 100 documents from ClueWeb09 batch query service
  - Query: Question + Answer
  - Long context: 2048 characters, short context: 200 characters
- Phase 2: Top pseudo-documents that contain the answer using Lucene
  - Query: Question
  - 20 long context & 100 short context per question

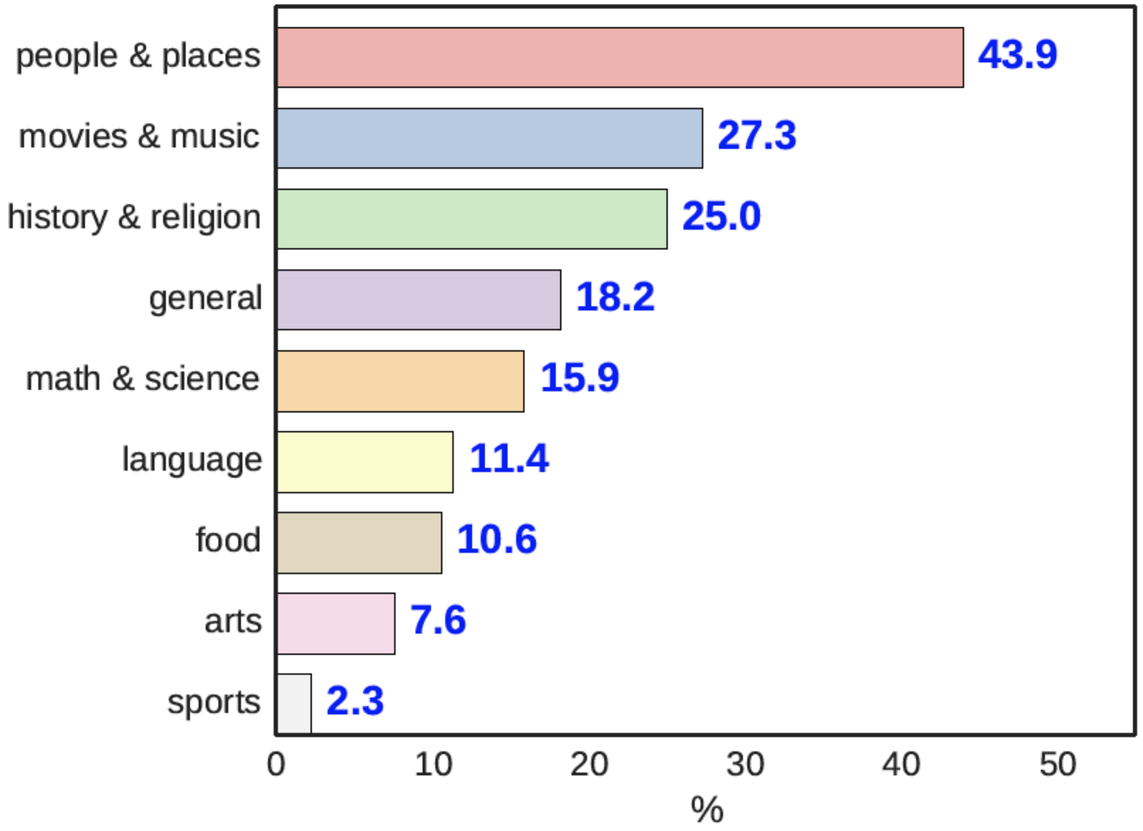
Dhingra et al., 2017. Quasar: Datasets for Question Answering by Search and Reading

Quasar-T: Statistics

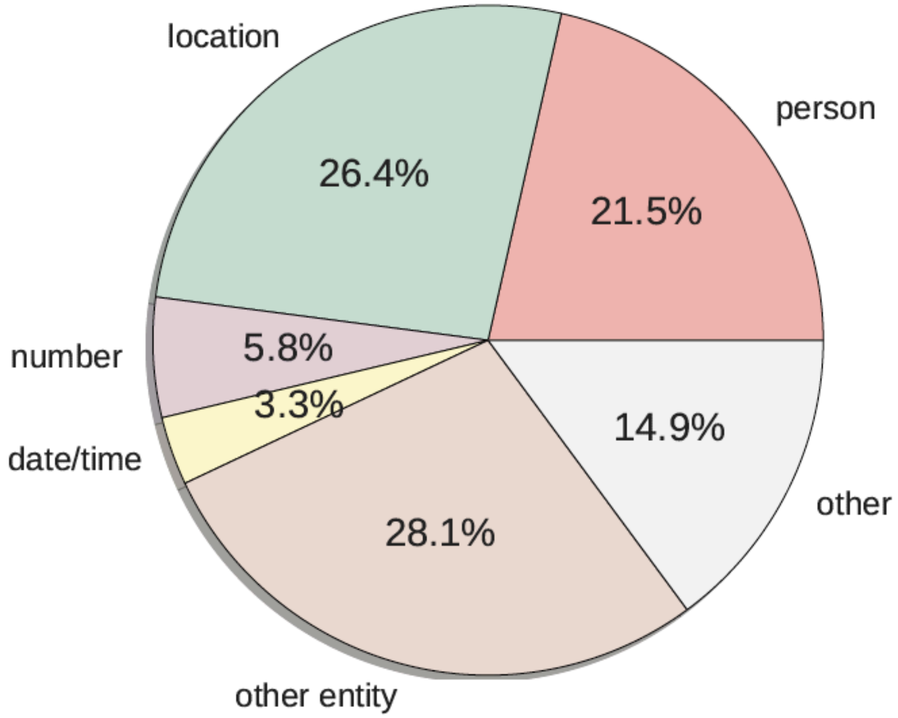
	Total	Single-Word Answer	Answer in Short Context	Answer in Long Context
Train	37,012	18,726	25,465	26,318
Validation	3,000	1,507	2,068	2,129
Test	3,000	1,508	2,043	2,102

Dhingra et al., 2017. Quasar: Datasets for Question Answering by Search and Reading

Quasar-T: Distribution



Question genres



Answer categories

Dhingra et al., 2017. Quasar: Datasets for Question Answering by Search and Reading

Quasar-T: Evaluation

- SQuAD Metrics
  - Exact match (EM)
  - F1 over words in the answer(s).

- Exact match measures whether the two strings, after preprocessing, are equal or not.

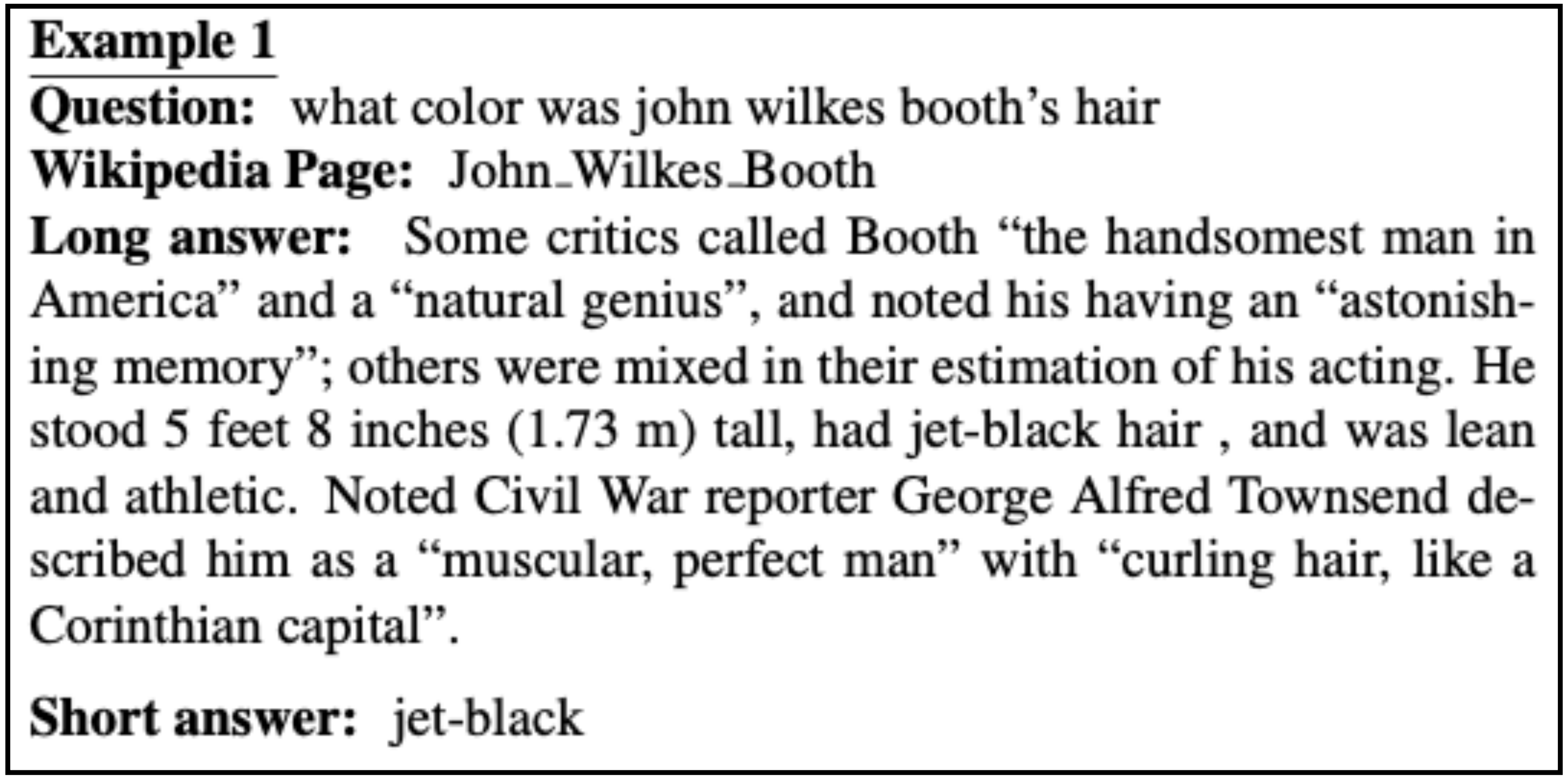
- F1 measures the overlap between the two bags of tokens in answers, after preprocessing

Dhingra et al., 2017. Quasar: Datasets for Question Answering by Search and Reading

Natural Questions [Kwiatkowski et al., 2019]

Motivation

- Large-scale end-to-end training data for QA
- "Natural" questions from search engine query logs



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

Kwiatkowski et al., 2019. Natural Questions: A Benchmark for Question Answering Research

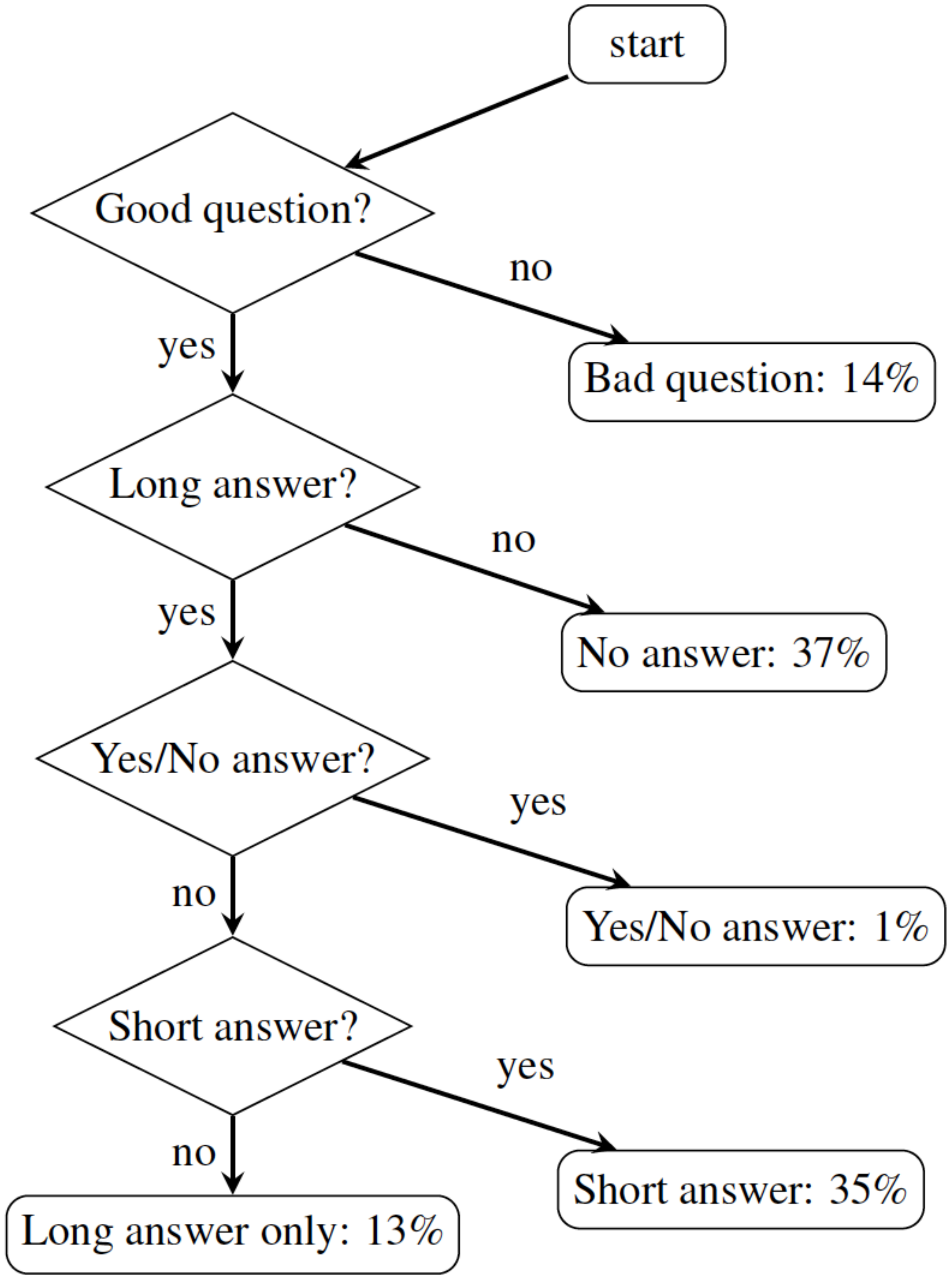
Natural Questions: Collection

- Question source: Google search queries
    - Queries of 8 words or more, by multiple users in a short period of time
- Answer source: Wikipedia page from top 5 search results
    - Long answer: paragraph, table, list (HTML bounding box)
    - Short answer: span(s), yes/no, NULL
- Annotation: a pool of ~50 annotators

Kwiatkowski et al., 2019. Natural Questions: A Benchmark for Question Answering Research

Natural Questions: Statistics

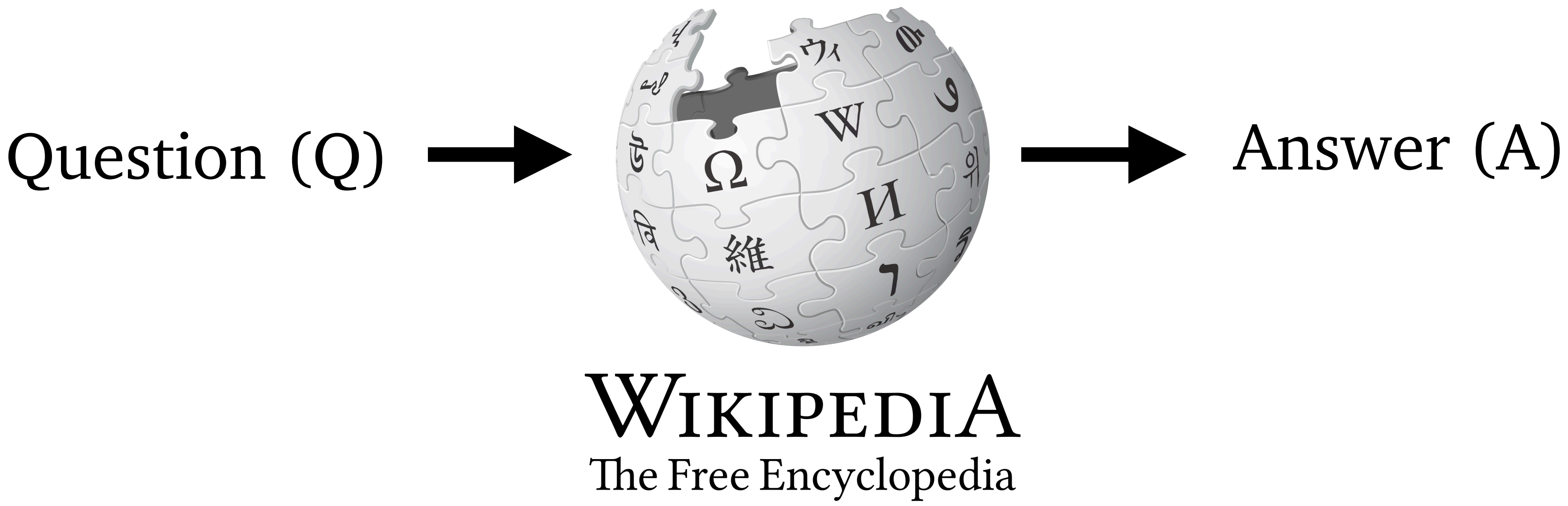
- Train: 307,373 examples with single annotations
- Dev: 7,830 examples with 5-way annotations
- Test: 7,842 examples with 5-way annotations (sequestered)



Kwiatkowski et al., 2019. Natural Questions: A Benchmark for Question Answering Research

Open-Domain QA Evaluation

$$\mathcal{D}_{\mathrm{QA}} = \{(Q_i, A_i)\}$$



Question (Q) → Wikipedia → Answer (A)

WIKIPEDIA

The Free Encyclopedia

[Chen et al., 2017; Lee et al., 2019]

- The correctness of the supporting evidence is not evaluated
- Dataset and Wikipedia dump may not be created at the same time

Open-Domain QA Datasets

used in ORQA [Lee et al., 2019]

- Natural Questions
  - Questions with short answers (<5 tokens)
- WebQuestions [Berant et al., 2013]
  - Questions sampled using Google Suggest API
  - Answers are Freebase entities
- CuratedTREC [Baudis & Sedivy, 2015]
  - Questions from TREC-QA; askers do not observe evidence doc.
- TriviaQA
  - Questions from the unfiltered set (i.e., all questions)
- OpenSQuAD [Rajpurkar et al., 2016]
  - Questions from SQuAD v1.1; askers do see the context (Wikipedia paragraph)

Dataset	Train	Val	Test
NQ	79,168	8,757	3,610
WebQ	3,417	361	2,032
TREC	1,353	133	694
TriviaQA	78,785	8,837	11,313
SQuAD	78,713	8,886	10,570

Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering