SIC Saarland Informatics Campus

UNIVERSITÄT DES SAARLANDES

max planck institut informatik

# Question Answering Systems

## Benchmarks that made a difference

**Rishiraj Saha Roy**
Max Planck Institute for Informatics, Germany

6

Saarland University, Summer Semester 2020

# Question of the day

How were some of the most popular QA benchmarks created?

# You'll find this covered in

1. SQuAD: 100,000+ Questions for Machine Comprehension of Text
   - Rajpurkar et al.
   - EMNLP 2016 NAACL ACL, EMNLP, COLING, EACL..
   - https://www.aclweb.org/anthology/D16-1264.pdf

Percy Liang

2. Semantic Parsing on Freebase from Question-Answer Pairs
   - Berant et al.
   - EMNLP 2013
   - https://www.aclweb.org/anthology/D13-1160.pdf

~~SEMPRE~~ + Web Questions benchmark

# Why are benchmarks great?

1 ▪ Large benchmarks drive concerted progress in the community

2 ▪ Standardizes task what exactly should we do?

3 ▪ Promotes use of uniform metrics for comparison

4 ▪ Enables fair comparisons

5 ▪ Avoids additional re-implementation efforts

Dataset

M1 M2 M3

Q? A? R? → D?

↑ NDCG, ERR, MRR ↓ bpref, ...

ImageNet

PTB

Q1 A1 R

Q2 A2 Test

corpora/KG

practical

code

M0 0.7

M1 0.5

M2 0.4

M3 0.6

R → R'

# Why are benchmarks great?

- Risk of overfitting
  - "Milking" benchmark
- Reduces creativity
  - What are the real problems?
- Incremental progress
- Reducing effort for contribution?
  - The "Pick task + pick benchmark + improve" paradigm

# Characteristics of good benchmarks

Pros >> Cons

1. Large!! (How large is large?) — relative; neural models → data-hungry ⇒ as large as you can; the larger the better; time + money
2. Vision of where the community wants to go – quantify difficulty!
3. Realistic! — Real question, Real users; simple QA → quickly saturation!!; complex, dialog
4. Diverse → length, topic (open-domain), complexity, ... → diversity (linguistic)
5. Low baseline (current) performance — simple; → Benchmark + Baseline 90%; power + responsibility!
6. Defines clear metric(s) — F1? MRR? (main) preamble + ....
7. Specifies train-dev-test splits — Wsdeyn n/o

| Method | M1 | M2 |
|---|---|---|
| | | ?? |

time + Scope / Room for improvement?

Train (B) Dev (B) Test (B) Benchmark B

# Research paper 1

large

## SQuAD: 100,000+ Questions for Machine Comprehension of Text

Text-QA ≅ Machine Reading comprehension

MRC/RC

Squad: 100,000+ questions for machine comprehension of text [PDF] arxiv.org

P Rajpurkar, J Zhang, K Lopyrev, P Liang - arXiv preprint arXiv …, 2016 - arxiv.org

SF Prof.

We present the Stanford Question Answering Dataset (SQuAD), a new reading comprehension dataset consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage. We analyze the dataset to understand the types of reasoning required to answer the questions, leaning heavily on dependency and constituency trees. We build a strong logistic regression model, which achieves an F1 score …

☆ ❝❝ Cited by 1586 Related articles All 16 versions »

Google Scholar

# The SQuAD effect

- Revived (factoid) text-QA
  - Make QA great again!
- Paved way for "open-domain QA"
  - Goal morphed over time
  - Changed connotations!
- Changed the perception of (text) QA
- The rise of the leaderboard

# “QA Billboard!”

Previously: dataset ≡ link on a website
↓
download

Concerted progress!

Time

60

https://rajpurkar.github.io/SQuAD-explorer/

SQuAD | Home | Explore 2.0 | Explore 1.1

# SQuAD2.0

The Stanford Question Answering Dataset

## What is SQuAD?

**S**tanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

**SQuAD2.0** combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

**SQuAD 1.1,** the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

## Getting Started

## Leaderboard

who stands where

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<br>*Stanford University*<br>(Rajpurkar & Jia et al. '18) | 86.831 | 89.452 |
| 1<br>Apr 06, 2020 | SA-Net on Albert (ensemble)<br>*QIANXIN* | **90.724** | **93.011** |
| 2<br>May 05, 2020 | SA-Net-V2 (ensemble)<br>*QIANXIN* | 90.679 | 92.948 |
| 2<br>Apr 05, 2020 | Retro-Reader (ensemble)<br>*Shanghai Jiao Tong University*<br>http://arxiv.org/abs/2001.09694v2 | 90.578 | 92.978 |
| 3<br>May 04, 2020 | ELECTRA+ALBERT+EntitySpanFocus (ensemble)<br>*SRCB_DML* | 90.442 | 92.839 |
| 4<br>Mar 12, 2020 | ALBERT + DAAF + Verifier (ensemble)<br>*PINGAN Omni-Sinitic* | 90.386 | 92.777 |
| 5<br>Jan 10, 2020 | Retro-Reader on ALBERT (ensemble)<br>*Shanghai Jiao Tong University*<br>http://arxiv.org/abs/2001.09694v2 | 90.115 | 92.580 |
| 6<br>Nov 06, 2019 | ALBERT + DAAF + Verifier (ensemble)<br>*PINGAN Omni-Sinitic* | 90.002 | 92.425 |

w/ exp + code

# SQuAD: Overview

Back at 15:15

- Large: 100,000+ question, answer pairs!
- From Stanford
- Leverages Wikipedia
- Relies on crowdworkers

crowdsourcing → AMT
CrowdFlower
odesk

prescribe

- Metrics: Exact match, F1

P, R, F

- Now version 2.0!

unanswerable

# Example

SQuAD: QA task

↓

Span prediction

↓

classifier

→ sentences
→ doc
→ passage

- tuple ?
- entity

given doc as input

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".

Q1 → What causes precipitation to fall?
**gravity**

Q2 → What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Q3 → Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

start end

arbitrary span

ans → not only entities

**Figure 1:** Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

# Before SQuAD

Hirschman et al 1999
- school RC question
- AI good
- smart as a 6th grader!

QA ≡ fill-in-the-blank
— is the president of USA
→ Cloze

question → I(Q)
D(S) sentence prediction

| Dataset | Question source | Formulation | Size |
|---|---|---|---|
| **SQuAD** | **crowdsourced** | **RC, spans in passage** | **100K** |
| MCTest (Richardson et al., 2013) | crowdsourced | RC, multiple choice | 2640 |
| Algebra (Kushman et al., 2014) | standardized tests | computation | 514 |
| Science (Clark and Etzioni, 2016) | standardized tests | reasoning, multiple choice | 855 |
| WikiQA (Yang et al., 2015) | query logs | IR, sentence selection | 3047 |
| TREC-QA (Voorhees and Tice, 2000) | query logs + human editor | IR, free form | 1479 |
| CNN/Daily Mail (Hermann et al., 2015) | summary + cloze | RC, fill in single entity | 1.4M |
| CBT (Hill et al., 2015) | cloze | RC, fill in single word | 688K |

**Table 1:** A survey of several reading comprehension and question answering datasets. SQuAD is much larger than all datasets except the semi-synthetic cloze-style datasets, and it is similar to TREC-QA in the open-endedness of the answers.

# Dataset collection

1. Passage curation
2. Question answer collection
3. Additional answers

# Passage curation

- Top 10000 articles from Project Nayuki
- Sampled subset
- Split into paragraphs
- Discard short paragraphs
- Ensure topic diversity
- Split into train-dev-set splits (why now?)

# QA collection

Form

- Use AMT + Daemo platform
- What is AMT?

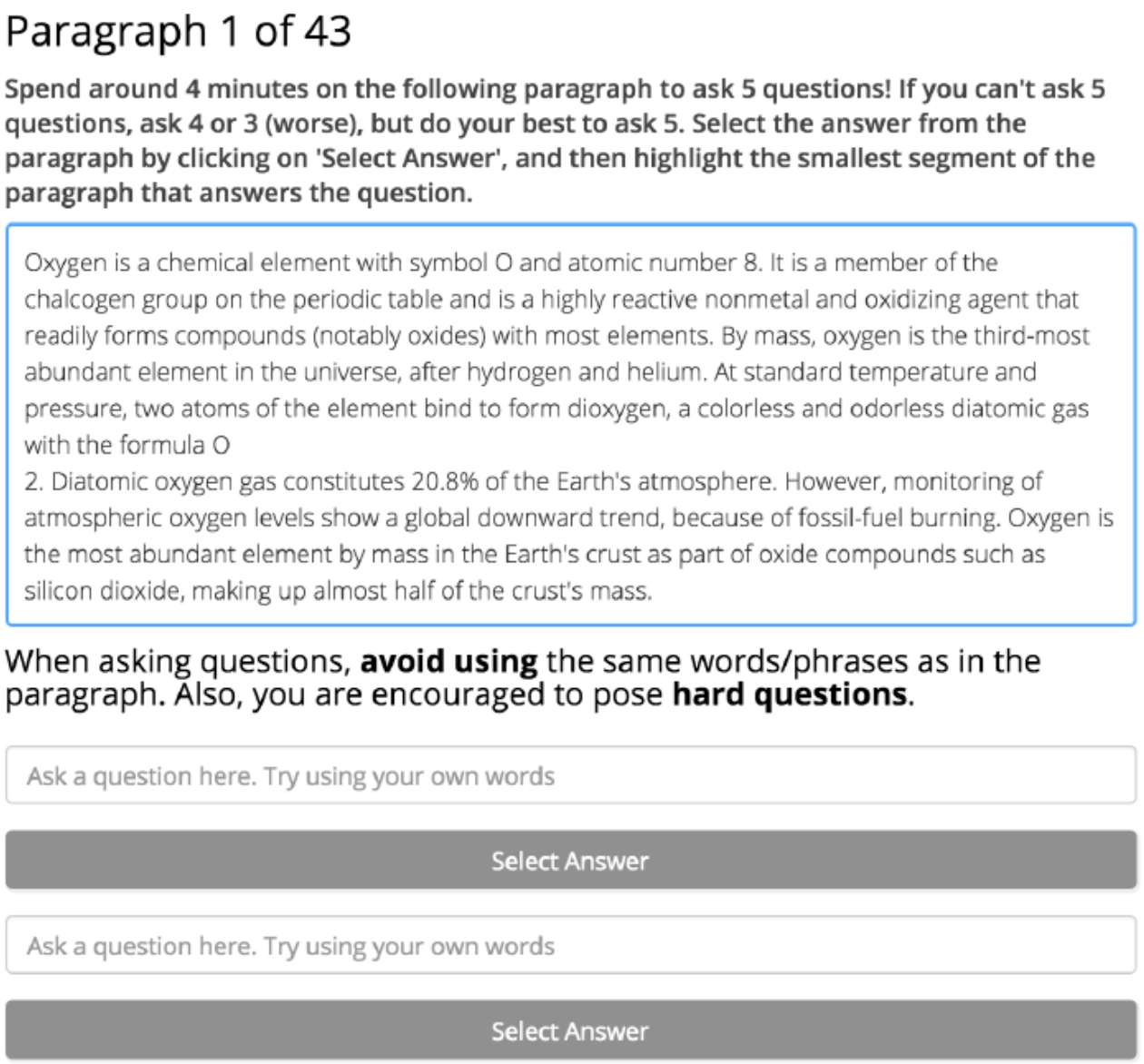



**Figure 2:** The crowd-facing web interface used to collect the dataset encourages crowdworkers to use their own words while asking questions.

Overview Features Pricing Help Developer Resources Customers Sign in as a Requester

# Amazon Mechanical Turk

## Access a global, on-demand, 24x7 workforce

Get started with Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development.

While technology continues to improve, there are still many things that human beings can do much more effectively than computers, such as moderating content, performing data deduplication, or research. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time consuming, expensive and difficult to scale, or have gone undone. Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the Internet (also known as 'microtasks').

Turker/
Worker

Requester

https://www.mturk.com/

# QA collection

backend

- Use AMT + Daemo platform

Mechanical Turker

- What is AMT?

1600 HITs

Guidelines + Examples

  - Filters: Spam control

97% acceptance

  - 5 questions
  - Payment/hour $9 → $10-12

- Own words! Disable ^C + ^V!
- Mark answer

Paragraph 1 of 43

Guidelines

**Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.**

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O
2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words

Select Answer

Ask a question here. Try using your own words

Select Answer

**Figure 2:** The crowd-facing web interface used to collect the dataset encourages crowdworkers to use their own words while asking questions.

# Additional answers

at least 2

- Questions are not individualized
- Easy to increase annotations + sanity check

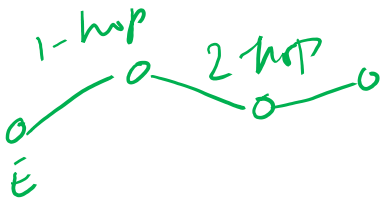
  rich

  reduce spam | mistake
- 5 questions in two minutes – calculate per-hour cost
- Shortest span (a b c d)
- Identify unanswerable questions!
- Smart idea!

# Dataset analysis

- Characterization – property of good benchmark!
- 1 Diversity in answers
- 2 Reasoning required to answer questions
- 3 Stratification by syntactic divergence

difficulty



# Answer type analysis

| Answer type | Percentage | Example |
|---|---|---|
| Date | 8.9% | 19 October 1512 |
| Other Numeric | 10.9% | 12 |
| Person | 12.9% | Thomas Coke |
| Location | 4.4% | Germany |
| Other Entity | 15.3% | ABC Sports |
| Common Noun Phrase | 31.8% | property damage |
| Adjective Phrase | 3.9% | second-largest |
| Verb Phrase | 5.5% | returned to Earth |
| Clause | 3.7% | to avoid trivialization |
| Other | 2.7% | quietly |

**Table 2:** We automatically partition our answers into the following categories. Our dataset consists of large number of answers beyond proper noun entities.

# Reasoning

| Reasoning | Description | Example | Percentage |
|---|---|---|---|
| Lexical variation (synonymy) | Major correspondences between the question and the answer sentence are synonyms. | Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle. | 33.3% |
| Lexical variation (world knowledge) | Major correspondences between the question and the answer sentence require world knowledge to resolve. | Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process. | 9.1% |
| Syntactic variation | After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications. | Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington. | 64.1% |
| Multiple sentence reasoning | There is anaphora, or higher-level fusion of multiple sentences is required. | Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance. | 13.6% |
| Ambiguous | We don't agree with the crowdworkers' answer, or the question does not have a unique answer. | Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment. | 6.1% |

# Syntactic divergence

Q: What department store is thought to be the first in the world?
S: Bainbridge's is often cited as the world's first department store.

Path:

first $\xleftarrow{\text{xcomp}}$ thought $\xrightarrow{\text{nsubjpass}}$ store $\xrightarrow{\text{det}}$ what

$\Downarrow$delete $\Downarrow$substitute $\Downarrow$insert

first $\xleftarrow{\text{amod}}$ store $\xleftarrow{\text{nmod}}$ cited $\xrightarrow{\text{nsubjpass}}$ Bainbridge's

Edit cost:

1 +2 +1=4

**Figure 3:** An example walking through the computation of the syntactic divergence between the question Q and answer sentence S.

# Research paper 2

## Semantic Parsing on Freebase from Question-Answer Pairs

[PDF] Semantic parsing on freebase from question-answer pairs [PDF] aclweb.org

J Berant, A Chou, R Frostig, P Liang - Proceedings of the 2013 …, 2013 - aclweb.org

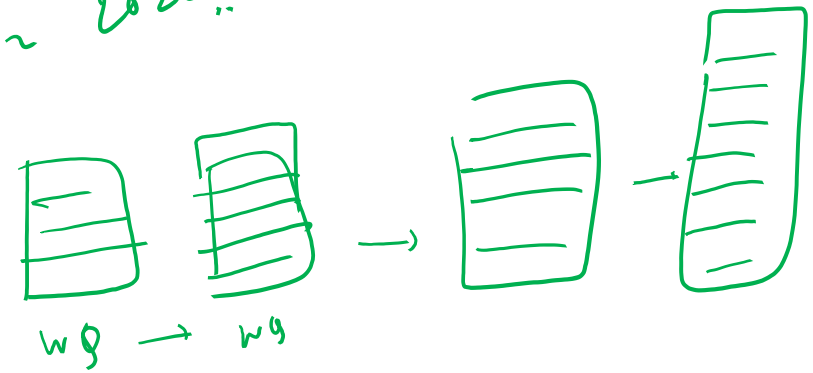
In this paper, we train a semantic parser that scales up to Freebase. Instead of relying on annotated logical forms, which is especially expensive to obtain at large scale, we learn from question-answer pairs. The main challenge in this setting is narrowing down the huge number of possible logical predicates for a given question. We tackle this problem in two ways: First, we build a coarse mapping from phrases to predicates using a knowledge base and a large text corpus. Second, we use a bridging operation to generate additional …

☆ 99 Cited by 915 Related articles All 18 versions »

# The WebQuestions effect

KB-QA

- Paved the way for the KG-QA community
- Passed the test of time ☺ 2013 still in use in 2020!!
- Introduced "in-paper leaderboard" for QA
  standard
- Real questions by real users
- Largest at the time (but noisy ☹ ) ~5K – 6K
  - Sparked improvements! Web Question SP
- Suitable for supervised + neural methods



# Key problem

- How to get factoid questions from the Web?
- Need experts? Explain KG? 917
- Leverage CQA? Super noisy!
- Search logs: rich resource
- But how to get them??

Master, Jake

# Key idea: The Google Suggest API

Google

popular

where was

- where was **jesus born** F
- where was **lost filmed** F
- where was **hitler born** F
- where was **dark filmed** F
- where was **pizza invented** F
- where was **troy**
- where was **jesus buried** ?
- where was **broadchurch filmed**
- where was **harry potter filmed**
- where was **vikings filmed**

Google Search | I'm Feeling Lucky

*Report inappropriate predictions*
Learn more

Home > Products > G Suite Developer > Cloud Search > Reference

☆☆☆☆☆

## Method: query.suggest

Send feedback

Provides suggestions for autocompleting the query.

**Note:** This API requires a standard end user account to execute. A service account can't perform Query API requests directly; to use a service account to perform queries, set up G Suite domain-wide delegation of authority.

### HTTP request

`POST https://cloudsearch.googleapis.com/v1/query/suggest`

The URL uses gRPC Transcoding syntax.

### Request body

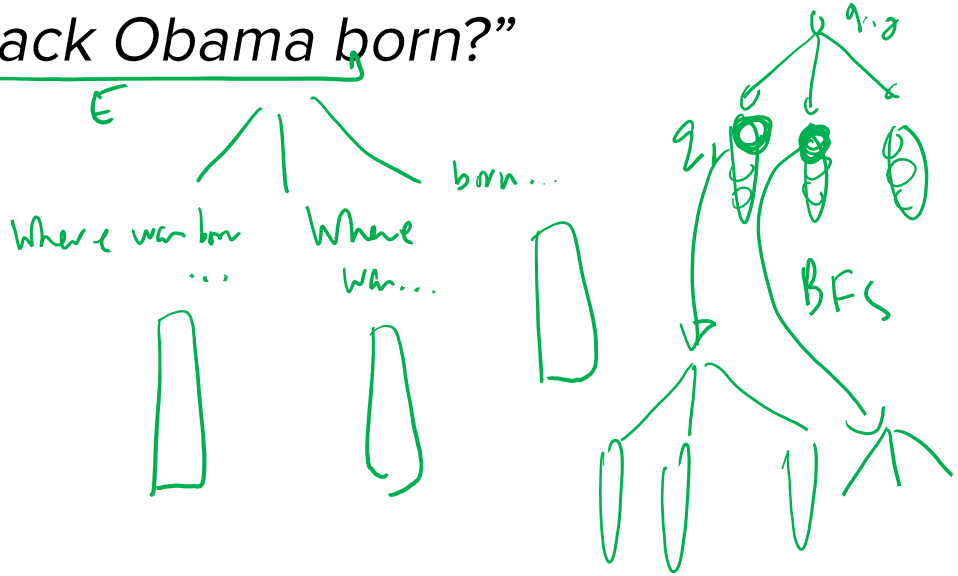
The request body contains data with the following structure:

JSON representation

```
{
  "requestOptions": {
    object (RequestOptions)
  },
  "query": string,
  "dataSourceRestrictions": [
    {
      object (DataSourceRestriction)
    }
  ]
}
```

# Data collection outline

- Start from seed question: *“Where was Barack Obama born?”*
- Use variations into Google API
- Do breadth-first search
- Expand queue
- Stop until 1M questions!



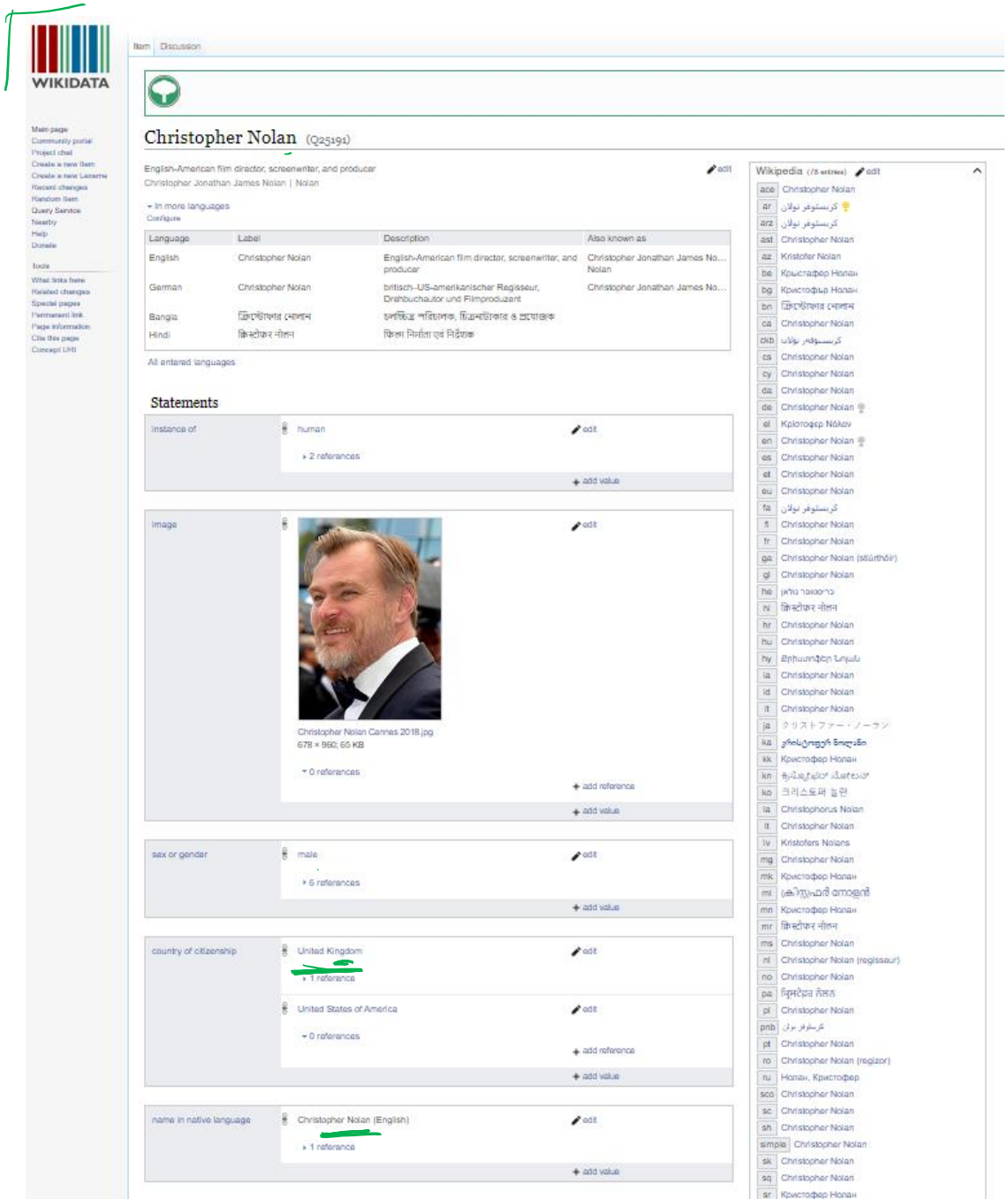
2+ SPARQL

2-2 pair

again

# AMT to the rescue

- Turker personifies average Web user!
- Answer questions (if you can!)
  - Entities
  - Values (literals)
  - Lists
  - Else mark unanswerable
- Use only the Freebase page!



# Size

| Dataset | # examples | # word types |
|---|---|---|
| GeoQuery | 880 | 279 |
| ATIS | 5,418 | 936 |
| FREE917 | 917 | 2,036 |
| WEBQUESTIONS | 5,810 | 4,525 |

3k 2k

Table 3: Statistics on various semantic parsing datasets. Our new dataset, WEBQUESTIONS, is much larger than FREE917 and much more lexically diverse than ATIS.

# Characterization

- Simple questions + count
  - complexity
- Includes qualifiers (CVT)
- Examples:
  - *"What music did Beethoven compose?"*
  - *"What is James Madison most famous for?"*
  - *"What movies does Taylor Lautner play in?"*
  - *"What kind of system of government does the United States have?"*
  - *"What number is Kevin Youkilis on the Boston Red Sox?"*

how - many

can't spend for production

n-ary facts

code results

SEMPRE

Q+A

+

Freebase

date

KG subset

R

FB ≠ DBP

FB Jun 06 ≠

FB Jun 10

# Conclusions

- Benchmarks drive progress in the community
- Notable QA benchmarks: SQuAD, WebQuestions, Hotpot QA, CC-QuAD
- Ideal benchmarks
  - Large
  - Real = Mimic real
  - Diverse
  - Visionary
- Be sensitive to what the community needs now