

The Quick Guide to SQuAD

All the basic information you need to know about the Stanford Question Answering Dataset (SQuAD).



Jerry Wei · Follow

Published in Towards Data Science · 4 min read · Oct 8, 2020



21



1



The Stanford Question Answering Dataset (SQuAD) is a set of question and answer pairs that present a strong challenge for NLP models.

Whether you're just interested in learning about a popular NLP dataset or planning to use it in one of your projects, here are all the basics you should know.



Photo by [Emily Morter](#) on [Unsplash](#)

What task does SQuAD present? As implied by its name, SQuAD focuses on the task of question answering. It tests a model's ability to read a passage of text and then answer questions about it (*flashback to reading comprehension on the SAT*). It's a relatively straightforward task; here's an example that the dataset's creators gave:

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

Sample question and answer pairs for a passage from the SQuAD dataset. Image credits to Rajpurkar et al., the original creators of the dataset.

How was SQuAD created? To compile SQuAD, the creators sampled 536 from the top 10,000 Wikipedia articles. From each of these sampled articles, they extracted a total of 23,215 individual paragraphs (making sure to filter for paragraphs that were too small). They split the dataset by articles such that 80% of articles went into the training set, 10% into a development set, and 10% into a testing set.

Annotating SQuAD. The most important part of creating a dataset — annotating it — was done by Mechanical Turk workers. *Classic! I’m seeing Mechanical Turk making a cameo in a lot of these NLP papers.* These workers

were selected only if they had a history of high quality work (as measured by the HIT approval rate). For each selected paragraph, the workers were asked to come up with and answer 5 questions on the content of the paragraph. They were provided a text field to type their question, and they could highlight the answers in the paragraph. The creators of SQuAD made sure that the questions that the workers came up with were in their own words, even disabling the copy-paste functionality. *Noooooo! Not my copy-paste tools!*

Paragraph 1 of 43

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 O₂.

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

Example of the procedure used to collect annotations from the Mechanical Turk workers. Image credits to Rajpurkar et al., the original creators of the dataset.

Dataset Analysis. A crucial part of a good dataset is understanding its properties. To do this, the creators explored three areas:

1. **Categories of answers.** Each answer was partitioned into one of the following categories: “date”, “other numeric”, “person”, “location”, “other entity”, “common noun phrase”, “adjective phrase”, “verb phrase”, “clause”, and “other”. They found that dates and numbers made up 19.8% of the answers, nouns made up 32.6%, noun-phrases made up 31.8%, and other categories made up the remaining 15.8%.
2. **Reasoning required.** The creators sampled questions from the development set and manually labeled questions into different categories of reasoning required to answer them. For example, the category “syntactic variation” means that the question is essentially paraphrased, requiring a rearrangement of words to find the answer. Below these bullet points, I’ve included the original table with all of the categories and the percentage of questions that fall into that category.
3. **Syntactic divergence.** The creators measured the syntactic divergence between a question and the sentence containing the answer in order to measure the difficulty of a question. Basically they created a metric that evaluates the number of edits needed to transform a question into the sentence with the answer. The specifics can be found in the original paper [here](#), but what’s important is that the dataset was found to have a diverse range of syntactic divergence. Always good to make sure the dataset is diverse!

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 <u>practical Carnot cycle</u> .	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%

Examples from the development set for each category of reasoning required to answer a question. Image credits to Rajpurkar et al., the original creators of the dataset.



Search Medium

Write



what makes SQuAD so good: Of course, with a task like question answering, there are tons of datasets out there. When comparing SQuAD with other datasets, there are a few primary differences:

- **SQuAD is big.** Other reading comprehension datasets such as MCTest and Deep Read are too small to support intensive and complex models. MCTest only has a total of 2,640 questions, and Deep Read only has a total of 600 questions. SQuAD has these datasets dominated with a whopping 100,000+ questions.
- **SQuAD is challenging.** In other document-based question answering datasets that focus on answer extraction, the answer to a given question occurs in multiple documents. In SQuAD, however, the model only has access to a single passage, presenting a much more difficult task since it isn't as forgiving to miss the answer.

- **SQuAD requires reasoning.** A popular type of dataset is the cloze dataset, which asks a model to predict a missing word in a passage. These datasets are large, and they present a somewhat-similar task as SQuAD. The key improvement that SQuAD makes on this aspect is that its answers are more complex and thus require more-intensive reasoning, thus making SQuAD better for evaluating model understanding and capabilities.

Concluding thoughts. SQuAD is probably one of the most popular question answering datasets (it's been cited over 2,000 times) because it's well-created and improves on many aspects that other datasets fail to address. I'd highly recommend anyone that wants to evaluate an NLP model to test it on SQuAD, as it's a great dataset for testing model understanding of language and even just performance in general.

Further reading:

- [Original paper for SQuAD](#)
- [Website for SQuAD \(including download links\)](#)
- [A list of other question answering datasets in case I haven't convinced you that SQuAD is some good stuff](#)

Artificial Intelligence

NLP

Machine Learning

Data Science

Data



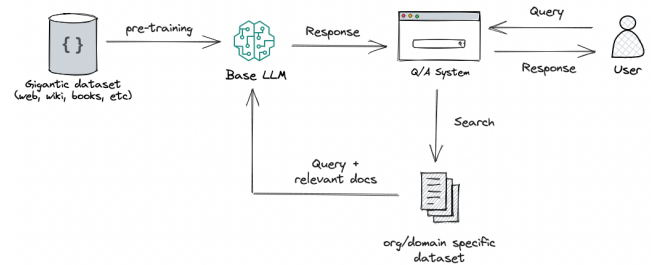
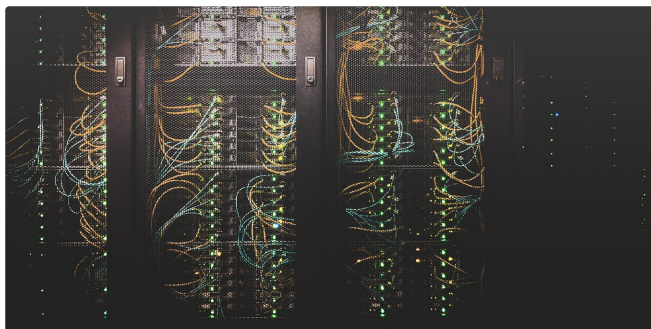
Written by Jerry Wei

Follow

345 Followers · Writer for Towards Data Science

Large language models, AI for health. Research Engineer at Google DeepMind.

More from Jerry Wei and Towards Data Science



Jerry Wei in Towards Data Science

VGG Neural Networks: The Next Step After AlexNet

AlexNet came out in 2012 and improved on traditional Convolutional Neural Networks...

4 min read · Jul 3, 2019



398



1



...



Heiko Hotz in Towards Data Science

RAG vs Finetuning—Which Is the Best Tool to Boost Your LLM...

The definitive guide for choosing the right method for your use case



19 min read · Aug 24



2.1K



16



...




 Jerry Wei in Towards Data Science

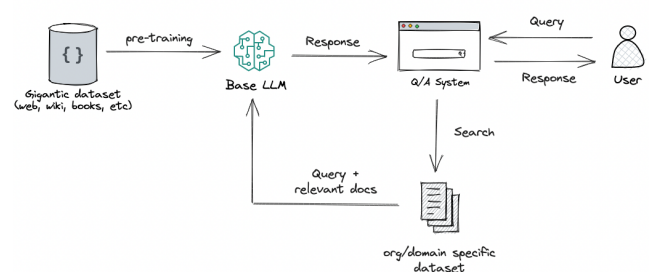
A Quick Guide to Using Command Line (Terminal)

Being able to use command line is necessary for high-level coding. Here are the essentials...

6 min read · Jul 10, 2019



See all from Towards Data Science





Sheldon L.

Evaluation on LLMs

The advent of large language models (LLMs) such as ChatGPT and others, has brought...

🌟 · 6 min read · Jun 3



21



Heiko Hotz in Towards Data Science

RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM...

The definitive guide for choosing the right method for your use case

🌟 · 19 min read · Aug 24



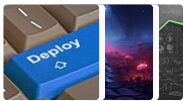
2.1K



16

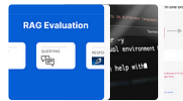


Lists



Predictive Modeling w/ Python

20 stories · 428 saves



Natural Language Processing

652 stories · 257 saves



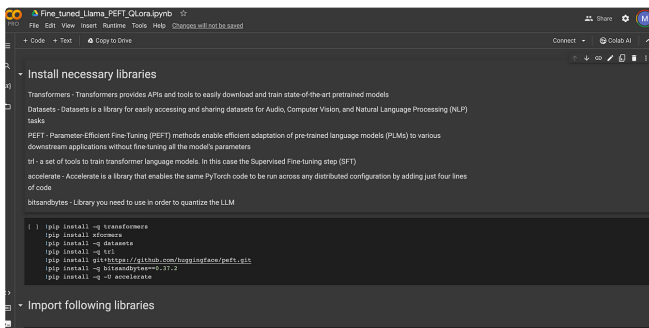
Practical Guides to Machine Learning

10 stories · 492 saves



ChatGPT prompts

24 stories · 433 saves



Maya Akim

Complete Guide to LLM Fine Tuning for Beginners

Fine-tuning a model refers to the process of adapting a pre-trained, foundational model...



M. Baddar in BetaFlow

Article # 1 : Question-Answering Over Documents via LLM - An...

Docs-QA overview




5 min read · Aug 13

👏 93 💬 1

🔖+ ⋮

es:

Feature Based Finetuning	PEFT
1. By updating only o/p layer	1.LORA
2. By Updating all layer	2.Adapters
	3.Prefix Tuning
	4. Prompt Tuning

 Ansuman Das

Finetuning Large Language Models

A Large Language Model is an advanced artificial intelligence (AI) system designed to...

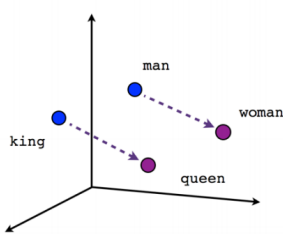
4 min read · Jul 19

👏 💬 🔖+ ⋮

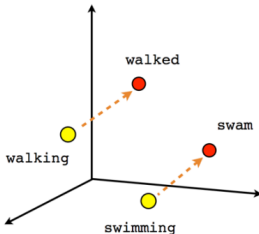
5 min read · Aug 3

👏 3 💬


🔖+ ⋮



Male-Female



Verb tense

 Maninder Singh

Accelerate Your Text Data Analysis with Custom BERT Word...

One thing is for sure the way humans interact with each other naturally is one of the most...

4 min read · Apr 23

👏 155 💬 🔖+ ⋮

See more recommendations