CoQA

A Conversational Question Answering Challenge

What is CoQA?

CoQA is a large-scale dataset for building **Co**nversational **Q**uestion **A**nswering systems. The goal of the CoQA challenge is to measure the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation.

CoQA is pronounced as **coca** (https://en.wikipedia.org/wiki/Coca).

CoQA paper (http://arxiv.org/abs/1808.07042)

CoQA contains 127,000+ questions with answers collected from 8000+ conversations. Each conversation is collected by pairing two crowdworkers to chat about a passage in the form of questions and answers. The unique features of CoQA include 1) the questions are conversational; 2) the answers can be free-form text; 3) each answer also comes with an evidence subsequence highlighted in the passage; and 4) the passages are collected from seven diverse domains. CoQA has a lot of challenging phenomena not present in existing reading comprehension datasets, e.g., coreference and pragmatic reasoning.

Download

Browse the examples in CoQA:

Browse CoQA (https://drive.google.com/open?id=1ik0d_nIsGdXLn8o7tYiiDWN6PK2XNy-D)

Download a copy of the dataset in json format:

Download Training Set (47 MB) (https://nlp.stanford.edu/data/coqa/coqa-train-

v1.0.json)

Download Dev Set (9 MB)
(https://nlp.stanford.edu/data/coqa/coqa-dev-v1.0.json)

Evaluation

To evaluate your models, use the official evaluation script. To run the evaluation, use `python evaluate-v1.0.py --data-file <path_to_dev-v1.0.json> --pred-file <path_to_predictions>`.

Evaluation Script (https://nlp.stanford.edu/data/coqa/evaluate-v1.0.py)

Sample Prediction File (on Dev Set)
(https://nlp.stanford.edu/data/coqa/drqa-pgnet-coqa-dev-hist1.txt.json)

FAQ (https://groups.google.com/forum/#!forum/coqa)

Once you are satisfied with your model performance on the dev set, you submit it to get the official scores on the test sets. We have two test sets, an in-domain set which constitutes the domains present in the training and the dev sets, and an out-of-domain set which constitutes unseen domains (see the paper for more details). To preserve the integrity of the test results, we do not release the test set to the public. Follow this tutorial on how to submit your model for an official evaluation:

Submission Tutorial (https://github.com/stanfordnlp/coqa-baselines/blob/master/codalab.md)

License

CoQA contains passages from seven domains. We make five of these public under the following licenses:

- Literature and Wikipedia passages are shared under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/)license.
- Children's stories are collected from MCTest (https://www.microsoft.com/en-us/research/publication/mctest-challenge-dataset-open-domain-machine-comprehension-text/)which comes with MSR-LA (https://github.com/mcobzarenco/mctest/blob/master/data/MCTest/LICENSE.pdf)license.
- Middle/High school exam passages are collected from RACE (https://arxiv.org/abs/1704.04683)which comes with its own (http://www.cs.cmu.edu/~glai1/data/race/)license.

- News passages are collected from the **DeepMind CNN dataset (https://arxiv.org/abs/1506.03340)**which comes with **Apache (https://github.com/deepmind/rc-data/blob/master/LICENSE)**license.

Questions?

Ask us questions at our **google group (https://groups.google.com/forum/#!forum/coqa)**or at **sivar@cs.stanford.edu (mailto:sivar@cs.stanford.edu)**or **danqi@cs.stanford.edu (mailto:danqi@cs.stanford.edu)**.

Acknowledgements

We thank the **SQuAD team (https://rajpurkar.github.io/SQuAD-explorer/)**for allowing us to use their code and templates for generating this website.

Tweet

Leaderboard

Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* **(Reddy & Chen et al. '18)** **(http://arxiv.org/abs/1808.07042)**	89.4	87.4	88.8
1 Mar 29, 2019	Google SQuAD 2.0 + MMFT (ensemble) *MSRA + SDRG*	**89.9**	**88.0**	**89.4**
2 Mar 29, 2019	ConvBERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	88.7	85.4	87.8
2 Mar 29, 2019	Google SQuAD 2.0 + MMFT (single model) *MSRA + SDRG*	88.5	86.0	87.8
3 Mar 28, 2019	ConvBERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	87.7	84.6	86.8

3 Jan 25, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	87.5	85.3	86.8
4 Jan 21, 2019	BERT + MMFT + ADA (single model) *Microsoft Research Asia*	86.4	81.9	85.0
5 Jan 03, 2019	BERT + Answer Verification (single model) *Sogou Search AI Group* **https://github.com/sogou/SMRCToolkit (https://github.com/sogou/SMRCToolkit)**	83.8	80.2	82.8
6 Jan 06, 2019	BERT with History Augmented Query (single model) *Fudan University NLP Lab*	82.7	78.6	81.5
7 Jan 31, 2019	BERT Large Finetuned Baseline (single model) *Anonymous*	82.6	78.4	81.4
8 Jan 21, 2019	BERT Large Augmented (single model) *Microsoft Dynamics 365 AI Research*	82.5	77.6	81.1
9 Dec 12, 2018	D-AoA + BERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	81.4	77.3	80.2
10 Mar 10, 2019	CNet (single model) *Anonymous*	80.9	77.1	79.8
11 Nov 29, 2018	SDNet (ensemble) *Microsoft Speech and Dialogue Research Group* **https://github.com/Microsoft/SDNet (https://github.com/Microsoft/SDNet)**	80.7	75.9	79.3
12 Feb 22, 2019	CQANet (single model) *Nanjing University*	80.2	76.5	79.1
13 May 09, 2019	CANet (single model) *Northwestern Polytechnical University*	80.1	75.7	78.9
14 Apr 13, 2019	BERT w/ 2-context (single model) *NTT Media Intelligence Laboratories*	79.8	75.9	78.7
15 Dec 30, 2018	BERT-base finetune (single model) *Tsinghua University CoAI Lab*	79.8	74.1	78.1

16 Apr 19, 2019	Bert-FlowDelta (single model) *Anonymous*	79.2	74.1	77.7
17 Feb 28, 2019	GraphFlow (single model) *Anonymous*	78.4	74.5	77.3
18 Nov 26, 2018	SDNet (single model) *Microsoft Speech and Dialogue Research Group* **https://github.com/Microsoft/SDNet (https://github.com/Microsoft/SDNet)**	78.0	73.1	76.6
19 Oct 06, 2018	FlowQA (single model) *Allen Institute for Artificial Intelligence* **https://arxiv.org/abs/1810.06683 (https://arxiv.org/abs/1810.06683)**	76.3	71.8	75.0
20 Jan 14, 2019	RNet + PGNet + BERT (single model) *Nanjing University*	74.7	70.0	73.3
21 Jan 31, 2019	XyzNet (single model) *Beijing Normal University*	74.3	68.8	72.7
22 Dec 30, 2018	DrQA + marker features (single model) *Stanford University*	71.6	65.1	69.7
23 Dec 10, 2018	BiDAF++ (single model) *Beijing University of Posts and Telecommunications*	71.1	65.5	69.5
24 Sep 27, 2018	BiDAF++ (single model) *Allen Institute for Artificial Intelligence* **https://arxiv.org/abs/1809.10735 (https://arxiv.org/abs/1809.10735)**	69.4	63.8	67.8
25 Nov 22, 2018	Bert Base Augmented (single model) *Fudan University NLP Lab*	68.4	61.8	66.5
26 Dec 17, 2018	RNet_DotAtt + seq2seq with copy attention (single model) *University of Science and Technology of China*	68.1	62.3	66.4
27 Dec 30, 2018	Simplified BiDAF++ (single model) *Peking University*	68.7	60.5	66.3

28 Aug 21, 2018	DrQA + seq2seq with copy attention (single model) *Stanford University* **https://arxiv.org/abs/1808.07042 (https://arxiv.org/abs/1808.07042)**	67.0	60.4	65.1
29 Aug 21, 2018	Vanilla DrQA (single model) *Stanford University* **https://arxiv.org/abs/1808.07042 (https://arxiv.org/abs/1808.07042)**	54.5	47.9	52.6