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).

(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_nlsGdXLn8o7tYiiDWN6PK2XNy-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://groups.google.com/forum/#!forum/coqa)

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.

Leaderboard

Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* **(Reddy et al. '18)** **(http://arxiv.org/abs/1808.07042)**	89.4	87.4	88.8
1 Dec 12, 2018	D-AoA + BERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	**81.4**	**77.3**	**80.2**
2 Nov 29, 2018	SDNet (ensemble model) *Microsoft Speech and Dialogue Research Group* **https://arxiv.org/abs/1812.03593** **(https://arxiv.org/abs/1812.03593)**	80.7	75.9	79.3
3 Nov 26, 2018	SDNet (single model) *Microsoft Speech and Dialogue Research Group* **https://arxiv.org/abs/1812.03593** **(https://arxiv.org/abs/1812.03593)**	78.0	73.1	76.6
4 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
5 Dec 10, 2018	BiDAF++ (single model) *Beijing University of Posts and Telecommunications*	71.1	65.5	69.5
6	BiDAF++ (single model)	69.4	63.8	67.8

Sep 27, 2018	*Allen Institute for Artificial Intelligence* **https://arxiv.org/abs/1809.10735 (https://arxiv.org/abs/1809.10735)**			
7 Nov 22, 2018	Bert Base Augmented (single model) *Fudan University NLP Lab*	68.4	61.8	66.5
8 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
9 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