A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC

Mark Yatskar

Allen Institute for Artificial Intelligence

marky@allenai.org

Abstract

We compare three new datasets for question answering: SQuAD 2.0, QuAC, and CoQA, along several of their new features: (1) unanswerable questions, (2) multi-turn interactions, and (3) abstractive answers. We show that the datasets provide complementary coverage of the first two aspects, but weak coverage of the third. Because of the datasets' structural similarity, a single extractive model can be easily adapted to any of the datasets and we show improved baseline results on both SQuAD 2.0 and CoQA. Despite the similarity, models trained on one dataset are ineffective on another dataset, but we find moderate performance improvement through pretraining. To encourage cross-evaluation, we release code for conversion between datasets at https://github.com/my89/co-squac.

1 Introduction

Question answering on textual data has served as a challenge problem for the NLP community (Voorhees, 2001; Richardson et al., 2013). With the development of large scale benchmarks and sufficiently simple evaluations (Trischler et al., 2016; Nguyen et al., 2016; Hermann et al., 2015) progress has been rapid. In recent evaluation on SQuAD (Rajpurkar et al., 2016), performance exceeded that of annotators (Wang et al., 2018; Hu et al., 2017; Wang et al., 2017).

In response to this development, there have been a flurry of new datasets. In this work, we analyze three such new proposed datasets, SQuAD 2.0 (Rajpurkar et al., 2018), QuAC (Choi et al., 2018), and CoQA (Reddy et al., 2018).[1] In each of these datasets, crowd workers are asked to (1) produce questions about a paragraph of text (context) and (2) produce a reply by either indicating there is no answer, or providing an extractive answer from the context by highlighting one contiguous span. QuAC and CoQA contain two other features: questions are asked in the form of a dialog, where co-reference to previous interactions is possible and directly answering yes/no is possible. CoQA also allows workers to edit the spans to provide abstractive answers.[2]

We compare these three datasets along several of their new features: (1) unanswerable questions, (2) multi-turn interactions, and (3) abstractive answers. Unanswerable question coverage is complementary among datasets; SQuAD 2.0 focuses more on questions of extreme confusion, such as false premise questions, while QuAC primarily focuses on missing information. QuAC and CoQA dialogs simulate different types of user behavior: QuAC dialogs often switch topics while CoQA dialogs include more queries for details. Unfortunately, no dataset provides significant coverage of abstractive answers beyond yes/no answers, and we show that a method can achieve an extractive answer upper bound of 100 and 97.8 F1 on QuAC and CoQA , respectively.

Motivated by the above analysis, we apply the baseline presented in QuAC (Choi et al., 2018), BiDAF++, a model based on BiDAF (Seo et al., 2016), augmented with self attention (Clark and Gardner, 2018) and ELMo contextualized embeddings (Peters et al., 2018) to all datasets. Experiments show that this extractive baseline outperforms existing extractive and abstractive baselines on CoQA by 14.2 and 2.7 F1 respectively. Finally, we show models can transfer between datasets with pretraining yielding moderate gains.[3]

[1] A review of other new datasets is in the related work.

[2] Also, SQuAD 2.0 and QuAC cover only Wikipedia text, CoQA covers six other domains and QuAC is the only one of these datasets that doesn't allow the questioner to see the context before formulating a question.

[3] To facilitate easy future cross-evaluation, we release tools for conversion between these dataset.

Proceedings of NAACL-HLT 2019, pages 2318–2323
Minneapolis, Minnesota, June 2 - June 7, 2019.

Dataset	Entity Salad	False Premise	Topic Error	Missing Information	Content Negation	Answerable Questions	Total Questions
CoQA	0.0	0.0	0.0	60.0	0.0	40.0	5 (0.5%)
SQuAD 2.0	21.3	21.3	13.5	16.1	16.1	10.9	230 (50.1%)
QuAC	5.5	0.0	16.4	71.2	0.0	6.8	73 (20.2%)

Table 1: Comparison of unanswerable questions on 50 random contexts from the development set of each dataset. SQuAD 2.0 contains a diverse set of circumstances that make questions unanswerable, QuAC focuses on information that could plausibly be in context material and CoQA does not significantly cover unanswerable questions.

Dataset	Topic Shift	Drill Down	Return to Topic	Clarification Question	Definition Question	Sentence Coverage	Total Questions
CoQA	21.6	72.0	2.9	0.0	0.7	63.3	722
QuAC	35.4	55.3	5.6	0.7	3.0	28.4	302

Table 2: Comparison of dialog features in 50 random contexts from the development set of each dataset. CoQA contains questions that drill into details about topics and cover 60% of sentences in the context while in QuAC dialog switch topic more often and cover less than 30% of sentences. Neither dataset has a significant number of returns to previous topics, clarifications, or definitional interactions.

2 Dataset Analysis

In this section we analyze unanswerable questions, dialog features, abstractive answers in SQuAD 2.0, QuAC, and CoQA. All analysis was performed by the authors, on a random sample of 50 contexts (300-700 questions) from the development set of each dataset.

2.1 Unanswerable Questions

In Table 1 we compare types of unanswerable questions across dataset. We identify five types of questions found between the datasets:

1. Entity Salad A nonsensical reference to entities found in the context or made-up entities (e.g. *"What infinite hierarchy implies that the graph isomorphism problem s NQ-complete?"*). Such questions are unanswerable for any context.

2. False Premise A fact that contradicts the context is asserted in the question (e.g. *"When is the correlation positive?"* but in the context says *"the correlation is strictly negative"*).

3. Topic Error A questions that references an entity in the context but the context does not focus on that entity (e.g *"How many earthquakes occur in California?"* when the article focus is actually about *"Southern California"*). Such questions potentially have answers, but it would be unlikely for the answer to be found in the context.

4. Missing Information A question who's answer could be plausibly in the context but is not (e.g. *"What is the record high in January?"* and the article is about temperature extremes). Such questions have an answer but it is not mentioned.

5. Content Negation A question which asks for the opposite information of something mentioned in the context (e.g. *"Who* **didnt** *cause the dissolution of the Holy Roman Empire?"*). Such questions either have answers that are the set of all entities other than the one mentioned or answers that could be found in some other context.

Results SQuAD 2.0 contains the highest diversity of unanswerable questions of all datasets analyzed. Some SQuAD 2.0 questions are unlikely to be asked without significant foreknowledge of the context material and do not occur in QuAC. [4] Both SQuAD 2.0 and QuAC cover a significant number of unanswerable questions that could be plausibly in the article. The difference in settings and distributions of unanswerable questions in SQuAD 2.0 and QuAC appear to be complementary: SQuAD 2.0 focuses more on questions simulating questioner confusion, while QuAC primarily focuses on missing information. [5]

2.2 Dialog Features

In Table 2 we analyze five dialog behaviors:

1. Topic Shift A question about something previously discussed (e.g. "Q: How does he try to take over? ... Q: Where do they live?").

2. Drill Down A request for more information about a topic being discussed (e.g. "A: The Sherpas call Mount Everest Chomolungma. Q: Is Mt. Everest a holy site for them?")

3. Topic Return Asking about a topic again after it had previously been shifted away from.

[4] Such questions resemble text from entailment datasets such as SNLI (Bowman et al., 2015) and seem more likely to arise if questioners are receiving very complex information and become confused.

[5] CoQA does not contain a significant number of unanswerable questions, and many of the ones that do exist are erroneously marked.

Dataset	Yes/No	Coref	Counting	Picking	Fluency	Max F1
CoQA	21.4	3.2	1.3	0.6	4.2	97.8
QuAC	21.1	0.0	0.0	0.0	0.0	100.0

Table 3: Comparison of abstractive features in 50 random contexts in the develoment set of each dataset. Both QuAC and CoQA contain yes/no questions while CoQA also contains answers that improve fluency through abstractive behavior. The extractive upper bound from CoQA is high because most absractivive answers involve adding a pronoun (Coref) or inserting prepositions and changing word forms (Fluency) to existing extractive answers, resulting in extremely high overlap with possible extractive answers.

4. Clarification Reformulating a question that had previously been asked.

5. Definition Asking what is meant by a term (e.g. "What are polygenes?")

Results QuAC and CoQA contain many similar features but at very different rates, offering complementary coverage of types of user behavior. CoQA dialogs drill down for details significantly more frequently and cover more than 60% of sentences in the context material (Sentence Coverage). QuAC dialogs shift to new topics frequently and cover less than 30% of sentences in the context. Both datasets contain only a small numbers of definition questions and returns to previous topics and few requests for clarification.

2.3 Abstractive Answers

Table 3 compares abstractive behavior in CoQA and QuAC. We observed five phenomena:

1. Yes/No Questions annotated with yes/no. In QuAC such questions and their corresponding yes or no are marked in addition to an extractive answer. In CoQA, the single token "yes" or "no" is simply asserted as the abstractive answer, with an extractive answer provided in the rationale (e.g. "Q: Is atmosphere one of them? A: yes").

2. Coref Coreference is added to previously mentioned entities in either context or question (e.g. "Q: How was France's economy in the late 2000s? A: **it** entered the recession").

3. Count Counting how many entities of some type were mentioned (e.g. "Q: how many specific genetic traits are named? A: five")

4. Picking A question that requires the answer to pick from a set defined in the question (e.g. "Q: Is this a boy or a girl? A: boy)

5. Fluency Adding a preposition, changing the form of a word, or merging two non-contiguous spans (e.g. "Q: how did he get away? A: **by** foot)

Results Both QuAC and CoQA have a similar rate of yes/no questions. QuAC contains no other abstractive phenomena while CoQA contains a small number of predominately insertions, often at the beginning of an extractive span, for coreference and or other fluency improvements. Because abstractive behavior in CoQA includes mostly small modifications to spans in the context, the maximum achievable performance by a model that predicts spans from the context is 97.8 F1. [6]

	Overall F1
DrQA (Extractive)	54.7
DrQA + PGNet (Abstractive)	66.2
BiDAF++ w/ 0-ctx	63.4
BiDAF++ w/ 3-ctx	**69.2**

Table 4: Development set performance by training BiDAF++ (Choi et al., 2018) models (extractive) on CoQA data with handling yes/no and no-answer questions as in QuAC. Despite being extractive, these models significantly outperform reported baselines, DrQA and DrQA + PGNet (Reddy et al., 2018).

	in-F1	out-F1	F1
DrQA	54.5	47.9	52.6
DrQA + PGNet	67.0	60.4	65.1
BiDAF++ w/ 3-ctx	**69.4**	**63.8**	**67.8**

Table 5: Test set results on CoQA. We report in domain F1 (in-F1), out of domain F1 on two held out domains, Reddit and Science (out-F1) and the overall F1 (F1).

3 New Extractive Baseline for CoQA

Our analysis strongly implies that beyond yes/no questions, abstractive behavior is not a significant component in either QuAC or CoQA. As such, QuAC models can be trivially adapted to CoQA.

We train a set of BiDAF++ baselines from the original QuAC dataset release (Choi et al., 2018) by optimizing the model to predict the span with maximum F1 overlap with respect to annotated abstractive answers.[7] If the abstractive answer is ex-

[6] To compute the upper bound, if abstractive answer is exactly "yes", "no", or "unknown", we consider the upper bound to be 100. Otherwise, we use the CoQA evaluation script to find a span in the context that has maximum F1 with respect to the abstractive answer.

[7] We use the implementation on http://allennlp.org, and do not modify any hyper-parameters except the the maximum dialog length and that models were allowed to train up to 65 epochs.

	F1	HEQQ	HEQD
BiDAF++ w/ 2-ctx	60.6	55.7	4.0
Train SQuAD 2.0	34.3	18.0	0.3
Train CoQA	31.2	19.2	0.0
Ft from SQuAD 2.0	62.6	58.3	**5.9**
Ft from CoQA	**63.3**	**59.2**	5.1

Table 6: Cross dataset transfer to QuAC development set. Models do not transfer directly (rows 3 and 4), but after fine tuning improve performance (rows 5 and 6).

actly "yes" or "no", we train the model to output the whole rationale span, and classify the question as yes/no with the appropriate answer. At evaluation time, if the model predicts a question is a yes/no question, instead of returning the extracted span, we simply return "yes" or "no".

Results Table 4 and Table 5 summarize our results for training BiDAF++ with varying contexts on CoQA. Beyond the difference of underlying base question-answer models (DrQA (Chen et al., 2017) vs. BiDAF (Seo et al., 2016) with self attention (Clark and Gardner, 2018)), BiDAF++ has two core differences with respect to DRQA+PGNet: (1) instead of appending previous questions and answers to input question tokens, BiDAF++ marks answers of previous questions directly on the context, and (2) BiDAF++ uses contextualized word embeddings through ELMo (Peters et al., 2018). These differences, in combination with appropriate handling of yes/no and unanswerable questions significantly improves on the existing extractive baseline **(+14.2 F1)** and even on the existing abstractive baseline **(+2.7 F1)**.

4 Cross-Dataset Experiments

In this section we consider whether models can benefit from transfer between SQuAD 2.0, QuAC, and CoQA, and show that the datasets, while ineffective for direct transfer, can be used as pre-training. In all experiments, we use BiDAF++, either with two context or no context, depending on if we are training for dialog settings or not, with default configurations. Models are trained by initializing from other models trained on different datasets and we do not decrease initial learning rates from just training directly on the target dataset. When SQuAD 2.0 is used to initialize models that use context, we randomly order questions in SQuAD 2.0 and train as if questions were asked in the form of a dialog. [8]

[8]Likely a better strategy exists but we would like to demonstrate transfer in the simplest way. We only report development numbers as these experiments are meant to be exploratory.

	In Domain F1
DrQA + PGNet	66.2
BiDAF++ w/ 2-ctx	67.6
SQuAD 2.0	41.4
QuAC	29.1
Ft from SQuAD 2.0	**69.2**
Ft from QuAC	68.0

Table 7: Cross dataset transfer to CoQA development set. Models do not transfer directly (rows 3 and 4), but after fine tuning improve performance (rows 5 and 6). For an explanation of why BiDAF++ outperforms DrQA + PGNet, see Section 3.

	F1	EM
Baseline	67.6	65.1
BiDAF++	70.5	67.4
CoQA	38.1	32.4
QuAC	25.4	16.8
Ft from CoQA	**72.5**	**69.4**
Ft from QuAC	69.5	66.8

Table 8: Cross dataset transfer to SQuAD 2.0 development set. BiDAF++ (Choi et al., 2018) outperforms the baseline, a different implementation of the same model (Rajpurkar et al., 2018) likely because of better hyper parameter tuning.

Results Tables 6-8 summarize our results. Across all of the datasets, BiDAF++ outperforms other baselines, and there exists at least one other dataset that significantly improves performance on a target dataset on average **+2.1 F1**. Experiments do not support that direct transfer is possible.

5 Related Work

Other proposals exist other than the three we analyzed that expand on features in SQuAD (Rajpurkar et al., 2016). For example, maintaining question independence of context to reduce the role of string matching and having long context length (Joshi et al., 2017; Kociský et al., 2017), higher level reasoning (Khashabi et al., 2018; Clark et al., 2018; Yang et al., 2018), multi-turn information seeking interactions, in either table settings (Iyyer et al., 2017; Talmor and Berant, 2018; Saha et al., 2018), regulation settings (Saeidi et al., 2018), or Quiz Bowl settings (Elgohary et al., 2018). Other work considers multi-modal contexts where interactions are a single turn (Tapaswi et al., 2016; Antol et al., 2015; Lei et al., 2018) or multi-turn (Das et al., 2017; Pasunuru and Bansal, 2018). These efforts contain alternative challenges than ones we analyze in this paper.

Acknowledgement

We thank Eunsol Choi, Hsin-Yuan Huang, Mohit Iyyer, He He, Yejin Choi, Percy Liang, and Luke Zettlemoyer for their helpful discussions in formulating this work. Also, Siva Reddy and Danqi Chen for help evaluating on CoQA and all reviewers for their comments.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Empirical Methods in Natural Language Processing*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. *Proceedings of the Association for Computational Linguistics*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Computer Vision and Pattern Recognition*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2017. Reinforced mnemonic reader for machine reading comprehension. *arXiv preprint arXiv:1705.02798*.

Mohit Iyyer, Wen tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the Association for Computational Linguistics*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, abs/1712.07040.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv*, abs/1611.09268.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. In *EMNLP*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Conference of the North American Chapter of the Association for Computational Linguistics*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the Association for Computational Linguistics*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *ArXiv*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Signh, Tim Rocktschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of Empirical Methods in Natural Language Processing*.

Amrita Saha, Vardaan Pahuja, Mitesh M Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Association for the Advancement of Artificial Intelligence*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *Proceedings of the International Conference on Learning Representations*.

A. Talmor and J. Berant. 2018. The web as knowledgebase for answering complex questions. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Ellen M Voorhees. 2001. The trec question answering track. *Natural Language Engineering*, 7(4):361–378.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1705–1714.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.