TOWARDS YET ANOTHER CHECKLIST FOR NEW DATASETS

Stefan Larson
Vanderbilt University
stefan.larson@vanderbilt.edu

ABSTRACT

The goal of this opinion paper is to start (or more accurately, continue) a conversation on a *checklist*-style artifact for researchers to use when introducing a new dataset. While other checklists exist and are widely used, this paper proposes the inclusion of checklist-style questions to encourage dataset developers (as well as consumers) to consider data quality, diversity, and evaluation, especially as it relates to machine learning model robustness and generalizability.

1 INTRODUCTION

The development of new datasets is a critical enabler of progress in machine learning research. New datasets can introduce new challenges, allowing researchers to expand the abilities of machine learning systems. New datasets can also scale up the amount of data available to existing tasks, enabling the development and application of deeper models (consider for instance the role of the ImageNet dataset Deng et al. (2009) in the rise of convolutional neural network architectures in the 2010s).

This paradigm is not without its problems, however, as machine learning datasets often contain unwanted patterns which lead to model biases. Moreover, the use of datasets as standardized benchmarks is closely entwined with the notion of "leaderboard chasing", whereby researchers publish incremental improvements on a select few datasets for a given task. Leaderboard chasing may lead to tunnel vision, whereby a researcher (or an entire research community) may focus on a particular task or setting that at best is a necessarily simplified version of reality and at worst a counterproductive, abstracted version of real life phenomena. Indeed, the term *incremental* is often seen as a pejorative especially given the recent breadth of findings indicating that machine learning models trained on standard benchmark datasets do not generalize well to new input distributions (e.g., Yatskar (2019); Harrigian et al. (2020); Lazaridou et al. (2021)) or do not perform well on minor alterations to in-distribution data (e.g., Glockner et al. (2018); Iyyer et al. (2018); Larson et al. (2020)).

Despite this, the machine learning community is laudable for being introspective and observing that the leaderboard chasing nature of research has downsides. For instance, Sculley et al. (2018) observed that rapid advancements in machine learning often came at the cost of lack of careful rigor. Lack of careful rigor also goes hand-in-hand with lack of reproducibility, and to combat this the machine learning research community has introduced guidelines and checklists aimed at facilitating rigorous and reproducible research. Examples include the 2021 NeurIPS Paper Checklist Guidelines, the NLP Reproducibility Checklist Dodge et al. (2019) now used by the *ACL conferences, the Checklist for Responsible Data Use in NLP Rogers et al. (2021), the 2021 AAAI Reproducibility Checklist, and the ARR Responsible NLP Research Checklist. These checklists do focus on datasets, either partially or exclusively (in the case of Rogers et al. (2021)), yet this paper argues that perhaps it is time to append new checklist-style questions to these lists:

My primary argument is that a checklist could be used to encourage the creation and use of more diverse datasets, those that could facilitate more evaluations into the important problems of out-of-distribution robustness and generalizability, and are free of spurious annotation artifacts, to the extent to which removal of such unwanted artefacts is reasonable. Toward this end, I discuss 9 checklist items that I argue would aid dataset creators in developing, analyzing, and presenting new datasets.

This paper follows a line of introspective recent work on improving machine learning research methodology like Linzen (2020) and Bowman & Dahl (2021), among many others. And if it has not already been made evident by the template used for this paper, this author is mainly familiar with NLP research, but the ideas in this paper extend to other fields such as computer vision, too.

2 IDEAS FOR A CHECKLIST

The rest of this paper presents ideas for questions that could be included in reproducibility or responsibility checklists. In general, the following questions containing more discussion tend to be more unorthodox, for lack of a better word.

2.1 TRAIN-TEST SPLITS

Does the dataset have a provided train/dev/test split? If so, how was the data split across the different folds? Does the dataset have a provided "challenge" split?

Datasets are often introduced with a provided train/dev/test split. Provided splits are helpful to researchers because they standardize the train, development, and evaluation data, and thereby facilitate comparison across models and researchers. (We note that often datasets are released with just train and test splits, and leave any development spltting up to the consumer of the dataset.) Typical approaches to splitting data involve randomly partitioning the data over the various splits or folds (e.g., in the case of k-fold cross validation). This approach is taught in introductory machine learning and data science courses and textbooks.

Recently, though, machine learning researchers have questioned the effectiveness of randomized data splits in titles such as *We need to talk about standard splits* Gorman & Bedrick (2019), *We need to talk about random splits* Søgaard et al. (2021), and *We need to talk about train-dev-test splits* van der Goot (2021). The primary issue raised is that randomized splits may lead to overstated or inflated evaluation scores due to data "twinning" artifacts whereby similar or near-duplicate data appears in both training and testing splits.

While there may be circumstances where randomized data splitting is desired, there certainly are cases where it is not appropriate. For instance, in prediction tasks that have a temporal element, it may be more appropriate to split based on some chronological cutoff date. Indeed, Karimi et al. (2015) notes that data may exhibit statistical dependence with time, for instance in social media data, where events being discussed or language use (in the case of NLP) may be dependent on certain time horizons.

It can be useful to use other-than-random splits even for data that does not have a temporal element. For instance, in a study of eight text-to-SQL datasets, Finegan-Dollak et al. (2018) found that standard splits caused many query forms to appear in both training and testing, hindering the ability to evaluate how well systems can formulate *new* SQL queries from natural language inputs. Finegan-Dollak et al. (2018) then introduced new splits on the text-to-SQL datasets using query form as a basis for splitting.

More recently, Søgaard et al. (2021) intentionally apply "biased" or "adversarial" splitting techniques to textual data, for instance splitting data by length (shorter sentences in one split, longer sentences in the other) or by maximizing the Wasserstein distance between splits, arguing that these relatively straightforward heuristic or adversarial splitting techniques yielded more accurate estimates of the true error rates versus randomized splits. Similarly, Wecker et al. (2020) introduce a novel splitting algorithm that makes use of k-means clustering to partition a dataset into folds.

Researchers constructing new datasets should be encouraged to consider the questions presented at the start of this sub-section, as should paper reviewers and consumers of datasets. Thinking about possible pitfalls to applying randomized splits to a dataset can also help researchers identify potential biases in the data. Consideration of these types of questions will also promote the development of new tools for investigating datasets (e.g., tools for detecting data twinning, tools for creation of adversarial or challenge splits, etc.).

2.2 ANNOTATION ARTIFACTS

What efforts were made to detect and/or minimize potential annotation artifacts in the dataset?

Annotation artifacts are patterns in data that can lead to the often-unwanted learning of "spurious cues" by a model. In textual data, such spurious cues may include the presence or absence of certain lexical features (e.g., tokens, punctuation, etc.), and sample length (e.g., the number of tokens in an input). In computer vision tasks, spurious cues could include seemingly irrelevant contextual information like the presence or absence of snow on the ground when classifying images of dogs Ribeiro et al. (2016).

Within NLP, artifacts have been studied in Levy et al. (2015); Schwartz et al. (2017); Gururangan et al. (2018); Glockner et al. (2018); Poliak et al. (2018); Tsuchiya (2018); Aharoni & Goldberg (2018); Paun et al. (2018); Niven & Kao (2019); Geva et al. (2019), who all note the prevalence of such artifacts in data generated via crowdsourcing, and finding that high capacity models like BERT have a tendency to overfit to these artifacts.

In addition to running benchmarks for primary metrics like accuracy, dataset developers could also provide initial analyses to see if there are any major annotation artifacts.

2.3 LABOR COMPENSATION

If hired labor was used, were they paid? How much?

Ideally, the answer to this question is at least the affirmative. But further, details shedding insight into the financial compensation to any workers used in the creation and annotation of the dataset is useful. If very little was paid to crowd workers, for instance, it would not be unreasonable to be worried that the quality or diversity of the dataset is less than ideal.

2.4 DATA DIVERSITY

What are the sources of your dataset? Does your dataset include data from diverse sources (e.g., different geographic or cultural "distributions")? Does your dataset consist of a diverse set of features?

The first two of these questions are aimed at prompting discussion or consideration of the need for data that is representative of different peoples, cultures, geographic regions, etc. As a concrete example, a face recognition dataset consisting of only white faces may be problematic in that it does not include enough diversity to enable effective application of models to other skin colors. Developers of datasets like this hypothetical one ought to be prompted to discuss their design decisions.

Notions of data diversity are closely tied to out-of-distribution robustness, generalizability, and annotation artifacts. A text classification dataset may have been created by scraping tweets containing a specific hashtag. It may be the case that twitter users who use that hashtag are from a particular geographic location, and hence dataset designers should be prompted to consider whether such data is diverse enough to endow models with the power to generalize well to other distributions.

The third question is also related to robustness, generalizability, and annotation artefacts, yet is focused more on what *features* appear in the dataset. For example, a text corpus consisting of just 1,000 unique tokens is less lexically diverse than a dataset of 25,000 unique tokens. Researchers who introduce new datasets should try to measure the diversity of their dataset, where appropriate. Moreover, new tools could be developed by researchers to aid in this type of analysis.

2.5 BASELINE BENCHMARKS

What baseline models or solutions were used to establish an initial benchmark for the dataset? What metrics were used in the benchmark experiments?

It is unlikely that there is a correct answer to this first question for the majority of datasets. To be impactful, a new dataset will typically introduce a new task (or subtask) or a new data distribution (e.g., textual data from a low-resource language) for an existing task, and therefore existing models may underperform on the dataset. However, if an "easy" or "simple" baseline achieves high perfor-

mance, then it may indicate that the dataset's task is easily solvable (alternatively, it may indicate the presence of spurious cues or annotation artifacts). Therefore, discussion the motivation behind the inclusion of any baseline models used should be encouraged.

Metrics are necessary for researchers to compare and evaluate models. The choice of an appropriate metric (or metrics) is extremely important, and discussion of the motivation behind the selection of a metric for benchmark performance evaluation should be encouraged.

2.6 Out-of-Distribution, Generalizability, and Robustness

If appropriate, does the dataset include a way to evaluate out-of-distribution performance?

As mentioned in the introduction, the "leaderboard chasing" paradigm in machine learning research often sees researchers laser-focused on optimizing models on a select few datasets. What's more, these datasets often only provide a means for measuring model performance on *in-distribution* data. In contrast, *out-of-distribution* (OOD) performance is extremely important in real-world settings, as production models often must be able to discriminate between inputs that belong to the training label set versus those that do not. Moreover, such models must be able to generalize to data that may belong to the training label set, yet was generated by a different mechanism than the original training data (i.e., distribution shift).

Consider the 30-year-old ATIS, a dataset commonly used for benchmarking intent classification and slot-filling models for task-driven dialog systems. Recent models now achieve accuracy and F1 scores in the high 90s on this dataset, yet Larson et al. (2020) found that model performance dropped substantially on paraphrased inputs where certain tokens were not allowed, indicating that the models were not robust and could not generalize well to new data. Or consider the RVL-CDIP dataset Harley et al. (2015) used in the document analysis research community for benchmarking document classifiers. RVL-CDIP consists of tobacco industry documents from between the 1970s and early-2000s. Current models now report accuracies in the mid-90s on this 16-class dataset, yet in a to-be-published work-in-progress, my colleagues and I found that model performance drops by up to 30 points on a newly-collected test set.

To put it plainly, findings such as these indicate a crisis. The research community can take steps toward overcoming this crisis by encouraging dataset developers to think of ways to create out-of-distribution evaluation sets for measuring (1) model performance on what has been called *out-of-application* Bohus & Rudnicky (2005) data that does not belong to any of a training set's label classes, and/or (2) model performance on covariate or distribution shifted data.

The onus for all of this does not rest only on dataset developers. Indeed, the question at the top of this section can also be re-crafted to researchers developing new models as: *Do the experiments you use to benchmark your new model include an analysis on out-of-distribution performance?*

2.7 Data Preprocessing

What, if any, preprocessing steps were performed on the dataset?

This question is certainly relevant for promoting reproduciblity. Moreover, knowledge of and access to—via code— preprocessing steps can aid developers or consumers of the dataset in spotting potential pitfalls with regards to unwanted annotation artefacts.

2.8 Sensitive Data

Does the dataset contain sensitive (e.g., personally identifiable information (PII)) data? In what ways were sensitive data handled?

This checklist category has been discussed in depth by Rogers et al. (2021), but I include it here because it may potentially impact data preprocessing and annotation artefacts.

Datasets containing personally identifiable information like names, phone numbers, addresses, religious and political affiliation, etc., could be subject to action requests due to laws such as the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR), and often it is best to minimize or eliminate the amount of PII data in a dataset, which could be ac-

complished by anonymizing or scrubbing PII. Dataset designers should therefore be encouraged to discuss how they handle sensitive information in their dataset.

2.9 Crowdsourcing Prompts

If you used crowdsourcing, do you make available the crowdsourcing prompts used in the collection and annotation of the dataset?

Of course, this question is mainly relevant for datasets that used crowdsourcing in their construction.[1] Making crowdsourcing prompts available along with the release of a dataset is related to the reproducability concept in machine learning research, and can enable other researchers (1) use similar prompts, (2) constructively criticize, and/or (3) improve upon in their research. Visibility into the nature of data collection prompts (crowdsourcing or otherwise) also can enable researchers identify the causes of potential biases and annotation artifacts in data.

3 Conclusion

This paper sketches some questions that I argue should be included in research checklists like those for NeurIPS and ARR. The questions outlined are tailored to encouraging the development of diverse datasets for facilitating the development of robust, generalizable machine learning models. Should a checklist be prescriptive and enforce this type of development of new datasets, even if the "enforcement" is soft encouragement? Perhaps, and this is the type of question through which I hope to start a discussion. What should a dataset *be* besides a collection of data?

References

Roee Aharoni and Yoav Goldberg. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 719–724, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2114. URL https://aclanthology.org/P18-2114.

Dan Bohus and Alexander I. Rudnicky. Sorry and I didn't catch that! - an investigation of non-understanding errors and recovery strategies. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pp. 128–143, Lisbon, Portugal, September 2005. Special Interest Group on Discourse and Dialogue (SIGdial). URL https://www.aclweb.org/anthology/2005.sigdial-1.14.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://aclanthology.org/D19-1224.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

[1] Other "prompt-like artifacts like search queries used, or websites scraped are also relevant and may be worth sharing as part of releasing a new dataset.

Long Papers), pp. 351–360, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1033. URL https://aclanthology.org/P18-1033.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL https://aclanthology.org/D19-1107.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 650–655, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2103. URL https://aclanthology.org/P18-2103.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2786–2791, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1267. URL https://aclanthology.org/P19-1267.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2015. URL https://www.cs.cmu.edu/~aharley/icdar15/harley_convnet_icdar15.pdf.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, November 2020. doi: 10.18653/v1/2020.findings-emnlp.337. URL https://aclanthology.org/2020.findings-emnlp.337.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1875–1885, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1170. URL https://aclanthology.org/N18-1170.

Sarvnaz Karimi, Jie Yin, and Jiri Baum. Squibs: Evaluation methods for statistically dependent text. *Computational Linguistics*, 41(3):539–548, September 2015. doi: 10.1162/COLI_a_00230. URL https://aclanthology.org/J15-3006.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K. Kummerfeld. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8097–8106, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.650. URL https://aclanthology.org/2020.emnlp-main.650.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in natural language models. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/file/f5bf0ba0a17ef18f9607774722f5698c-Paper.pdf.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 970–976, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1098. URL https://aclanthology.org/N15-1098.

Tal Linzen. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5210–5217, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.465. URL https://aclanthology.org/2020.acl-main.465.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL https://aclanthology.org/P19-1459.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585, 12 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00040. URL https://doi.org/10.1162/tacl_a_00040.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines for natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. URL https://arxiv.org/pdf/1805.01042.pdf.

Marco Tulio Ribeiro, Sameer Sing, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of KDD 2016*, 2016. URL https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4821–4833, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.414. URL https://aclanthology.org/2021.findings-emnlp.414.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 15–25, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1004. URL https://aclanthology.org/K17-1004.

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/pdf?id=rJWF0Fywf.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1823–1832, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.156. URL https://aclanthology.org/2021.eacl-main.156.

Masatoshi Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1239.

Rob van der Goot. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4485–4494, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.368. URL https://aclanthology.org/2021.emnlp-main.368.

Hanna Wecker, Annemarie Friedrich, and Heike Adel. ClusterDataSplit: Exploring challenging clustering-based data splits for model performance evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 155–163, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.15. URL https://aclanthology.org/2020.eval4nlp-1.15.

Mark Yatskar. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2318–2323, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1241. URL https://aclanthology.org/N19-1241.