A guide to the dataset explosion in QA, NLI, and commonsense reasoning

Anna Rogers
Center for Social Data Science
University of Copenhagen
Copenhagen, Denmark
arogers@sodas.ku.dk

Anna Rumshisky
Dept. of Computer Science
Univ. of Massachusetts Lowell
Lowell, USA
arum@cs.uml.edu

Abstract

Question answering, natural language inference and commonsense reasoning are increasingly popular as general NLP system benchmarks, driving both modeling and dataset work. Only for question answering we already have over 100 datasets, with over 40 published after 2018. However, most new datasets get "solved" soon after publication, and this is largely due not to the verbal reasoning capabilities of our models, but to annotation artifacts and shallow cues in the data that they can exploit.

This tutorial aims to (1) provide an up-to-date guide to the recent datasets, (2) survey the old and new methodological issues with dataset construction, and (3) outline the existing proposals for overcoming them. The target audience is the NLP practitioners who are lost in dozens of the recent datasets, and would like to know what these datasets are actually measuring. Our overview of the problems with the current datasets and the latest tips and tricks in the dataset construction methodology will also be useful to the researchers working on future benchmarks. The tutorial slides are available online at https://www.annargrs.github.io/dataset-explosion.

1 Tutorial description

High-level verbal reasoning tasks are increasingly used as de-facto Turing test proxies in evaluating the language capabilities of NLP systems. In particular, question answering (QA), natural language inference (NLI) and commonsense reasoning are included in evaluation suites and featured in most papers proposing new architectures. Accordingly, these tasks are seeing an explosion of datasets: there are already over 100 datasets only for QA, with over 40 published since 2018. This makes the choice of data for a given study a research-intensive task in itself.

The goals of the tutorial are as follows:

- to provide an up-to-date guide to the recent datasets for training verbal reasoning systems;
- to survey the old and new methodological issues;
- to outline the existing proposals for overcoming them, and to highlight the biggest remaining challenges.

This tutorial would be useful to NLP practitioners who simply want to pick a dataset and focus on modeling work, while being aware of potential issues that often go unnoticed. It would also be useful to the researchers working on new datasets and looking for the latest tips and tricks for overcoming the common pitfalls.

2 Details and Prerequisites

The tutorial will be of the *cutting-edge* type. The tutorial slides are available online at https://www.annargrs.github.io/dataset-explosion.

Prerequisites. We assume basic familiarity with the standard machine learning evaluation workflow and the three tasks that we are covering (question answering, commonsense reasoning, natural language inference). We also assume some familiarity with the methodology of crowdsourcing NLP datasets.

Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts, pages 27–32
Barcelona, Spain (Online), December 12th, 2020.

3 Reading list

The core approaches to machine reading comprehension and several widely-used datasets are covered in the survey by Qui et al. (2019). For NLI, we refer the reader to the surveys on resources and approaches (Storks et al., 2019b), as well as issues with the current data (Schlegel et al., 2020). A survey on benchmarks and approaches is also available for commonsense reasoning (Storks et al., 2019a).

4 Tutorial outline

The tutorial will present three hours of content with a thirty minute break.

Motivation. We will start by discussing the place of high-level reasoning tasks in the current NLP system evaluation paradigm: how the focus shifted away from the the low-level tasks such as POS-tagging, and how the low-level linguistic competences seems to be coming back (Ribeiro et al., 2020).

The dataset explosion. This first part of the tutorial will provide an overview of the main types of datasets for QA, NLI, and commonsense reasoning. For each sub-type, we will discuss representative dataset examples.

The field of question answering encompasses both open-world QA and reading comprehension (RC). Open-world QA focuses on factoid questions, with the answers typically extracted from web snippets or Wikipedia. The questions usually come from search engine queries (Kwiatkowski et al., 2019) and quiz data (Joshi et al., 2017). Bordering on open-world QA is the task of QA on structured data, such as tables and databases (Jiang et al., 2019).

Most current reading comprehension datasets are extractive (Rajpurkar et al., 2016; Dua et al., 2019), i.e. the correct answer is contained within the text itself, and the task is to find the correct span. Multiple-choice questions are harder to generate, as they need good confounds, and often come from curated test collections (Lai et al., 2017). Freeform answers remain rare (Bajaj et al., 2016), as evaluating them faces the general problem of evaluating language generation. Most RC datasets are single-domain, with a few exceptions (Reddy et al., 2018).

For NLI, we will organize the discussion in terms of domains covered by the current datasets: single-domain (Bowman et al., 2015), multi-domain (Williams et al., 2017), specialized domains (Romanov and Shivade, 2018). In both QA and NLI there have also been attempts to recast datasets from other tasks as QA/NLI problems (McCann et al., 2018; White et al., 2017), and researchers working on NLI often rely on the datasets for the related problem of RTE (Dzikovska et al., 2013).

Commonsense reasoning datasets come in different formats: multi-choice reading comprehension (Ostermann et al., 2018), extractive reading comprehension (Zhang et al., 2018), story completion (Mostafazadeh et al., 2017) and also as multi-choice questions for a single sentence input (Levesque et al., 2012). The task of commonsense reasoning is supposed to involve a combination of context-internal knowledge with context-external world knowledge, and we will briefly mention the major sources of such knowledge that are typically recommended in commonsense challenges, such as scripts (Wanzare et al., 2016), frames (Baker et al., 1998), and entity relations (Speer et al., 2017).

Reality check. One of the reasons there are so many new datasets is that most of them get "solved" very soon after publication, as it happened with CoQA (Reddy et al., 2018). However, this is not necessarily a testimony to the linguistic power of deep learning. It is becoming increasingly clear that, given the opportunity, our models exploit annotation artifacts and shallow lexical cues, achieving a high performance but not a high degree of language understanding. The second part of the tutorial will synthesize a string of papers exposing such issues (Niven and Kao, 2019; McCoy et al., 2019; Geva et al., 2019; Wallace et al., 2019).

To give a few examples, for QA it has been shown that human-level performance on SQuADcan be achieved while relying only on superficial cues (Jia and Liang, 2017), and 73% of the NewsQA can be solved by simply identifying the single most relevant sentence (Chen et al., 2016). A system trained on one QA dataset does not tend to perform well on another one, even if it is in the same domain (Yatskar, 2019). Research on adversarial attacks suggests that it is possible to find dataset-specific phrases that

will force a QA system to output a certain prediction when added to any input. For example, a SQuAD-trained QA system can be hacked in this way to always predict "to kill Americam people" as the answer to any question (Wallace et al., 2019).

In NLI, 67% of SNLI (Bowman et al., 2015) and 53% of MultiNLI (Williams et al., 2017) can be solved without looking at the premise (Gururangan et al., 2018). The HANS dataset showed that models trained on MNLI (Williams et al., 2017) actually learn to rely on shallow cues and can be fooled by syntactic heuristics (McCoy et al., 2019). Furthermore, the models trained on such datasets are unaware of lexical knowledge that would have enabled them to solve simple WordNet-based permutations of the original data (Glockner et al., 2018).

In commonsense reasoning, by definition, the challenge is to get the system to make decision based on both the current context and some general knowledge about the world. However, in the challenge of SemEval2018-Task 11 (Ostermann et al., 2018) most participants did not use any extra knowledge sources, and one of them still achieved 0.82 accuracy vs 0.84 achieved by the ConceptNet-based winner. It is argued that large pre-trained language models already possess much of such knowledge: for instance, BERT (Devlin et al., 2018) achieved over 86% on SWAG (Zellers et al., 2018).

We will also mention the widespread methodological problem of under-reporting of environment factors that may make as much difference as the proposed architecture changes. The effect of such factors as random seed, hardware, library versions has been discussed for several QA datasets (Crane, 2018).

Methodology developments and challenges. For existing datasets, simply removing annotation artifacts will not solve the problem, as it creates other exploitable artifacts (Gururangan et al., 2018). Among the recent improvements in the dataset collection methodology are complex queries that require aggregating information from several sources (Dua et al., 2019; Kocisky et al., 2018; Yang et al., 2018). Reliance on shallow patterns could be reduced by paraphrasing, including adversarial paraphrasing with a model-in-the-loop as an oracle that would reject questions that were too easy (Dua et al., 2019). Another alternative is balanced datasets with as many question types and genres (Rogers et al., 2020). Diversity can also be somewhat improved with partly synthesized data (Labutov et al., 2018), but any templates or annotator examples themselves are potential sources of bias.

Questions are more difficult if they are collected independently from the text (Kwiatkowski et al., 2019), written from summaries (Trischler et al., 2016) or hints (Choi et al., 2018). Finally, unanswerable questions (Rajpurkar et al., 2018) in conjunction with adversarial inputs should also force the model to go beyond lexical pattern-patching.

A radically different direction is shifting to exclusively out-of-distribution evaluation (Linzen, 2020), e.g with adversarial (McCoy et al., 2019) and multi-dataset (Fisch et al., 2019) evaluation. However, for that we still need to be aware of the training distribution, which becomes particularly challenging because with very large pre-trained models it is hard to guarantee that the test examples were not seen in pre-training (Brown et al., 2020).

5 Diversity efforts

The tutorial will be presented by an all-female team with a senior researcher and a post-doc as the lead organizer.

The survey will focus on English datasets, but we will provide references to the existing datasets in other languages that we are aware of.

6 Organizers

Anna Rogers, University of Copenhagen arogers@sodas.ku.dk

Research interests: distributional and cognitive semantics, interpretability and evaluation of deep learning models, computational social science.

Organization: LREC T4 tutorial on compositionality in distributional semantics, CogALex-V Shared Task on the Corpus-Based Identification of Semantic Relations, the Third Workshop on Evaluating Vector Space Representations for NLP (NAACL 2019), the First Workshop on Insights from Negative Results in NLP (EMNLP 2020).

Anna Rumshisky, University of Massachusetts Lowell arum@cs.uml.edu

Research interests: distributional semantics, biomedical and social NLP, temporal reasoning, machine learning/deep learning for NLP

Organization: Program Chair for NAACL 2021, Organizer for LREC T4 tutorial on compositionality in distributional semantics, SemEval-2017 task 6 (#HashtagWars: Learning a Sense of Humor), Clinical Natural Language Processing Workshop at COLING 2016, NAACL 2019 and EMNLP 2020, SemEval-2019 task 11 (Normalization of Medical Concepts in Clinical Narrative), The Third Workshop on Evaluating Vector Space Representations for NLP (NAACL 2019), the First Workshop on Insights from Negative Results in NLP (EMNLP 2020).

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv: 1611.09268 [cs]*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, 17-21 September 2015.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, June.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium.

Matt Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2368–2378.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China, November. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *EMNLP*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. 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)*, pages 650–655.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. 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)*, pages 107–112.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A New Factoid QA Data Set Matching Trivia-Style Question-Answer Pairs with Freebase. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 318–323.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria. 2018. Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in Simulated Worlds. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 833–844, Melbourne, Australia.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561.

Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? *arXiv:2005.00955 [cs]*, May.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The Natural Language Decathlon: Multitask Learning as Question Answering. *arXiv:1806.08730 [cs, stat]*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy.

Nasrin Mostafazadeh, Michael Roth, Nathanael Chambers, and Annie Louis. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-Level Semantics*, pages 46–51.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana.

Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A Survey on Neural Machine Reading Comprehension. *arXiv:1906.03824 [cs]*, June.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge. *arXiv:1808.07042 [cs]*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8722–8731.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2020. Beyond Leaderboards: A survey of methods for revealing weaknesses in Natural Language Inference data and models. *arXiv:2005.14709 [cs]*, May.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019a. Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches. *arXiv:1904.01172 [cs]*, April.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019b. Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. *arXiv:1904.01172 [cs]*, November.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A Machine Comprehension Dataset. *arXiv:1611.09830 [cs]*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. *EMNLP*.

Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. 2016. DeScript : A Crowdsourced Corpus for the Acquisition of High-Quality Script Knowledge. In *Language Resources and Evaluation Conference*, pages 3494–3501.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium.

Mark Yatskar. 2019. 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*, pages 2318–2323.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv:1810.12885 [cs]*.