Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr[1], **Clara Vania**[2], **Nikita Nangia**[3], **Maarten Sap**[4]
Mark Yatskar[5], **Samuel R. Bowman**[3] and **Yoav Artzi**[1]

[1]Cornell University [2]Amazon [3]New York University
[4]University of Washington [5]University of Pennsylvania

{suhr, yoav}@cs.cornell.edu
{nikitanangia, bowman}@nyu.edu vaniclar@amazon.co.uk
msap@cs.washington.edu myatskar@seas.upenn.edu

Abstract

Crowdsourcing from non-experts is one of the most common approaches to collecting data and annotations in NLP. Even though it is such a fundamental tool in NLP, crowdsourcing use is largely guided by common practices and the personal experience of researchers. Developing a theory of crowdsourcing use for practical language problems remains an open challenge. However, there are various principles and practices that have proven effective in generating high quality and diverse data. This tutorial exposes NLP researchers to such data collection crowdsourcing methods and principles through a detailed discussion of a diverse set of case studies.

1 Tutorial Description

Crowdsourcing from non-experts is one of the most common approaches to collecting data and annotations in NLP. It has been applied to a plethora of tasks, including question answering (Rajpurkar et al., 2016; Choi et al., 2018), textual entailment (Williams et al., 2018; Khot et al., 2018), instruction following (Bisk et al., 2016; Misra et al., 2018; Suhr et al., 2019a; Chen et al., 2019a), visual reasoning (Antol et al., 2015; Suhr et al., 2017, 2019b), and commonsense reasoning (Talmor et al., 2019; Sap et al., 2019b). Even though it is such a fundamental tool, crowdsourcing use is largely guided by common practices and the personal experience of researchers. Developing a theory of crowdsourcing use for practical language problems remains an open challenge. However, there are various principles and practices that have proven effective in generating high quality and diverse data. This tutorial exposes NLP researchers to such data collection crowdsourcing methods and principles through a detailed discussion of a diverse set of case studies.

The selection of case studies focuses on challenging settings where crowdworkers are asked to write original text or otherwise perform relatively unconstrained work. Through these case studies, we discuss in detail processes that were carefully designed to achieve data with specific properties, for example to require logical inference, grounded reasoning or conversational understanding. Each case study focuses on data collection crowdsourcing protocol details that often receive limited attention in research presentations, for example in conferences, but are critical for research success. We introduce the task of each case study, and do not assume prior knowledge. Where possible, we highlight common trends, or otherwise key differences between the discussed case studies.

Relevance to the NLP Community Crowdsourcing techniques are commonly used, but rarely discussed in detail. This tutorial provides a detailed description of crowdsourcing decisions in complex scenarios and the reasoning behind them. NLP researchers aiming to develop new datasets, tasks and data collection protocols will find the content directly applicable to their own work. A strong understanding of data collection practices and the range of decisions they include will also aid researchers using existing dataset to critically assess the data they use, including its limitations.

Post-tutorial Materials The tutorial videos, slides and other material will be made available publicly online following the tutorial.

2 Structure and Content Overview

The tutorial spans three hours (180 minutes), and is divided into eight sections:

Introduction (10 min) A brief introduction to the tutorial structure, its goals, and the case studies.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 1–6
Punta Cana and Online, November 10–11, 2021.

Background (20 min) A high-speed recap of established crowdsourcing concepts and terms. We refer back to the content of this section in the case studies. This section includes the basic structure of a Mechanical Turk task (HIT), typical incentive mechanisms, typical communication mechanisms, typical worker qualification and screening mechanisms, as well as relevant results about the demographics and expressed preferences of crowdworkers and the crowdworker community.

Case Study I: MultiNLI (45 min) We discuss the MultiNLI (Williams et al., 2018) corpus, with primary focus on experiments from subsequent papers that extend or evaluate the data collection protocol used to create this dataset. MultiNLI is built around the task of natural language inference (a.k.a. textual entailment; Dagan et al., 2006; MacCartney, 2009): given two sentences, the task is to identify (roughly) whether the first sentence entails the second. We start with this case study not because of any unique success of the data collection protocol, but because MultiNLI and the natural language inference task have emerged as a popular testbed for data collection methods and for relevant data analysis methods in NLP. Topics include:

- The development of a simple crowdworker-writing protocol for natural language inference data (Marelli et al., 2014; Bowman et al., 2015; Williams et al., 2018)
- Known issues with artifacts, social bias, and debatable judgments in data collected under this protocol (Rudinger et al., 2017; Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018; Pavlick and Kwiatkowski, 2019)
- Experiments evaluating data collection feasibility under variants of the base task definition (Chen et al., 2020; Bowman et al., 2020)
- Studies evaluating the feasibility of collecting data for the same task using alternative protocols (Nie et al., 2020; Kaushik et al., 2019; Bowman et al., 2020; Vania et al., 2020; Parrish et al., 2021)

Case Study II: NLVR (25 min) Natural Language for Visual Reasoning comprises two datasets, NLVR (Suhr et al., 2017) and NLVR2 (Suhr et al., 2019b), both study natural language sentences grounded in visual context.[1] The task is to determine whether a caption is true or false about a paired image. The data was collected to require reasoning about object quantities, comparisons between object properties, and spatial relations between objects. NLVR2 is used as evaluation data for numerous language-and-vision systems (e.g., Tan and Bansal, 2019; Chen et al., 2019c). Both datasets were crowdsourced with a contrastive captioning designed to elicit linguistically complex sentences and to naturally balance the datasets between true and false examples. NLVR2 also uses a tiered system during crowdsourcing including distinct pools of annotation tasks for experienced workers and new workers.

Case Study III: CerealBar (25 min) CerealBar (Suhr et al., 2019a) is a game designed for studying collaborative natural language interactions, released alongside a dataset of interactions between human players.[2] CerealBar emphasizes collaboration through natural language instruction between agents with differing abilities. Each of the agents can be a human user or a learned model. CerealBar has been used to design and train systems that follow instructions by grounding them in the surrounding environment and acting in the environment. The game rules were explicitly designed with the intent of eliciting rich collaborative interactions across many instructions, for example by allowing a pair of players that is scoring well to continue playing for longer, thereby collecting more data from successful collaborations. The CerealBar data collection process included a development of a community of players, which has demonstrated behavioral and linguistic change over the crowdsourcing process.

Case Study IV: QuAC (25 min) Question Answering in Context is a dataset for studying information seeking dialogs between a student and a teacher (Choi et al., 2018). Given a subject heading, a student questions a teacher, who responds by copying spans from a Wikipedia article. The goal of the pair is to maintain a dialog of sufficient length without encountering too many unanswerable questions. The task is to play the role of the teacher: answering questions of an interested student. The collection protocol is unique in that two unreliable workers had to be coordinated for sufficient time to accomplish a meaningful dialog. QuAC collection relied on several strategies to keep

[1] http://lil.nlp.cornell.edu/nlvr/

[2] http://lil.nlp.cornell.edu/cerealbar/

partners from leaving interactions, such as allowing workers to simultaneously participate in multiple related dialogs, a feedback system teachers used to help students formulate questions, and scaling incentives that included punitive elements.

Case Study V: SOCIALIQA (25 min) SOCIALIQA (Sap et al., 2019b) is the first large-scale benchmark to test model emotional and social reasoning through 38k questions about everyday situations. The distributional nature of social commonsense knowledge requires the answer candidates to cover the plausible and likely, as well as the plausible but unlikely, as opposed to right/wrong answer candidates as common in other QA benchmarks. SOCIALIQA introduces a question-switching technique for crowdsourcing these unlikely answers, to overcome the possible stylistic artefacts in negative answers (e.g., negations, out-of-context responses; Schwartz et al., 2017). Additionally, to achieve large-scale and broad coverage, SOCIALIQA used a multi-stage crowdsourcing pipeline to expand seed events from the ATOMIC (Sap et al., 2019a) commonsense knowledge graph into full-fledged social situations.

Summary (5 min) A brief summary of the tutorial, including the main takeaways from the different cases studies and repeating themes.

3 Breadth

The set of case studies covers a broad and diverse set of task types, including large-scale inference tasks (e.g., NLI), small-scale interactive tasks (e.g., CerealBar), and multi-modal grounded tasks (e.g., NLVR). The aim of this broad distribution is to cover the most common task and data scenarios in NLP. We focus on details that are rarely discussed fully in papers. The set of case studies covers a broad and diverse set of task types, including large-scale inference tasks (e.g., NLI), small-scale interactive tasks (e.g., CerealBar), and multi-modal grounded tasks (e.g., NLVR). The aim of this broad distribution is to cover the most common task and data scenarios in NLP. The case studies cover the research of four distinct research labs. For each case study, we will also discuss related work from other authors as is relevant. For example, the MultiNLI case study will include extensive discussion of followup work and the SocialIQA case study will discuss related commonsense resources. In addition, we will discuss relevant existing work to provide all necessary background (e.g., Dumitrache et al., 2018; Chen et al., 2019b; Ramírez et al., 2019).

4 Prerequistites

Broad familiarity with NLP tasks, empirical evaluation methods, and data collection practices. We introduce all the necessary terms and the specifics of each case study.

5 Reading List

We recommend reviewing the 2015 NAACL tutorial on crowdsourcing.[3] While we focus on unconstrained and complex case studies, the 2015 tutorial provides an overview of basic terms and methods that is a complementary background to our material. However, we review the required material in the background section, and do not assume a familiarity with the content of this prior tutorial. We also recommend reading the main papers describing each of the case studies (Williams et al., 2018; Suhr et al., 2017, 2019b,a; Choi et al., 2018; Sap et al., 2019b).

6 Presenters

Alane Suhr
PhD Student, Cornell University
suhr@cs.cornell.edu
https://alanesuhr.com
Alane's research focuses on grounded natural language understanding. Alane has designed crowdsourcing tasks for collecting language data to study situated natural language understanding. Alane co-presented a tutorial in ACL 2018.

Clara Vania
Applied Scientist, Amazon
vaniclar@amazon.co.uk
https://claravania.github.io/
Her research focuses on crowdsourcing, transfer learning, and multilingual NLU. Recently, she has been working on semi-automatic data collection for natural language inference and crowdsourcing methods for question answering.

Nikita Nangia
PhD student, New York University
nikitanangia@nyu.edu
https://woollysocks.github.io
Nikita's work focuses on crowdsourcing methods

[3] http://crowdsourcing-class.org/tutorial.html

and data creation for natural language understandingsoft. Her recent work explores using incentive structures to illicit creative examples. Nikita co-organized a tutorial on latent structure models for NLP at ACL 2019.

Maarten Sap
PhD student, University of Washington
msap@cs.washington.edu
http://maartensap.com/
His research focuses on endowing NLP systems with social intelligence and social commonsense, and understanding social inequality and bias in language. His substantial experience with crowdsourcing includes the collecting of the SOCIALIQA commonsense benchmark as well as the creation of knowledge graphs with inferential knowledge (ATOMIC, Social Bias Frames).

Mark Yatskar
Assistant Professor, University of Pennsylvania
myatskar@seas.upenn.edu
https://markyatskar.com/
His research focuses on the intersection of natural language processing and computer vision. Mark's work has resulted in the creation of datasets such as imSitu, QuAC and WinoBias and recent research has focused on gender bias in visual recognition and coreference resolution.

Sam Bowman
Assistant Professor, New York University
bowman@nyu.edu
https://cims.nyu.edu/~sbowman/
Sam works on data creation, benchmarking, and model analysis for NLU and computational linguistics. Sam has had a substantial role in several NLU datasets, including SNLI, MNLI, XNLI, CoLA, and BLiMP, and his recent work has focused on experimentally evaluating methods for crowdsourced corpus construction.

Yoav Artzi
Associate Professor, Cornell University
yoav@cs.cornell.edu
https://yoavartzi.com/
Yoav's research focuses on learning expressive models for natural language understanding, most recently in situated interactive scenarios. Yoav led tutorials on semantic parsing in ACL 2013, EMNLP 2014 and AAAI 2015.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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. Association for Computational Linguistics.

Samuel R Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. Collecting entailment data for pretraining: New protocols and negative results. In *Proceedings of EMNLP*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019a. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Dan S. Weld. 2019b. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019c. UNITER: Learning universal image-text representations. *ArXiv*, abs/1909.11740.

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*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer, New York, NY.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. In *Joint Proceedings SAD 2018 and CrowdBias 2018*, CEUR Workshop Proceedings, pages 11–18. CEUR-WS.

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. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University, Stanford, CA.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3D environments with visual goal prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R Bowman. 2021. Does putting a linguist in the loop improve nlu data collection? In *To appear in Findings of EMNLP*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. 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 the Conference on Empirical Methods in Natural Language Processing*.

Jorge Ramírez, Simone Degiacomi, Davide Zanella, Marcos Baez, Fabio Casati, and Boualem Benatallah. 2019. Crowdhub: Extending crowdsourcing platforms for the controlled evaluation of tasks designs. *arXiv preprint 1909.02800*.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *EMNLP*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *CoNLL*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019a. Executing instructions in situated collaborative interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019b. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. 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. European Language Resources Association (ELRA).

Clara Vania, Ruijie Chen, and Samuel R. Bowman. 2020. Asking crowdworkers to write entailment examples: The best of bad options. In *Proceedings of AACL*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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. Association for Computational Linguistics.