Stanford CS 329X - Human-Centered NLP

Lecture 7: Data Collection and Curation

04/26/23

Lecturer: Diyi Yang, Mitchell Gordon *Readings: N/A*
Scribe: Brennan Megregian

1 Introduction

Data annotation is an essential part of every NLP project, as data is used to provide training data for your system, and to evaluate how well your system is working. In genearl, this lecture outlined and focused on the importance of proper data collection and curation techniques for Natural Language Processing tasks.

2 What makes a good dataset?

Understanding features that contribute to a 'good' and useful dataset is key when considering using or curating one. The following are examples of these features from Bowman and Dahl [2021]:

- **Validity.** A dataset should correspond well to the task, domain, and language it is designed for. Good performance on the dataset should imply robust in-domain performance on the task. A good evaluation dataset should have comprehensive coverage of language variation, test cases isolating all necessary task skills, and no artifacts that let bad models score highly.
- **Reliable Annotation.** The labels in the dataset should be correct and reproducible. Try to avoid having examples that are carelessly mislabeled, that have no clear correct label due to unclear or underspecified task guidelines, and that have no clear correct label under the relevant metric due to legitimate disagreements in interpretation among annotators.
- **Statistical Power.** Benchmarks should be able to detect qualitatively relevant performance differences between systems. Since our systems continue to improve rapidly, though, we should expect to be spending more time in the long tail of our data difficulty distributions.
- **No Social Bias.** Benchmarks should reveal plausibly harmful social biases in systems, and shouldnt incentivize the creation of biased systems. We need to better encourage the development and use auxiliary bias evaluation metrics.

3 How do we get a good dataset?

Before you consider obtaining or building a dataset, first it is important to know your end goal before you start collecting and annotating data points [Bowman and Dahl, 2021].

Tips for data collection, and how to try to get the desired data carefully:

1. Good performance on the benchmarks should imply robust in-domain performance on the task. We need more work on dataset design and data collection methods.

2. Benchmark examples should be accurately and unambiguously annotated. Test examples should be validated thoroughly enough to remove erroneous examples and to properly handle ambiguous ones.

Tips for data curation, and how to modify you collected dataset (augment it, fill in gaps etc.) so its more [difficult, fair, usable, etc.]:

1. Benchmarks should offer adequate statistical power. Benchmark datasets need to be much harder and/or much larger.

2. Benchmarks should reveal plausibly harmful social biases in systems, and should not incentivize the creation of biased systems. We need to better encourage the development and use auxiliary bias evaluation metrics.

4 Annotation procedure

A typical annotation process usually has 3-4 steps:

1. Explain the dataset, annotation instruction, label definitions, etc.

2. Use some examples to help annotators better their task.

3. Actual task - provide labels for multiple examples.

4. Optionally, can involve a survey to get annotators' feedback.

4.1 Labeling Instruction

Describe the task, and the label definitions. Show what the annotators will see in each labeling round. Explain every visualization on the UI. Explain the entire process. Collect student's consent.

Highlight Warning. Annotators are noisy. Warn them beforehand that you might reject their work if their label quality is bad. Important, otherwise annotators will be surprised when they are rejected, and will complain.

Pilot Study. Run pilot studies e.g. ask your friends to go through the annotation first, tell them to ask you questions on things that are unclear.

4.2 Training Process

- The training interface should be the same as the actual labeling task interface.
- Train people with examples that have different labels.
- Use a combination of simple examples (show a typical task), and edge cases (help them make decisions on ambiguous cases).
- Training examples have groundtruth labels.
- Provide clear feedback when people are correct/incorrect.
- Only allow them to proceed if an annotator gets all training labels correct.

4.3 Actual Labeling Process

Once people pass training, they can proceed with the actual task. Always allow annotators to review the annotation requirement in a popup window.

5 Case Study: Story behind the LitBank Dataset

In this video, David Bamman discusses the process of building and curating a dataset of novels in an effort to analyze the meaning of culture [Bamman, 2022]. In particular, Bamman outlines the steps taken to ensure the dataset was diverse and representative of different populations. This involved a more deliberate selection of data examples, choosing from a range of different categories including bestsellers, Pulitzer prize nominees, a variety of different genres, and specifically looking at novels written by black and other minority authors.

Analyses of novels and films demonstrate key perceptions and portrayals of men versus women:

- Contemporary novels favor heteronormative pairs [Kraicer and Piper, 2019].
- Men often have more agency and power than women in film [Sap et al., 2017].
- Women are depicted as the linchpins of information flow [Sims and Bamman, 2020].

6 Data documentation and sharing

6.1 Data Card

Data Cards are for fostering transparent, purposeful and human-centered documentation of datasets within the practical contexts of industry and research. They are structured summaries of essential facts about various aspects of ML datasets and provide explanations of processes and rationales that shape the data and consequently the models, such as the following from Sap et al. [2017]:

- The authors of the dataset.
- The purpose for which the dataset was created.
- The problem space the dataset curation was aiming to solve.
- The motivations for the creation of the dataset.
- What sort of tasks the dataset is intended to be used for/applied to.
- How to use the dataset and known caveats.
- The primary data type.
- A snapshot of the datset including its size, whether it includes human annotated labels (and if so how many), etc.
- How to interpret a datapoint.
- An example of a datapoint.
- The source of the collected data and the methods used for collection.
- The labeling process and method - e.g. did humans annotate the labels?
- Analysis on data distribution.

Data Cards have many, many relevant and useful information. They help us decide when we can/cannot use a dataset. Its supported by mainstream libraries like Hugging Face. But this is too much information and a lot of data creators dont pay attention.

Data statements will help alleviate issues related to exclusion and bias in language technology, lead to better precision in claims about how NLP research can generalize and thus better engineering results, protect companies from public embarrassment, and ultimately lead to language technology that meets its users in their own preferred linguistic style and furthermore does not misrepresent them to others [Bender and Friedman, 2018].

7 Task Ambiguity

Human annotators guess based on personal belief and won't always agree with the consensus gold label. NLP models guess based on a model of the *typical* annotator and may agree with the gold label *more* often.

7.1 Addressing Task Ambiguity: Iterative Design

Run pilot studies to gather potential edge cases. If you have a fixed definition for a subcategory, add them as part of your instruction [Bragg et al., 2018]. But sometimes you wont be able to capture all the edge cases, or you dont want to force people to converge this early [Chang et al., 2017]. Collect additional justification from people. Make the decision boundary later later, or use uncertainty in other ways.

7.2 Task Ambiguity By Population

Significant inconsistency can exist in rater behavior within and across various subgroups. This leads to unreliability of gold labels: Majority-based and/or instruction based gold label may be unreliable for a significant portion of the data, if the replication per item is low. In many cases, people within a target group (e.g. Muslim) should have more voice power in labeling. Based on studies showing whether Americans and Indians choose to label potentially toxic posts as either 'safe' or 'unsafe', US raters produced ratings that are significantly similar to each other, compared to Indian raters on average [Aroyo et al., 2023]. Female raters also produced ratings that are very similar to each other, and significantly dissimilar to the ratings produced by male raters; Male raters also showed high variance in their disagreement [Aroyo et al., 2023].

Addressing Task Ambiguity By Population. One way to consider addressing this variation due to differences in population is to model each annotator and their population individually; Given an example, the model guesses what each annotator might do and decides what is the best population for labeling [Gordon et al., 2022].

8 Learning from Limited Data

- **Transfer learning.** Leverage data from a different-but-related task.
- **Few/zero-shot learning.** Generalize to new tasks after seeing a few (or no) examples of that task.
- **Multitask learning.** Use information learned on different tasks for mutual benefit.
- **Data augmentation.** Modify labeled data to with class-preserving transformations.
- **Semi-supervised learning.** Learn from labeled and unlabeled data.

9 Summary

- We need data thats representative, reliable, unbiased, large and difficult, but data collection is harder than we think.
- Data sources, annotator distribution, task definition, etc. all have significant impact on labeling results.
- Most popular labeling platform is MTurk, but should carefully design for its limitations.
- Also, naturally collected data is hardly perfect, so data curation is important.
- Data-centric AI is the discipline of systematically engineering the data used to build an AI system.

10 Fireside Chat with Mitchell Gordon

In his talk, Mitchell Gordon discusses the concept of 'Jury Learning', which integrates dissenting voices into machine learning models [Gordon et al., 2022]. Jury learning aims to answer questions such as whose voices or labels machine learning algorithms should try to emulate. The jury learning model aggregates different groups of people into individual pseudo-humans. Mitchell argues that current datasets already have implicit juries, as they somehow decide which labels given by different human annotators are 'correct'. However, jury learning aims to make these juries within datasets *explicit*. Firstly, the content-based jury is selected (i.e. the 'jury' members are selected based on the content and characteristics in the datapoints). For each juror, there must be 100 people in the dataset that fit their characteristics. The model is then trained to emulate individual people, then some majority decision/vote over jurors is taken to get a final verdict. The goal of this task is to better represent real human-decisions, specifically in human-centered tasks and labeling.

References

Lora Aroyo, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. The reasonable effectiveness of diverse evaluation data, 01 2023.

David Bamman. Building datasets for the analysis of culture. In *Sharing Stories and Lessons Learned Workshop at EMNLP*, 2022.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL `https://aclanthology.org/Q18-1041`.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding?, 2021.

Jonathan Bragg, Mausam, and Daniel S. Weld. Sprout: Crowd-powered task design for crowdsourcing. New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359481. doi: 10.1145/3242587.3242598. URL https://doi.org/10.1145/3242587.3242598.

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 23342346, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3026044. URL https://doi.org/10.1145/3025453.3026044.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. ACM, apr 2022. doi: 10.1145/3491102.3502004. URL https://doi.org/10.1145%2F3491102.3502004.

Eve Kraicer and Andrew Piper. Social characters: The hierarchy of gender in contemporary english-language fiction. *Journal of Cultural Analytics*, 01 2019. doi: 10.22148/16.032.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1247. URL https://aclanthology.org/D17-1247.

Matthew Sims and David Bamman. Measuring information propagation in literary social networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 642–652, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.47. URL https://aclanthology.org/2020.emnlp-main.47.