

## Chapter 1: Conclusions on Natural Language Processing Data

In this thesis, we create natural language processing datasets using three types of users: **crowd-workers**, **experts**, and a **hybrid** combination. We argue that improving data quality with reliable data **generators** and **annotators** is paramount towards establishing new NLP tasks. As examples, we propose a new task, cultural adaptation, that uses verified cultural experts for the creation of gold labels (Chapter ??). Additionally, we introduce a novel self-annotated deception dataset by working with top players from the Diplomacy community (Chapter ??). Last, we create the largest goal-oriented dialogue dataset by pairing Amazon customer support associates with crowd workers (Chapter ??).

These tasks would not be possible by using **found** or **crowd-sourced** data. Several projects show the limitations of creating large datasets in this way. Using text-to-speech to automatically generate questions scales at the expense of diversity and realism in the data (Chapter ??). Using an **expert** to design, but not generate, a formulaic dataset for assessing coreference resolution creates unlikely phrases (Chapter ??). Using the **crowd** to generate question rewrites can increase the amount of training data for question answering, but requires extensive quality control (Chapter ??).

## 1.1 Creating Timeless Natural Language Processing Datasets

Datasets that have withstood the test of time in natural language processing were painstakingly created and quality controlled. The Penn Treebank (Marcus et al., 1993) was collected and refined for years using graduate students in linguistics as annotators. The annotation process had extensive experimental design, annotators underwent extensive training, and the data was evaluated for disagreements. That effort caused graduate students today to learn about it.

The granularity and quantity of NLP datasets continues to increase as machine learning expands to new languages and tasks. Quality control is usually an afterthought in a conference paper paradigm that rewards quantity. However, this mindset introduces room for error, potentially with real-life repercussions (Wallace et al., 2021). The importance of NLP to modern day life in communication, information gathering, and commerce means that decisions made in an academic context can have wide-ranging implications. Authoritative, realistic, and diverse datasets are less likely to contain errors or artifacts and more likely to be used in years to come than larger datasets derived from Wikipedia or crowd-sourced knowledge.

Recent work questions conventional wisdom about data in NLP. Rodriguez et al. (2021) question the paradigm of using quantitative leaderboards in question answering, given the disparity of question difficulties. van der Goot (2021) question the paradigm of using a development set for model tuning. Kummerfeld (2021) question the qualification requirements for Mechanical Turk workers. Last, Karpinska et al. (2021) question the output of Mechanical Turk workers for evaluation. We

question if the human generating the language data has been forgotten.

## Bibliography

- Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.
- Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.
- Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.
- Eric Wallace, Tony Z. Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on nlp models. In *Conference of the North American Chapter of the Association for Computational Linguistics*.