Chapter 1: The Case for Upfront Investment in Data

Computation can solve tasks across multiple areas of scientific inquiry: natural language processing, computer vision, biology, etc. Solving tasks for all these domains—translating a sentence between languages, distinguishing a cat from a dog, classifying a mutation—has two abstract and intertwined dependencies: model-building and data collection.[1] The relationship is intertwined since today's models are optimized to draw statistical conclusions from significant amounts of data through machine learning. But, even the most cutting edge modeling techniques are heavily dependent on having *realistic* and *accurate* data for solving a task. These large datasets are primarily gathered from online repositories or created through low-cost crowd-sourcing (Deng et al., 2009; Rajpurkar et al., 2016; Budzianowski et al., 2018), which are often *artificial* or *inaccurate*. We argue that a new paradigm of high-quality, expert-reliant data collection can lead to long-term improvements in Natural Language Processing (NLP) and enable complex, novel tasks.

[1] Mitchell (1997) defines a machine learning model as, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". This E depends on data collection.

1.1 Defining Data: Annotation and Generation

In the overview, we discuss the two tasks necessary for data collection and explain the importance of data quality for computer science as a field.

Data creation can be broadly categorized into two categories: **generation** and **annotation**. We define **generation** as the creation of a data item that is not previously available (e.g., sequencing a genome, creating a new image, gathering a new sentence from a user, or automatically creating a sentence) (Atkins et al., 1992; Goodfellow et al., 2014; Zhu et al., 2018). We define **annotation** as the application of a label to an existing data item (e.g., classifying a part of the genome, labeling an image as a cat, or describing the sentiment of a sentence) (Deng et al., 2009; Finin et al., 2010; Kozomara and Griffiths-Jones, 2014). In many fields, data must be both **generated** to be representative of the task and then accurately **annotated** to be effective.

1.2 Quantity over Quality as a Paradigm

The demand of neural models for quantity has caused models to be trained on large, noisy data (Brown et al., 2020). The building blocks of other research areas—gene sequences in biology and individual pixels in computer vision—are not readily human interpretable by default. Even in more human-intuitive fields, like natural language processing, data have reached the scale where its veracity—the certainty and completeness of the data—cannot be assumed (Qiu et al., 2016), despite the

early assertions by Atkins et al. (1992). They posit that, "there is in fact little danger of obfuscation for the major parameters that characterize a corpus: its size (in numbers of running words), and gross characterizations of its content."[2] However, the objectivity of size is questionable; a corpus consisting of the same word repeated a million times clearly differs from one with a million unique words.

This focus on quantitative metrics evaluation metric has shaped NLP data creation during the past decade (Rodriguez et al., 2021). A dataset paper will comment on the amount of words, sentences, questions, etc., but with no assessment of their quality. But, the sheer quantity of data masks biases and artifacts, as they are no longer obvious to the naked human eye (Pruim et al., 2015; Gururangan et al., 2018; Gor et al., 2021). Since current approaches to machine learning often obscure how decisions are made by a model, the quality of the data is not immediately questioned as a culprit when a false prediction is made.

The current paradigm of crowd-sourcing—"obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers" (Merriam-Webster)—for dataset creation has been the main impetus of unreliability in data. Specifically, natural language processing has generally depended on low-cost crowds following the popularity of ImageNet (Deng et al., 2009). However, the entirely crowd-sourced annotations still have notable problems after a decade of updates (Yang et al., 2020) and should serve as a cautionary tale. A re-prioritization to working with users that have a reputation incentive to generate realistic and reliable data is

[2]Additionally, they crucially comment that the evaluation of corpora has not been standardized.

a solution to this problem.

1.3 The Nuance of Using Text as Data

We introduce the Natural Language Processing tasks covered in our work, challenges faced in NLP due to trade-offs of annotation speed and quality, and the two parts of data collection which impact the quality.

A large focus of NLP is on building models that exploit patterns in language data to solve a variety of tasks: question answering, conversational agents, machine translation, information extraction, etc. However, in the current paradigm of machine learning, models answer questions or make translations based on existing training data. This makes *realistic* data a prerequisite for any model that aims to *realistically* solve a language task.

But, the prevalence of neural model in NLP has prioritized data size over realism. Chapter **??** describes the history of data collection in NLP and explains why this dependence has grown over time. At the extreme end, GPT-3 is trained on 499 *billion* tokens, *de facto* training a neural model based on the entire Internet (Brown et al., 2020). However, not everything on the Internet is relevant or accurate! This is significant since training data containing low-quality data unsurprisingly leads to models learning controversial or false conclusions, with high levels of confidence (Wolf et al., 2017; Wallace et al., 2019). Therefore, missing or false data in the data **generation** process undermines the ability of NLP to solve language tasks in a realistic manner.

Furthermore, many tasks in NLP depend on accurate **annotation** of the raw

data. As a thought experiment, if all verbs are labeled as nouns and all nouns are labeled as verbs in the training data, a perfectly designed language model would be confidently wrong in its predictions. Crowd-sourcing with generalists (Buhrmester et al., 2011) assumes that enough unspecialized workers will answer a question correctly. This is a valid assumption for unambiguous, multiple-choice annotation with a large amount pool of annotators. However, many annotation tasks, such as span-annotation or candidate selection, have so many parameters that they are akin to language **generation** and cannot be easily verified through IAA (Karpinska et al., 2021). Therefore, NLP **annotation** needs to be accurate, at least in aggregate.

1.4 Data Quality as a New Paradigm

Investing in reliable data—as defined by its **generation** and **annotation** dimensions—upfront has two benefits. First, this improvement in the quality and diversity of data is a prudent long-term investment as high-quality datasets can have shelf-lives of decades (Marcus et al., 1993; Miller, 1995) while model architectures are frequently supplanted (Vaswani et al., 2017; Peters et al., 2018; Devlin et al., 2019). Second, using experts for data generation can enable tasks not otherwise possible; generalists cannot annotate medical images nor generate sentences in a language which they do not speak.

We use experts in three experiments to collect NLP corpora and contrast them with past automated and crowd-sourced ones. First, we show the limitations of using automated methods of data collection (Chapters **??** and **??**). Second, we show that

crowd-sourcing can **generate** data in a flexible but inaccurate manner (Chapter **??**). Third, we show the merits of using experts as **annotators** for data evaluation for a subjective and novel named entity adaptation task (Chapter **??**). Fourth, we describe an experiment that uses experts for both **generation** and **annotation** to study deception through the medium of a board-game (Chapter **??**). Last, we discuss a hybrid approach—using verified experts paired with external, low-cost data sources (Vukovic and Bartolini, 2010) (Chapter **??**) that can mitigate some of the accuracy issues while scaling in size and cost.

Bibliography

Sue Atkins, Jeremy Clear, and Nicholas Ostler. 1992. Corpus design criteria. *Literary and linguistic computing*, 7(1):1–16.

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. In *Proceedings of Advances in Neural Information Processing Systems*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of Empirical Methods in Natural Language Processing*.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on psychological science: a journal of the Association for Psychological Science*, 6 1:3–5.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Timothy W. Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680.

Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021. Toward deconfounding the influence of subject's demographic characteristics in question answering. In *Emperical Methods in Natural Language Processing*, page 6.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT*.

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*.

Ana Kozomara and Sam Griffiths-Jones. 2014. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*.

Merriam-Webster. Crowdsourcing. In *Merriam-Webster.com dictionary*.

George A. Miller. 1995. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Tom Mitchell. 1997. Introduction to machine learning. *Machine Learning*, 7:2–5.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Raimon H. R. Pruim, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. 2015. Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *NeuroImage*, 112:267–277.

Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1):67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*.

Maja Vukovic and Claudio Bartolini. 2010. Towards a research agenda for enterprise crowdsourcing. In *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*, pages 425–434. Springer.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of Empirical Methods in Natural Language Processing*.

Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.