Considerations for the social impact of natural language processing

Amandalynne Paullada

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Fei Xia, Chair

Trevor Cohen

Hannaneh Hajishirzi

Meliha Yetişgen

Program Authorized to Offer Degree:

Linguistics

University of Washington

Abstract

Considerations for the social impact of natural language processing

Amandalynne Paullada

Chair of the Supervisory Committee:
Professor Fei Xia
Department of Linguistics

Natural language processing (NLP) technologies have transformed how people access information and communicate with one another. It has thus become critical to take stock of the social impact of natural language processing technologies. In this thesis, I review practices at different stages of development for NLP systems and examine some of the issues that arise in turn, considering the social and political contexts that shape how systems are developed and deployed.

This thesis contributes three case studies of natural language processing technologies which exemplify many of the key issues in data collection practices and real-world system usage. The first two case studies situate computational models of text and machine translation in the complex social and political contexts that have informed the development of these applications. The third case study involves a reflection on original work in building and evaluating a system for representing biomedical relationships learned from text. In addition to the findings from these case studies, I contribute a practice-based framework for reflecting on factors that influence social impact at various stages of NLP system development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

Countless people have supported me on this journey, and I couldn't possibly name them all. I am grateful and humbled to have had the support and the shelter to complete a dissertation amid the ongoing COVID-19 pandemic, the record-breaking forest fire smoke and record-breaking heat, and all the other ambient threats to well-being on Earth.

First, I want to thank my advisor, Fei Xia, for giving me the freedom to forge my own academic path and the independence to pursue what really interests me — and, in the writing of this dissertation, for the guidance in putting all the pieces together and for the patience and moral support to get me across the finish line. Thanks as well to Trevor Cohen for taking me on as a research assistant, thereby giving me a profoundly interesting topic, and for your steady support as I have progressed in my studies. Thanks to my committee members, Meliha Yetişgen, Hannaneh Hajishirzi, and Annie Chen for all their insights that have helped to sharpen my work. Any lingering shortcomings of the present work are mine alone.

A chance meeting with Deb Raji in Summer 2019 (at the house where I'm finishing my dissertation) changed the trajectory of my studies for the better. Thank you, Deb, for being a constant positive force and for really believing in me and my ideas at a time when it felt like no one did. Thank you to Deb and our incredible co-conspirators (Emily M. Bender, Emily Denton, and Alex Hanna) for creating a thoughtful and rigorous space to explore the social impacts of 'artificial intelligence' together and for being the incredible role models you are.

A number of people helped me find my footing in the earlier years of my PhD. Thank you Byron Wallace for your generous advice, and for introducing me to

Trevor. Thank you Waleed Ammar for giving me a chance I never thought I'd have, and to the folks at AI2, in particular Kyle Lo, Sergey Feldman, and Mark Neumann, who taught me so much.

Thank you Anna Lauren Hoffmann for your vital teachings, and for being a radical and inspirational guide to a vocabulary and an ontology that I had been anxiously and clumsily grasping for.

Thank you to my linguist friends who first planted the possibility of graduate studies in my mind, including Mark Norris and Kelsey Kraus. Thank you, Kristen Sheets, for sharing your excellent taste and for fun field trips on more than one US coast. Yadav Gowda: I blame you for my mispronunciation of "Raynaud's" during my defense. Does this bus to go the bridge?

To the ones I left behind in Boston, thank you for letting me go.

Thank you to the friends, classmates, and colleagues that have enriched my life in Seattle. Thank you, Amanda, for much needed perspective on all things life, love, and scholarship. Thank you, Chris Haberland, for being the first one to ask me 'What if you did both?' when I couldn't choose between dissertation topics. It's proven to be one of the most generative questions of my research thus far. Thank you for helping me get to where I am today. Thank you, Rachael Tatman, for helping me become a researcher and for a sorely needed pep talk in the days leading up to my defense. Thank you, Julian Michael, for giving me a place to stay when I needed one, and for long and often hilarious conversations. Thank you to Marina Shah, Angie McMillan-Major, and the rest of the CLAMS crew for late night shenanigans. Thank you, Naomi Shapiro, for TV reruns, tasty snacks, and your boundless empathy and warmth. Thank you Courtney Mansfield, Kristen Howell, Levin Kim, Olga Zamaraeva, and so many other classmates for coffee-fueled co-working sessions, commiseration, and co-celebration. To the UW Linguistics 2016

incoming class (Ajda Gokcen, Ben Jones, and Jiahui Huang): We did it. Jiahui, you are so sorely missed.

Las chicas de la group chat (Ana Marasović, Lucy Wang, Swabha Swayamdipta): thank you for encouraging me to take myself more seriously, and for being ready to party when it's time to not be so serious. Mako, Mika, Lucy, Bryan, and Will: thank you for extravagant meals, energizing jogs, entertaining puzzles, and constant reassurance. Thank you for putting up with the absolute worst version of me, day in and day out, for the last year and a half.

Thank you to Robin Blythe, one of my earliest research collaborators and a frequent source of entertainment and moral support from the other side of the globe. Thank you for knowing me better than almost anyone else in the world and still being my friend in spite of it.

To Gary Berg and Dale Berg: thank you for art and music, for grief and joy, and for some of the most lively virtual socializing of my life thus far. Thank you, Gary, for giving me a place to write, physically and spiritually.

Thank you to my family near and far. 한국에 계신 사랑하는 가족한테 많이 감사합니다. Thank you for long hikes, exquisite meals, and exciting road trips. Thank you Tina, Dave, Logan, Brookelyn, and Deedy for holidays together and for keeping me grounded while I grind. To the memory of my aunt Bonnie, who is impossible to describe in just one sentence, thank you for love and laughter. Your presence is felt always. Thank you, Leila, for long conversations over tea and for all your wise observations about the world and about our family. Megan, James, Oskar & Gus: thank you and all your cats and dogs and horses and birds for being my refuge and my joy.

엄마 and Dad: thank you for worldwide adventures together and for riotous video calls the times we were apart. Thank you for being easily impressed by the

frivolous things I spend my time on. Thank you for forcing me to apply to college. Turns out I actually like school.

The first two years of my PhD were funded through teaching assistantships in the Linguistics department. Years 3-5 were funded by a research assistantship supported by U.S. National Library of Medicine Grant (R01LM011563), led by PI Trevor Cohen. My final quarter of support came from a Dissertation Fellowship from the University of Washington Linguistics Department.

COLOPHON

This thesis was typeset in *Charter* using pdflatex on Overleaf[1] and and a slightly customized version of the UW Thesis template[2]. It was written on a Mid 2014 13-inch MacBook Pro running OS Sierra version 10.12.6 — by many accounts, and certainly by my father's, a rather outdated setup. Most of the tables and figures were made with the support of Tables Generator[3] and Diagrams.net[4].

I wrote and revised this thesis at tables and on couches in a handful of rooms in houses and apartments in the area known as Seattle, Washington, USA, on land that has been stewarded by the Duwamish and other Coast Salish peoples for over 10,000 years. Virtually all (no pun intended) of the collaborative work and mentorship reflected in this thesis was conducted remotely via the Internet, through video calls, collaborative typesetting software, lots of emails, and the occasional Tweet.

---

[1] `https://www.overleaf.com/`

[2] Accessed in 2021 at `https://github.com/UWIT-IAM/UWThesis`

[3] `https://www.tablesgenerator.com/latex_tables`

[4] `https://app.diagrams.net/`

Chapter 1

INTRODUCTION

> "What needs to be emphasized is that technologies are developed and used within a particular social, economic, and political context. They arise out of a social structure, they are grafted on to it, and they may reinforce it or destroy it, often in ways that are neither foreseen nor foreseeable."
>
> — Ursula Franklin, *The Real World of Technology*[1]

As human society has become progressively interdependent, techniques for gathering and processing information have become essential to communication and coordination. The increased digitization of our behaviors and social relations has powered new industries that aim to not only capture this data, but also make predictions based on it. Data-driven technologies have become ubiquitous, and are thus implicated more and more in our daily routines. We use data-driven technologies to find places to eat, papers to read, songs to listen to, and answers to medical questions, among many other applications. However, these very systems have been implicated in perpetuating social inequities (O'neil, 2016; Eubanks, 2018; Noble, 2018; Benjamin, 2019). Particularly as automated data-driven systems grow in scale and influence, it is critical to examine the social impact of these systems, as well as the role that the development of these systems has played in shaping their impact.

The digital processing of text and speech has become an integral part of life in the so-called 'information age'. Recent successes in natural language processing (NLP) have relied, in part, on the availability of large quantities of data. The 'big data' revolution was hailed as 'the end of theory' (Anderson, 2008) but this claim ignores the fact that a radically

[1] Franklin (1999, p. 51)

empiricist paradigm does have epistemological commitments, or in other words, 'big data is theory' (Crawford et al., 2014). While the field of NLP has long been characterized by debates between rationalist and empiricist approaches (Gold, 2011), the latest turn toward 'big data' seems to be capitalizing on the availability of relatively cheap-to-acquire data at a massive scale.

The 'big data' paradigm in NLP has, while yielding promising advances, also produced negative externalities that must be addressed as the field's impact grows (Bender et al., 2021). Additionally, it is conspicuous that some of the biggest corporate sponsors of NLP, such as Google and Facebook, are monetized through the extraction and repackaging of human behavioral data at a massive scale, and thus stand to truly gain from the epistemological commitment of 'big data.' Particularly in light of community concern about the social impact of NLP, it is important to critically investigate NLP's impact more concretely, taking stock of the full breadth of NLP's various stakeholders and real-world applications (Blodgett et al., 2020). Researchers have also proposed that a *sociotechnical* lens is necessary for understanding and anticipating the breadth of benefits, harms, and likely uses of technologies (Selbst et al., 2019).

In this thesis, I argue that in order to assess the social impact of NLP systems, we need to critically examine the full development pipeline for NLP systems as well as the real-world contexts in which these systems are created and embedded. I advocate for an analysis that is attentive to the values that shape decisions about what applications to build and where these applications are used. I echo Winner (1980) and Franklin (1999) in emphasizing that technology does not exist in a vacuum; it is inherently situated within a social, political, economic context at every stage from conception to development to deployment. Thus, in this work, I apply a sociotechnical lens to the study of NLP systems — that is, going beyond an understanding of NLP systems as comprised merely of source code and data, and analyzing their impact holistically, looking at the human practices implicated in each step of system design and development. Following Agre (1997)'s call to keep 'one foot planted in the craft work of design and the other foot planted in the reflexive work of

critique,' I attempt to use my own experiences as a practitioner to explore and reflect on the challenges in building and evaluating NLP systems in a socially conscious manner.

In the remaining sections, I lay out the research questions that guide the present work; I follow by sketching out the current sociotechnical ecosystem and previewing some of the social impacts of NLP. Then, I motivate a pipeline-based analysis of NLP systems and provide an overview of the steps in this pipeline. Finally, I provide an overview of the thesis.

1.1 Research questions

In this work, we seek to answer the following questions:

- What has been the social impact of data-driven natural language processing technologies?
- What role does the design and development of these systems play in shaping that impact?
- What considerations should be made during the development process to anticipate and address this impact?

1.2 Background

1.2.1 "A new social order"

The last few decades have seen a proliferation of technologies that are predicated on the collection of massive amounts of personal data (Kitchin, 2014; Sadowski, 2020). Much of this work has been exploitative, non-consensual, not participatory, and comes at a great cost to the people whose data is being collected to power the information economy. Indeed Katz (2017) argues that the latest wave of so-called artificial intelligence is merely a rebranding of such extractive technologies.

Institutions both public and private have long engaged in the project of collecting information from and about people for surveillance and management (Scott, 2008). However, the spread of information and communication technologies (ICTs), most prominently personal computers and smartphones, has facilitated the collection of massive amounts of personal data in continuous, real-time streams (Kitchin, 2014), and this data is increasingly commodified (Fourcade and Healy, 2017). Entrepreneur Andrew Ng makes this point rather bluntly: 'At large companies, sometimes we launch products not for the revenue, but for the data. We actually do that quite often... and we monetize the data through a different product' (quoted in Sadowski (2020), p. 30). The insidious creep of 'smart' sensing technology into the public and private sphere has drawn people into a data collection network in which they are often unaware of the extent to which they are being surveilled. Couldry and Mejias characterize this pervasive datafication of daily life as 'nothing less than a new social order, based on continuous tracking, and offering unprecedented new opportunities for social discrimination and behavioral influence (Couldry and Mejias, 2019, p. 1).'

Meanwhile, recent disciplinary shifts in NLP are marked by a return toward an empiricist, data-driven paradigm, both predicated on and justified by the availability of large text collections. As natural language processing (NLP) technologies, which have a primary goal of extracting information from unstructured linguistic data, have seen a surge in popularity and impact in recent decades, it is critical to ask of real-world NLP technologies: Who is looking at language data? What are they looking for, and why?

1.2.2 The social impact of natural language processing

Language technologies have transformed our experience of the world and of each other. Some of these technologies have become so standard as to seem invisible or to be second nature. Over the course of a single day in 2021, I have used web-based search engines to find information, have seen social media posts that were subject to automatic translation

and moderation, and have composed SMS messages in an interface that suggests completions to my sentences. Each of these technologies involves natural language processing. However, many of these technologies do not work equally well for everyone (Tatman and Kasten, 2017), and have been shown to promote harmful biases against women and people of color (Noble, 2018). Toxicity classifiers have been shown to be biased against speech by people of color and by queer people (e.g. Oliva et al., 2021), and machine translation has been shown to produce slurs in translated text where the original text contained none (Moorman, 2014). Particularly as NLP technologies become more widespread, researchers have called for a more thorough consideration of the social impact of natural language processing (Hovy and Spruit, 2016). Recent years have also seen prominent NLP conferences, such as EMNLP 2020, NAACL 2021, and ACL-IJCNLP 2021 instantiate a requirement for authors to publish ethical impact statements in their research publications.

A lot of the 'blame' for the negative impacts of NLP technologies has tended to fall on the data, claiming that if we only had better datasets or de-biased models, we could mitigate these issues. However, limiting our focus to data is misguided — we should also take into account broader questions about what why we have collected that data and what we intend to use it for, as well as other pieces of the pipeline. Rather than locating the potential negative impacts of NLP systems solely within representational issues in data, there are growing calls to question more critically and more broadly how sociotechnical systems predicated on data shift power (Blodgett et al., 2020; Denton et al., 2020).

1.3 Stages of development for machine learning systems

As described earlier, data-gathering technologies are increasingly pervasive and are reorganizing our social relations. Many scholars have urged that an understanding of the impact of technology should involve reflecting on the *design* of these systems, and the values and assumptions invoked in that process (Sengers et al., 2005). Recent work has also proposed a framework for identifying sources of harm brought about at different

stages in the machine learning life cycle (Suresh and Guttag, 2021).

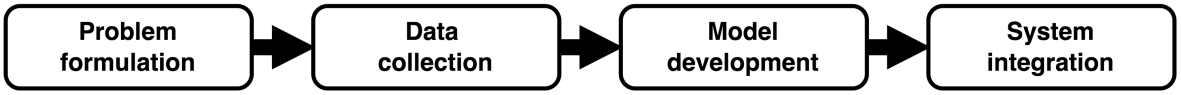


Figure 1.1: Simplified pipeline for a data-driven machine learning system

There are a variety of ways of breaking the development of machine learning systems into stages. Wagstaff (2012) refers to 'necessary preparation,' 'machine learning contribution,' and 'impact'. Sim et al. (2021) refer to the 'before' 'during' and 'after' phases of a project, with different considerations for each. In general, the process can be characterized as a cycle of formulating a question, gathering or soliciting data, developing a model, and finally releasing the resulting artifact. Following these frameworks and leveraging my own experience as a practitioner, I delineate phases of development into the following: problem formulation, data collection phases, model building phase, and system integration[2] (Figure 1.1). In practice, these phases may occur repeatedly or out of order. For example, problems may be re-formulated based on limitations of available data, or as a system collects data during its course of operation, the model parameters may be revised.

I elaborate on the stages here, using the following running example to illustrate each stage:

Person A runs a popular blog. Recently, she has noticed a surge in misogynistic comments being posted anonymously to her blog. Her blog's followers have expressed that this is distressing to them, and has led them to feel less inclined to participate in discussions on the blog. Person A would like to consider an automated solution to this problem.

[2]Here, I use Madeline Elish's term 'integrate' rather than the often-used 'deployment', to emphasize the importance of situating work in a context and weaving it into an existing environment, rather than simply releasing it without such considerations (Brown, 2021).

1.3.1 Problem formulation

Building a system tends to be motivated by some problem that people want to solve. In this stage, one formulates a question and considers previous approaches to the problem, including those that predate technological solutions. In the case of machine learning, this also entails framing the question as a machine learning problem, i.e. one that consists of an input space and an output space, and a mapping between each. This stage is important, as clarifying the goals of the system "...can help uncover latent biases in your mental model of what kind of people there are in the world and how you believe they move through it (Schnoebelen, 2017)."

Person A decides to build a classifier for detecting misogynistic comments on her blog. She chooses to formulate this as a binary classification problem, i.e., given a particular comment, determine whether it is misogynistic or not. Note that there are non-machine-learning interventions available as well – she can disallow comments entirely, or she can prevent anonymous comments, but she prefers to allow her audience to comment without these restrictions.

1.3.2 Data collection

Data-driven machine learning systems necessarily rely on collections of data for training and evaluation. A system designer must decide what data they need, where to get the data, and how to acquire it. They should ensure that it is sourced legally and that care is taken to not violate rights of human subjects. When existing data is not available, the designer may need to solicit annotations from other people. We elaborate on this more in Ch. 3.

Person A is aware that many datasets exist for training classifiers to address the problem of 'toxicity,' but she is particularly interested in the issue of detecting misogyny in blog comments, since this is the issue she is facing. Because she is the blog owner, she feels that she can use the comments directly as training data. Should she inform her followers that this is the case? Will she label the comments herself, or will she enlist the help of someone else? In that case, how

can she prepare others for the task of labeling alarming data? What counts as misogynistic, and what will she do about tricky edge cases?

1.3.3 Model development

In this stage, system builders select an architecture and evaluation metrics, and use the data collected in the previous stage. In contemporary machine learning practice, the modeling stage arguably receives lots of attention, while data work and downstream impact considerations are undervalued (Wagstaff, 2012; Heinzerling, 2019; Sambasivan et al., 2021).

Person A does not have access to a lot of compute power, so she decides to use a simple logistic regression classifier and a bag-of-words model of text. She is aware that this system may not be sensitive to nuances in text, including sarcasm or distinctions between use and mention of possibly toxic terms. Should Person A favor a system that has a high false negative rate (thereby increasing the risk that a truly bad comment leaks through) or a high false positive rate (thereby inadvertently silencing some commenters)?

1.3.4 System integration

After a model has been designed, built, and tested, it may be integrated into a broader user-facing system. Users may or may not choose to adopt the system, and in some cases, users might not even be aware an NLP system is being applied to their data (e.g. social media widgets that translate user posts).

Person A still has to decide what actions will result from the use of her classification system for blog comments. When will classifications take place — before the user posts the comment, or after? Will every comment be subject to classification, or only those that are flagged by other users? What will happen to comments that are classified as 'misogynistic' — will they be automatically deleted from her blog, or will they be demoted to the bottom of the page and/or given a content warning for other users?

1.4 Thesis Overview

As our toy example of *Person A* has shown, there are a variety of considerations at each step in the system development pipeline, some of which may be unique to her particular use case, and some of which may generalize to a broader array of concerns in NLP system development. We explore some of these concerns more in-depth with three case studies.

First, we provide some definitions for useful concepts and an overview of current issues in the stages of development in Chapter 2. In Chapter 3, we focus our attention on the data collection stage, and explore some of the limitations of prevailing approaches to data set development and use in the broader machine learning community, which has informed practices in natural language processing.

The next few chapters present the three case studies. In Chapter 4, I give a brief overview of vector space models of text, highlighting key affordances of these models and considering the implications of the training data used in their develompent. In Chapter 5, I examine machine translation as a case study in the tensions involved in reconciling the notion of 'socially good' applications of natural language processing technology with troubling trends in the use of such technology in high stakes, often unjust scenarios. I note how the history of this task, from its initial conception and the primary theoretical assumptions involved, has implications for the societal impact of the development and deployment of machine translation technology. Chapters 6 and 7 document the development and validation of a system for representing biomedical relationships for practical applications such as identifying similar concepts and proposing hypotheses to connect concepts in previously unknown ways. Chapter 8 provides some reflection and analysis of the case studies, and Chapter 9 concludes the work.

Portions of the work in Chapters 2, 3, and 8 appeared in the publication 'Data and its (dis)contents: a survey of dataset development and use in machine learning,' which first appeared at the NeurIPS 2020 Workshop on Machine Learning Retrospectives, Surveys, and Meta-analyses (ML-RSA 2020) and is now published as a journal article (Paullada

et al., 2021) in *Patterns*, written in collaboration with Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna.

Versions of Chapter 5 appeared as a publication at the 2020 Resistance AI workshop and as an article on The Gradient (Paullada, 2021).

Portions of the work in Chapters 4, 6, and 7 appeared in the publication 'Improving Biomedical Relationship Retrieval with Structural Dependencies' (Paullada et al., 2020), co-authored with Trevor Cohen and Bethany Percha, presented at the 2020 Workshop on Biomedical Natural Language Processing (BioNLP).

A Note on Language: I use 'I' and 'we' interchangeably in this thesis. I use 'we' in particular when describing work that has been the product of a collaboration or, at times, to refer more broadly to the community of scholars I am part of, and 'I' when describing work done individually, including the contributions of this thesis overall.

Chapter 2

BACKGROUND

In this chapter, we define some key concepts for data-related principles. Then, we provide some background on current trends and issues in the stages of machine learning system development, and some proposals for addressing these issues. We focus in particular on natural language processing.

2.1 Key concepts

Here, I provide brief definitions for concepts that will be useful in situating some of the issues I discuss and my suggestions.

2.1.1 Benchmarks, tasks, capabilities, and datasets

We follow Schlangen (2021) in distinguishing between *benchmarks*, *tasks*, *capabilities*, and *datasets*. While his work focused on natural language processing, we broaden these definitions to include aspects of other machine learning applications. In this context, a *task* is constituted of an input space and output space and an expected mapping between them. Schlangen notes that there are typically both *intensional* and *extensional* descriptions of tasks. An intensional description corresponds to a high-level, theoretical relationship between input and output (e.g. 'translation' from a string in one language to a string in another language), where an extensional realization is comprised of the set of input-output pairs (i.e., actual pairs of strings). Thus tasks are exemplified by *datasets*, i.e. sets of input-output pairs that conform, if valid, to the intensional definition of the task. Schlangen illustrates this relationship with the following informal example: a set of images of giraffes paired with sentences describing each image is not a valid exemplification of the task of

image description, but could be valid for the narrower task of *giraffe image description*.

Schlangen further notes that *benchmark tasks*, as they are currently conceived, are meant to test particular cognitive *capabilities* of interest. For example, a *language benchmark* may involve tasks, instantiated as datasets, such that solving a particular task requires capabilities believed to belong to a competent language user, e.g. the ability to disambiguate words in context or infer entailment relationships between sentences. Whether a given cognitive task T, involving capability C, can be validly exemplified by a given dataset D, and whether a system that performs well at dataset D can be said to be doing well at task T (and thus models capability C), is a topic of active debate in the variety of disciplines that machine learning has intervened in.[1]

In referring to dataset exemplars that pair instances (input) and labels (output), we follow a convention from machine learning of referring to the latter as *target labels*, which are those labels that are used as the learning target, and which have typically been produced by human annotators or, in some cases, automated labeling heuristics. These are also often referred to in the literature as 'gold standard' or 'ground truth' labels, but we wish to emphasize their role as training targets that are neither objective nor necessarily representative of reality.

2.1.2 Licenses and legal concepts for data use

Here, we provide definitions for some legal concepts that are relevant to the production and proliferation of digital media: fair use, Creative Commons, Free Cultural Works, and public domain works.

In the United States legal context, fair use is a legal doctrine that allows for the unlicensed use of copyrighted materials in certain circumstances, and is meant to promote "freedom of expression." Copyrighted materials, such as images, video, or audio, can be

[1]For perspectives on this debate in the area of natural language understanding, refer to Bender and Koller (2020) and Michael (2020).

legally used without permission from the copyright owner for research or scholarly purposes, for parody, among other uses. Legal scholars have advocated that the use of copyrighted material for text and data mining should be recognized as fair use, arguing that the potential social benefits would be hampered by strict copyright laws (Sag, 2019).

The Creative Commons license model was created to enable copyright holders to allow permissive uses of their content. While Fair Use is determined on a case-by-case basis, Creative Commons is intended to be a simple model for creators to share their content with some restrictions, ranging from disallowing commercial use (CC BY NC) to requesting that any derivative work also be made available under the same license (CC BY ShareAlike). The Free Cultural Works movement promotes the radical reuse of media with no restrictions, and provides definitions both for Free Cultural Works and Free Cultural licenses that enable this kind of use. A list of such licenses is available at `https://freedomdefined.org/Licenses`.

Works in the public domain have no copyright. These include works whose copyright has expired. Because of their unrestricted use limitations, public domain resources such as the Gutenberg Corpus and the ENRON corpus have been a popular source of linguistic data.

2.2 Current approaches to machine learning development

Here we summarize some of the issues present in the phases of machine learning system development.

2.2.1 Problem formulation

A popular text book in statistical natural language processing notes in the preface: 'Increasingly, businesses, government agencies and individuals are confronted with large amounts of text that are critical for working and living, but not well enough understood to get the enormous value out of them that they potentially hide...' (Manning and Schutze, 1999).

Indeed, it has been these first two entities who have shaped the core scientific questions, research practices, and overall impact of NLP. Some of the most ardent sponsors of natural language processing research in the United States have a clear incentive to amass large amounts of data for further processing.

At a finer grained level, the conventional NLP pipeline consists of structured prediction tasks like parsing and part of speech tagging, but there are also so-called 'natural language understanding' applications which can be broken down into a variety of other tasks. Modern benchmarks like GLUE and DynaBench are composed of datasets constructed such that they realize some task of interest, such as question answering or hatespeech detection. Earlier, I mentioned the pervasive nature of government and industry in shaping the goals of NLP. Many of these tasks embody practical problems for these organizations — hatespeech detection only exists because social media platforms wish to moderate, at scale, the content that users post online. This is not to say that this is a poor choice of focus, but rather to call out the ways that a field's imagination is shaped by these key influences.

How does the problem formulation stage impact the ensuing stages? The problematization that guides decisions about what data to collect and how to label it can lead to the creation of datasets that formulate pseudoscientific, often unjust tasks. For example, several papers in recent years that attempt to predict attributes such as sexuality and other fluid, subjective personal traits from photos of human faces presuppose that these predictions are possible and worthwhile to make. However, these datasets enable a reliance on meaningless shortcuts that support the apparent 'learnability' of the personal traits in question. For example, an audit by Agüera y Arcas et al. (2018) found that a model trained to predict sexual orientation from images of human faces, harvested from online dating profiles, was really learning to spot stereotypical choices in grooming and self-expression, which are by no means universally correlated with homosexuality. Gelman et al. discuss how such a study strips away context and implies the existence of an "essential homosexual nature" (Gelman et al., 2018, p. 271). The task rests on a pseudoscientific essentialism of human traits; thus, the *intensional* task definition of 'determine sexuality (output) from

images of human faces (input)' has no valid *extensional* realization in the form of any dataset, because the underlying causal assumption for the task is unsound. Similar issues occur in datasets meant to provide a foundation for determining 'criminality' from faces (Agüera y Arcas et al., 2017). Not only are these task formulations problematic, but once sensitive data has been collected, it can be misused.

As the previous examples show, when machine learning models can leverage spurious cues to make predictions well enough to beat a baseline on the test data, which is typically drawn from the same distribution as the training data, the resulting systems can appear to legitimize spurious tasks that do not map to real world capabilities. The mapping between inputs and target labels contained in datasets is not always a meaningful one, and the ways in which data are collected and tasks are structured can lead models to rely on faulty heuristics for making predictions.

2.2.2 Dataset development and use

Datasets in machine learning have typically come from other fields (e.g. many of the datasets in the UCI Machine Learning Repository) or have resulted from scraping the web. Many resources are expert-crafted (e.g. Treebanks) but progressively more have been developed by soliciting labels from crowdworkers on platforms such as Amazon Mechanical Turk. We elaborate on the variety of issues pertaining to datasets, including legal issues, labor issues, and practical issues, in the following chapter. Understanding properties of the data can inform model selection, choice of metrics, and delineation of appropriate deployment contexts in the following stages.

2.2.3 Model development

Here we discuss current trends in model development and evaluation, largely focusing on the broader model development ecosystem.

As mentioned previously, the model development phase has earned the most attention

in machine learning, and tends to be the most incentivized area of focus (Wagstaff, 2012; Sambasivan et al., 2021). Leaderboard-driven development in machine learning is intended to incentivize competitive modeling and to faciliate comparisons across competing systems.

Benchmark datasets play a critical role in orienting the goals of machine learning communities and tracking progress within the field (Dotan and Milli, 2020; Denton et al., 2020). Yet, the near singular focus on improving benchmark metrics has been critiqued from a variety of perspectives in recent years. Natural language processing researchers have exhibited concern with the focus on accuracy on benchmark leaderboards, with several calls to include more comprehensive evaluations that include reports of energy consumption, model size, and hyperparameter tuning details in addition to standard top-line metrics (Ethayarajh and Jurafsky, 2020; Dodge et al., 2019; Schwartz et al., 2019; Strubell et al., 2019). From a fairness perspective, researchers have called for the inclusion of disaggregated evaluation metrics for relevant subgroups, in addition to standard top-line metrics, when reporting and documenting model performance (Mitchell et al., 2019). Sculley et al. (2018) examine the incentive structures that encourage singular focus on benchmark metrics — often at the expense of empirical rigor — and offer a range of suggestions including incentivizing detailed empirical evaluations, including negative results, and sharing additional experimental details.

The excitement surrounding leaderboards and challenges can also give rise to a misconstrual of what high performance on a benchmark actually entails. In response to the recent onslaught of publications misrepresenting the capabilities of BERT language models, Bender and Koller (2020) encourage natural language processing researchers to be attentive to the limitations of tasks and include error analysis in addition to standard performance metrics.

NLP practice is often focused a on evaluating systems during the model development stage, particularly the use of benchmarks and metrics. However, models that score well on such evaluations may prove to be unusable to real-world users (Heuer and Buschek,

2021). Failing to ask 'who needs or wants this' before building a system can cause friction during the deployment stage.

2.2.4 System integration

Recent years have seen growing calls to consider the impact of a machine learning artifact once it has been implemented and publicized. Recently, NLP conference submission guidelines have begun to incorporate ethical reflections into research papers, taking cues from the ACM Future of Computing Academy's proposal to adapt the peer review process to include broader considerations of research impact (Hecht et al., 2018). It is urgent to prompt researchers and practitioners to consider the impact of their work in particular contexts as well.

Frameworks for promoting considerations of impact in the machine learning field are still under development, and there are some clear shortcomings to the current approach. A recent survey of newly implemented NeurIPS impact statements found that the imagined stakeholders were narrow in scope, often overlooking the most vulnerable stakeholders from their analysis, and imagining potential 'benefits' of a system most often as those accrued by companies and governments (Boyarskaya et al., 2020). The imagined harms included impediments to deployment or adoption, or mass disasters, and many statements involved boilerplate language. As such, the authors suggest 'context-aware frameworks of harm' for better anticipation of potential harms.

2.3 Summary

We have introduced some key concepts that we will return to throughout the thesis, as well as elaborated on some current concerns with the stages of development. In the following chapter, we analyze issues with dataset collection and usage practices in more depth.

Chapter 3

DATA-CENTRIC ISSUES AND PROPOSALS

> "Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings."
>
> — Mimi Ọnụọha, 'The Point of Collection'[1]

> "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."
>
> — Geoffrey Bowker, 'Memory Practices in the Sciences'[2]

In this chapter, we illustrate issues endemic to how datasets are produced, disseminated, and used in the field of machine learning. We focus in particular on the areas of computer vision and natural language processing. We summarize concerns relating to the design, collection, maintenance, distribution, and use of machine learning datasets as well as broader disciplinary norms and cultures that pervade the field.

3.1 Background

Datasets form the basis for training, evaluating and benchmarking machine learning models. As a result, they have played a foundational role in the advancement of the field. Furthermore, the ways in which we collect, construct and share these datasets inform the kinds of problems the field pursues and the methods explored in algorithm development.

[1] Ọnụọha (2016, §5)

[2] Bowker (2005, p. 184)

Datasets have been seen as the limiting factor for algorithmic development and scientific progress (Halevy et al., 2009; Sun et al., 2017), and a select few benchmark datasets, such as the ImageNet benchmark for visual object recognition (Deng et al., 2009) and the GLUE benchmark for English textual understanding (Wang et al., 2019), have been the foundation for some of the most significant developments in the field. Benchmark datasets have also played a critical role in orienting the goals, values, and research agendas of the machine learning community (Dotan and Milli, 2020). In recent years, machine learning systems have been reported to achieve 'super-human' performance when evaluated on such benchmark datasets. However, recent work from a variety of perspectives has surfaced not only the shortcomings of some machine learning datasets as meaningful tests of human-like reasoning ability, but also the troubling realities of the societal impact of how these datasets are developed and used. Together, these insights reveal how this apparent progress may rest on faulty foundations.

As the machine learning field turned to approaches with larger data requirements, the sort of skilled and methodical annotation applied in dataset collection practices in earlier eras was spurned as 'slow and expensive to acquire', and a turn toward unfettered collection of increasingly large amounts of data from the Web, alongside increased reliance on crowdworkers, was seen as a boon to machine learning (Halevy et al., 2009; Deng et al., 2009). Enormous scale has been mythologized as beneficial to generality and objectivity, but all datasets have limitations and biases (boyd and Crawford, 2012). In particular, prevailing data practices tend to abstract away the human labor, subjective judgments and biases, and contingent contexts involved in dataset production. However, such details are important for assessing whether and how a dataset might be useful for a particular application, for enabling more rigorous error analysis, and for acknowledging the significant difficulty required in constructing useful datasets.

The machine learning field has placed large scale datasets at the center of model development and evaluation. As systems trained in this way are deployed in real-world contexts that affect the lives and livelihoods of real people, it is essential that researchers, advocacy

groups and the public at large understand both the contents of the datasets and how they affect system performance. In particular, as the field has focused on benchmarks as the primary tool for both measuring and driving research progress (Schlangen, 2021), understanding what these benchmarks measure (and how well) is urgently necessary.

In the following sections, we provide an overview for some of the issues in dataset practices that have been surfaced recently, as well as some of the proposals that have been made to address these issues. We conclude with some key takeaways for the full system development pipeline.

3.2 Issue: Datasets are hastily collected and insufficiently documented

A host of concerns regarding the practices of dataset collection, annotation, and documentation have been raised in recent years. In combination, these concerns reflect a pervasive undervaluation of data work (Sambasivan et al., 2021) and what Jo and Gebru (2020) describe as a *laissez-faire* attitude regarding dataset development. Rather than collecting and curating datasets with care and intentionality — as is more typical in other data-centric disciplines — machine learning practitioners often adopt an approach where anything goes. As one data scientist put it, "if it is available to us, we ingest it" (Holstein et al., 2019).

The common practices of scraping data from internet search engines, social media platforms, and other publicly available online sources faced significant backlash in recent years. For example, facial analysis datasets have received push-back due to the inclusion of personal Flickr photos without data subject's knowledge (Solon, 2019).

Dataset annotation practices have also come under scrutiny within recent years. Much of this has focused on how subjective values, judgments, and biases of annotators contribute to undesirable or unintended dataset bias (van Miltenburg, 2016; Misra et al., 2016; Ghai et al., 2020; Hube et al., 2019; Sap et al., 2019). More generally, several researchers have identified a widespread failure to recognize annotation work as *interpretive*

work, which in turn can result in a conflation of *target* labels in a collected dataset and *real-world* objects, for which there may be no single 'ground truth' label (Miceli et al., 2020; Aroyo and Welty, 2015). As discussed further in Section 3.4, data annotation tasks are often mediated through crowdwork platforms such as Amazon Mechanical Turk (AMT). These platforms, by design, position annotators as interchangeable workers, rather than individuals who bring to bear their own subjective experiences and interpretations to the task. Divergences in judgements across different annotator pools (Geva et al., 2019) as well as between AMT annotators and other communities (Sen et al., 2015) has been empirically explored.

Recent work has revealed inconsistencies and biases in the data that hints at larger annotation patterns that mischaracterize the real world tasks these datasets are meant to abstractly represent, and the broader impact of data curation design choices in determining the quality of the final dataset. For example, Tsipras et al. (2020) found that the annotation pipeline for ImageNet does not reflect the intention of its development for the purpose of object recognition in images. They note that ImageNet, constructed with the constraint of a single label per image, had its labels largely determined by crowdworkers indicating the visual presence of that object in the image. This has led to issues with how labels are applied, particularly to images with multiple objects, where the class of interest could include a background or obscured object that would be an unsuitable result for the image classification task of that particular photo. Furthermore, the nature of image retrieval for the annotation tasks biases the crowdworkers' response to the labeling prompt, making them much less effective at filtering out unsuitable examples for a class category.

A recent survey of 164 publications involving machine learning applications that leverage Twitter data found that details about the data collection and labeling process were heavily under-specified and inconsistent (Geiger et al., 2020). Scheuerman et al. (2020) found a widespread under-specification of annotation processes relating to gender and racial categories within facial analysis datasets.

The lack of rigorous and standardized dataset documentation practices has contributed

to reproducibility concerns. For example, recent work undertook the laborious task of reconstructing ImageNet, following the original documented dataset construction process in an effort to test the generalization capabilities of ImageNet classifiers (Recht et al., 2019). Despite mirroring the original collection and annotation methods — including leveraging images from the same time period — the newly constructed dataset was found to have different distributional properties. The differences were largely localized to variations in constructing target labels from multiple annotations. More specifically, different thresholds for inter-annotator agreement were found to produce vastly different datasets, indicating that so-called 'ground truth' labels in datasets do not correspond to truth.

3.2.1 Proposal: More rigorous documentation and reflection practices

Several dataset documentation and development frameworks have been proposed in recent years in an effort to address these concerns, with certain frameworks looking to not just capture characteristics of the output dataset but also report details of the procedure of dataset creation for better transparency and accountability (Bender and Friedman, 2018; Holland et al., 2018; Chmielinski et al., 2020; Hutchinson et al., 2021; Gebru et al., 2021). Denton et al. (2020) propose a research agenda in the 'data genealogy' paradigm that promotes critical assessment of the design choices with respect to the data sources, theoretical motivations, and methods used for constructing datasets. Prospective accounting for dataset contents using some of the documentation methods previously discussed can offset the potential of post-hoc documentation debt that can be incurred otherwise.

3.3 Issue: Datasets pose intellectual property concerns

Benchmark datasets are often mined from the internet, collecting data instances which have various levels of licensing attached and storing them into a single repository. Dif-

ferent legal issues arise at each stage in the data processing pipeline, from collection to annotation, from training to evaluation, from inference and the reuse of downstream representations such as word embeddings and convolutional features (Benjamin et al., 2019).

Benchmark datasets are drawn from a number of different sources, each with a different configuration of copyright holders and permissions for their use in training and evaluation in machine learning models. For instance, ImageNet was collected through several image search engines where licensing/copyright restrictions on data instances in those images are unknown (Russakovsky et al., 2015). The ImageNet project does not host the images on their website, and therefore sidestep the copyright question by claiming that they operate like a search engine (Levendowski, 2018, ftn. 36). PASCAL VOC was collected via the Flickr API, meaning that the images were all held through the Creative Commons license (Everingham et al., 2010). Open licenses like Creative Commons allow for training of machine learning models under fair use doctrine (Merkley, 2019). Faces in the Wild and Labeled Faces in the Wild were collected through Yahoo News, and via an investigation of the captions on the images we can see that the major copyright holders of those images are news wire services, including the Associated Press and Reuters (Berg et al., 2004). Other datasets are collected in a studio environment, where images were taken by dataset curators and therefore are copyright holders, which avoids potential copyright issues.

US copyright law is not well-suited to cover the range of uses of benchmark datasets, and there is limited case law establishing precedent in this area. Legal scholars have defended the use of copyrighted material for data science and machine learning by suggesting that this material's usage is protected by fair use, since it entails the non-expressive use of expressive materials (Sag, 2019).

3.3.1 Proposal: Better licensing

The machine learning and AI research communities have responded to this crisis by attempting to outline alternatives to licensing which make sense for research and bench-

marking practices more broadly. The Montreal Data License[3] outlines different contingencies for a particular dataset, including whether the dataset will be used in commercial versus non-commercial settings, whether representations will be generated from the dataset, whether users can annotate the label or use subsets of it, and more (Benjamin et al., 2019). This is a step forward in clarifying the different ways in which the dataset can be used once it has been collected, and therefore is a clear boon for AI researchers who create their own data instances, such as photos developed in a studio or text or captions written by crowdworkers. However, this does not deal with the larger issue of the copyright status of data instances scraped from the web, nor the privacy implications of those data instances.

3.4 Issue: Data labor is under-valued

As the machine learning community has turned to the cheap and scalable work forces offered by crowd sourcing platforms, there has been growing concern regarding the working conditions of those laboring on machine learning datasets. Data annotation is often cast as unskilled work —work *anyone* can perform — which in turn contributes to a dehumanizing and alienating work experience. For example, Irani (2015a) describes how crowd work platforms, such as Amazon Mechanical Turk, create a hierarchy of data labor, positioning crowd work as menial work relative to the innovative work of those leading dataset development. Miceli et al. (2020) discuss how, in commercial data annotation companies, power asymmetries and company hierarchies affect the work output of data annotation teams.

Framing data annotation as unskilled work frames crowd workers as essentially interchangeable, and creates the infrastructural conditions of precarity and invisibility (Suchman, 1995; Star and Strauss, 1999; Precarity Lab, 2020). For example, crowd-sourced data annotation is typically mediated through digital interfaces that distance the crowd workers from not only the dataset commissioners who manage the annotation tasks, but

[3] https://montrealdatalicense.com/

also their fellow workers, thereby rendering the workers and the labor concerns they might face invisible (Irani and Silberman, 2013; Irani, 2015b). Such concerns include low and unstable wages, unfair treatment by task requesters, and barriers to worker solidarity and collective action (Berg, 2016; Semuels, 2018; Gray and Suri, 2019).

3.4.1 Proposal: Build worker solidarity

In response to these growing concerns, guidelines and tools for task creators have been developed to help facilitate fair pay (Silberman et al., 2018; Whiting et al., 2019) and interventions oriented at crowd workers directly have been developed to support worker solidarity (Irani and Silberman, 2013; Salehi et al., 2015) and fair pay (Callison-Burch, 2014). Gray and Suri (2019) also discuss corporate interventions, such as providing collaborative online discussion spaces, offline shared workspaces, and portable reputation systems, as well as governmental responses such as the construction of worker guilds, unions, and platform cooperatives, and the provision of a social safety net work for these precarious workers.

As personal data is increasingly commodified by technology companies and harvested at scale to improve proprietary machine learning systems, often in ways that are by turns inscrutable or distasteful to the general public (Viljoen, 2020), recent proposals call for not only reframing personal data as labor (Posner and Weyl, 2019), but also for 'data strikes' in which users collectively withhold their data as a means to shift the power imbalance back towards subjects who are not compensated for the ambient collection of their data (Vincent et al., 2019).

3.5 Issue: Datasets contain harmful contents

In recent years there has been growing concern regarding the degree and manner of representation of different demographic groups within prominent machine learning datasets, constituting what have been called *representational harms* (Crawford, 2017). For example,

a glaring under-representation of darker skinned subjects, compared with lighter skinned subjects, has been identified within prominent facial analysis datasets (Buolamwini and Gebru, 2018) and in image datasets used to train self-driving cars to detect pedestrians (Wilson et al., 2019). Meanwhile, the images in object recognition datasets have been overwhelmingly sourced from Western countries (DeVries et al., 2019). NLP datasets have faced similar issues. For example, Zhao et al. (2018) found a stark underrepresentation of female pronouns in the commonly used OntoNotes dataset for English coreference resolution; similarly, Lennon (2020) found that feminine-coded names were vastly underrepresented in the CoNLL-2003 dataset used for named entity recognition. While the underrepresentation of marginalized groups in datasets has been met with calls for 'inclusion', Hoffmann (2020) provides a case for skepticism of this narrative, as it has the potential to merely uphold the very sort of power hierarchy that engenders such underrepresentation in the first place.

Stereotype-aligned correlations have also been identified in both computer vision and natural language processing datasets. For example, correlations between gender and activities depicted in computer vision datasets have been shown to reflect common gender stereotypes (Zhao et al., 2017; Burns et al., 2018; van Miltenburg, 2016). Dixon et al. (2018) found that a dataset for toxicity classification contained a disproportionate association between words describing queer identities and text labeled as 'toxic', while Park et al. (2018) found evidence of gender bias against women in similar datasets. Such disparities in representation stem, in part, from the fact that particular, non-neutral viewpoints are routinely implicitly invoked in the design of tasks and labeling heuristics. For example, a survey of literature on computer vision systems for detecting pornography found that the task is largely framed around detecting the features of thin, nude, female-presenting bodies with little body hair, largely to the exclusion of other kinds of bodies — thereby implicitly assuming a relatively narrow and conservative view of pornography that happens to align with a straight male gaze (Gehl et al., 2017).

In an examination of the person categories within the ImageNet dataset (Deng et al.,

2009), Crawford and Paglen (2019) uncovered millions of images of people that had been labelled with offensive categories, including racial slurs and derogatory phrases. In a similar vein, Birhane and Prabhu (2021) examined a broader swath of image classification datasets that were constructed using the same categorical schema as ImageNet, finding a range of harmful and problematic representations, including non-consensual and pornographic imagery of women. In response to the work of Crawford and Paglen (2019), a large portion of the ImageNet dataset has been removed (Yang et al., 2020). Similarly, Birhane and Prabhu (2021)'s examination prompted the complete removal of the TinyImages dataset (Torralba et al., 2008).

3.5.1 Proposal: Inspect datasets during and after production

Birhane and Prabhu (2021), summarized above, and Pipkin (2020) show how meticulous manual audits of large datasets are compelling ways to discover the most surprising and disturbing contents therein. Pipkin spent hundreds of hours watching the entirety of MIT's 'Moments in Time' video dataset (Monfort et al., 2019), finding shocking and unexpected footage of violence, assault, and death. They provocatively point out, in a reflection on the process of developing their artistic intervention *Lacework*, that the researchers who commission the curation of massive datasets may have less intimate familiarity with the contents of these datasets than those who are paid to look at and label individual instances, and as we discuss in §3.4, there is growing awareness of the need to better support the workers at the front lines of the often grim and undervalued work of data labeling.

3.6 Issue: Datasets harbor spurious cues that lead to brittle models

While deep learning models have seemed to achieve remarkable performance on challenging tasks in artificial intelligence, recent work has illustrated how these performance

gains may be due largely to 'cheap tricks'[4] rather than human-like reasoning capabilities. Geirhos et al. (2020) illustrate how the performance of deep neural networks can rely on *shortcuts*, or decision rules that do not extrapolate well to out-of-distribution data and are often based on incidental associations. Oftentimes, these shortcuts arise due to artifacts in datasets that allow models to overfit to training data and to rely on nonsensical heuristics to 'solve' the task — for example, detecting the presence of pneumonia in chest X-ray scans based on hospital-specific tokens that appear in the images (Geirhos et al., 2020). That is, in spite of high predictive performance, models are not performing the task according to its *intensional* description, and thus the datasets may not be exemplary of reasoning *capabilities*, as we discussed in Section 2.1.1.

Recent work has revealed the presence of shortcuts in commonly used datasets that had been conceived of as proving grounds for particular competencies, such as reading comprehension and other 'language understanding' capabilities. Experiments that illuminate data artifacts, or 'dataset ablations' as Heinzerling (2019) calls them, involve simple or nonsensical baselines such as training models on incomplete inputs and comparing performance to models trained on full inputs. Much recent work in NLP has revealed how these simple baselines are competitive, and that models trained on incomplete inputs for argument reasoning, natural language inference, fact verification, and reading comprehension — i.e., tasks structured such that no human could do much more than randomly guess the correct output — perform quite well (Niven and Kao, 2019; Schuster et al., 2019; Gururangan et al., 2018; Poliak et al., 2018; Kaushik and Lipton, 2018).[5] In many cases, this work has revealed how an over-representation of simple linguistic patterns (like negation or presence of certain words) in dataset instances belonging to one label class can serve as a spurious signal for models to pick up on.

---

[4]To borrow a term from Levesque (2014)

[5]Storks et al. (2019) and Schlegel et al. (2020) provide more comprehensive reviews of datasets and dataset ablations for natural language inference.

3.6.1 Proposal: Rethink annotation

Many of these issues result from the assumptions made in task design and in the underspecification of instructions given to human data labelers, and can thus can be addressed by rethinking the format that dataset collection takes. In light of this, recent work has proposed approaches to pre-empting spurious correlations by designing annotation frameworks that better leverage human 'common sense' (Srivastava et al., 2020) and more critical approaches to dataset creation and use for tasks such as reading comprehension (Gardner et al., 2019).

3.6.2 Proposal: Use tools to examine datasets

Recent work has proposed tools for using statistical properties of datasets to surface spurious cues and other issues with contents. The AFLITE algorithm proposed by Sakaguchi et al. (2020) provides a way to systematically identify dataset instances that are easily gamed by a model, but in ways that are not easily detected by humans. This algorithm is applied by Le Bras et al. (2020) to a variety of natural language processing datasets, and they find that training models on adversarially filtered data leads to better generalization to out-of-distribution data. Additionally, recent work proposes methods for performing exploratory data analyses based on training dynamics that reveal edge cases in the data, bringing to light labeling errors or ambiguous labels in datasets (Swayamdipta et al., 2020). Han et al. (2020) demonstrate the application of influence functions, originally introduced by Koh and Liang (2017) as a way to identify the influence of particular training examples on model predictions, to the discovery of data artifacts. The REVISE tool by Wang et al. (2020a) can be used to identify unequal representation in image description datasets by leveraging features of the images and the corresponding texts. Using their tool, they spot that images of outdoor sports activities are overwhelmingly labeled as men, and that in images where a person is too small for any sort of gender to be told at all, they are still labeled as men.

In response to a proliferation of challenging perturbations derived from existing datasets to improve generalization capabilities and lessen the ability for models to learn shortcuts, Liu et al. (2019) propose 'inoculation by fine-tuning' as a method for interpreting what model failures on perturbed inputs reveal about weaknesses of training data (or models). Recent papers also outline methodologies for leveraging human insight in the manual construction of counterfactual examples that complement instances in natural language processing datasets to promote better generalization (Gardner et al., 2020; Kaushik et al., 2020).

The case of VQA-CP (Teney et al., 2020b) provides a cautionary tale of when a perturbed version of a dataset is, itself, prone to spurious cues. This complement to the original Visual Question Answering (VQA) dataset, consisting of VQA instances redistributed across train and test sets as an out-of-domain benchmark for the task, was found to be easy to 'solve' with randomly generated answers. Cleverly designed sabotages that are meant to strengthen models' ability to generalize may ultimately follow the same patterns as the original data, and are thus prone to the same kinds of artifacts. While this has prompted attempts to make models more robust to any kind of dataset artifact, it also suggests that there is a broader view to be taken with respect to rethinking how we construct datasets for tasks overall.

These methods crucially rely on statistical patterns in the data to surface problem instances; it is up to human judgment to make sense of the nature of these problematic instances, whether they represent logical inconsistencies with the task at hand, cases of injustice, or both. Additionally, while a variety of recent papers have proposed methods for removing spurious cues from training data or 'de-biasing' models, recent work has shown that this can be damaging for model accuracy (Khani and Liang, 2021).

Considering that datasets will always be imperfect representations of real-world tasks, recent work proposes methods of mitigating the impacts of biases in data. Teney et al. (2020a) propose an auxiliary training objective using counterfactually labeled data to guide models toward better decision boundaries. He et al. (2019) propose the DRiFT

algorithm for 'unlearning' dataset bias.

Sometimes, noise in datasets is not symptomatic of statistical anomalies or labeling errors, but rather, a reflection of variability in human judgment. Pavlick and Kwiatkowski (2019) find that human judgment on natural language inference tasks is variable, and that machine evaluation on this task should reflect this variability.

3.7 Issue: Datasets pose privacy and consent issues to human subjects

Potential privacy violations arise when datasets contain biometric information which can be used to identity individuals, including faces, fingerprints, gait, and voice amongst others. However, at least in the US, there is no national-level privacy law which deals with biometric privacy. A patchwork of laws exist in Illinois, California, and Virginia which have the potential to safeguard the privacy of data subjects. However, only the Illinois Biometric Privacy law requires corporate entities to provide notice to data subjects and obtain their written consent (Khan & Hanna 2020, under submission). The EU General Data Protection Protection Regulation (GDPR), adopted in 2016, aimed at improving transparency about consumer data collection and empowering individuals to exert more control over what personal data was collected about them. For a look at personal data privacy laws worldwide, refer to Cortez.

Even in cases in which all data were collected legally from a copyright perspective — such as through open licenses like Creative Commons — many downstream questions remain, including issues about privacy, informed consent, and procedures of opt-out (Merkley, 2019). O'Sullivan (2020) discusses how technically legal uses of personal data that are not anticipated by or fully disclosed to the original owners of the data, e.g. the use of images scraped from the web to train facial recognition algorithms, constitute the ethical equivalent of data theft. Copyright guarantees are not sufficient protections for safeguarding privacy rights of individuals, as seen in the collection of images for the Diversity in Faces

and MegaFace datasets (Solon, 2019; Murgia, 2019).

Secure storage and appropriate dissemination of human-derived data is a key component of data ethics (Richards and King, 2014). To have a culture of care for the subjects of the datasets we make use of requires us to prioritize the well-being of the subjects in the dataset throughout collection, development *and* distribution. However, machine learning researchers developing such datasets rarely pay attention to this necessary consideration. Researchers will regularly distribute biometric information — for example, face image data — without so much as a distribution request form, or required privacy policy in place. Furthermore, the images are often collected without any level of informed consent or participation (Solon, 2019; Harvey and LaPlace, 2019; Raji and Fried, 2021). In the context of massive data collection projects, the potential harms extend beyond those that can be addressed with individual consent.

Even when these datasets are flagged for removal by the creators, researchers will still attempt to make use of that now illicit information through derivative versions and backchannels. For example, Peng et al. (2021) finds that after certain problematic face datasets were removed, hundreds of researchers continued to cite and make use of copies of this dataset months later. Without any centralized structure of data governance for the research in the field, it becomes nearly impossible to take any kind of significant action to block or otherwise prevent the active dissemination of such harmful datasets.

Additional security concerns arise due to the manner in which large-scale datasets are curated and disseminated through a web-scraping paradigm. For example, it was recently discovered that one of the URLs in the ImageNet dataset that originally pointed to an image of a bat instead linked to malware, potentially making dataset users vulnerable to hacking (O'Sullivan, 2020).

3.7.1 Proposal: Humanize data

Metcalf and Crawford (2016) go so far as to suggest the re-framing of data science as hu-

man subjects research, indicating the need for institutional review boards and informed content as researchers make decisions about other people's personal information. Particularly in consideration of an international context, where privacy concerns may be less regulated in certain regions, the potential for the data exploitation is a real threat to the safety and well-being of data subjects (Mohamed et al., 2020). As a result, those that are the most vulnerable are at risk of losing control of the way in which their own personal information is handled. Without individual control of personal information, anyone who happens to be given the opportunity to access their unprotected data to can act with little oversight, potentially against the interests or well-being of data subjects. This can become especially problematic and dangerous in the most sensitive contexts of personal finance information, medical data or biometrics (Birhane, 2020).

Recent work by Register and Ko (2020) illustrates how educational interventions that guide students through the process of collecting their own personal data and running it through machine learning pipelines can equip them with skills and technical literacy toward self-advocacy — a promising lesson for the next generation of machine learning practitioners and for those impacted by machine learning systems.

3.8 Issue: Dataset reuse strips away original context

Data management practices, such as the FAIR data principles (Wilkinson et al., 2016), assert the importance of making research datasets findable, accessible, interoperable, and reusable. Several scholars have written on the importance of reusable data and code for reproducibility and replicability in machine learning (Stodden and Miguez, 2014; Stodden, 2020), and the publication of scientific data is often seen as an unmitigated good, either in the pursuit of reproducibility (Pasquetto et al., 2017) or as a means of focusing research effort and growing research communities (e.g. through shared task evaluations (Belz and Kilgarriff, 2006)). Here, we want to consider the potential pitfalls of taking data which had been collected for one purpose and using it for one in which it was not intended,

particularly when this data reuse is morally and ethically objectionable to the original curators. Science and technology scholars have considered the potential incompatibilities and reconstructions needed in using data from one domain in another (Edwards, 2013). Indeed, Strasser and Edwards discuss several major questions for big data in science and engineering, asking critically "Who owns the data?" and "Who uses the data?" (2017, p. 341-343). Although in Section 3.3 we discuss ownership in a legal sense, ownership also suggests an inquiry into who the data have come from, such as the "literal [...] DNA sequences" of individuals (Strasser and Edwards, 2017, p. 342) or other biometric information.

Instances of data reuse in benchmarks are often seen in the scraping and mining context, especially when it comes to Flickr, Wikipedia, and other openly licensed data instances. Many of the instances in which machine learning datasets drawn from these and other sources which are serious privacy violations are well-documented by Harvey and LaPlace (2019), who discuss instances of scraping Flickr and other image hosting services for human images without explicit user consent.

Another concerning example of data reuse occurs when derivative versions of an original dataset are distributed —beyond the control of its curators— without any actionable recourse for removal. The DukeMTMC (Duke Multi-Target, Multi-Camera) dataset was collected from surveillance video footage from eight cameras on the Duke campus in 2014, used without consent of the individuals in the images and distributed openly to researchers in the US, Europe, and China. After reporting in the *Financial Times* (Murgia, 2019) and research by Harvey and LaPlace, the dataset was taken down on June 2, 2019. However, Peng et al. (2021) highlighted how the dataset and its derivatives are still freely available for download and used in scientific publications. It is nearly impossible for researchers to maintain control of datasets once they are released openly or are not closely supervised by institutional data repositories.

3.8.1 Proposal: Data stewardship

Researchers have thus proposed the use of better data stewardship protocols, such as promoting more standardized dataset management and citation practices (Peng et al., 2021) and the establishment of data consortia (Jo and Gebru, 2020). Best practices for sharing and managing datasets are a burgeoning area of research in natural language processing. In addition to a comprehensive accounting for the motivations and contents of abusive language datasets, Vidgen and Derczynski (2020) provide several suggestions for the responsible dissemination of such data, including the establishment of data trusts, platform-supported data sets, and the use of synthetic data.

3.9 Discussion

We have explored prominent issues in machine learning dataset practices, as well as some existing proposals for addressing these issues. A common thread pervading these issues seems to arise from a disciplinary culture that values rapid progress focused on large, cheaply acquired datasets that suffer from quality issues. Meanwhile, many of the proposals for addressing these issues suggest turning toward a more careful, systems-level and detail-oriented strategy for the collection, use, and stewardship of data.

We argue that fixes that focus narrowly on improving datasets by making them more representative or more challenging enrolls us in the Sisyphean task of finding and fixing dataset flaws rather than taking the necessary step back to address the more systematic issues at play. This renewed focus is essential to making progress as a field, so long as notions of progress are largely defined by performance on datasets.

We also recognize the need for a fundamental shift in the incentive structures that guide how machine learning practitioners prioritize dataset related tasks. The introduction of a "Datasets and Benchmarks Track" (Vanschoren and Yeung, 2021) at Neural Information Processing Systems Conference 2021, which will incentivize data-focused research, indicates a positive step in this direction.

It is also worth considering new models for data collection that involve enthusiastic, informed, reversible consent (`http://consentfultech.io/`) and voluntarily donated data aimed at helping to understand sensitive phenomena such as suicideality (e.g. the CLPsych 2021 shared task (Macavaney et al., 2021)).

3.10 Takeaways

In Chapter 2, we examined issues with common practices in the stages of machine learning development. In this chapter, we have explored data concerns in more depth. Now, we provide some recommendations based on our observations, and reflect on how actions at the different stages of development interact with each other. Figure 3.1 illustrates steps for each stage.

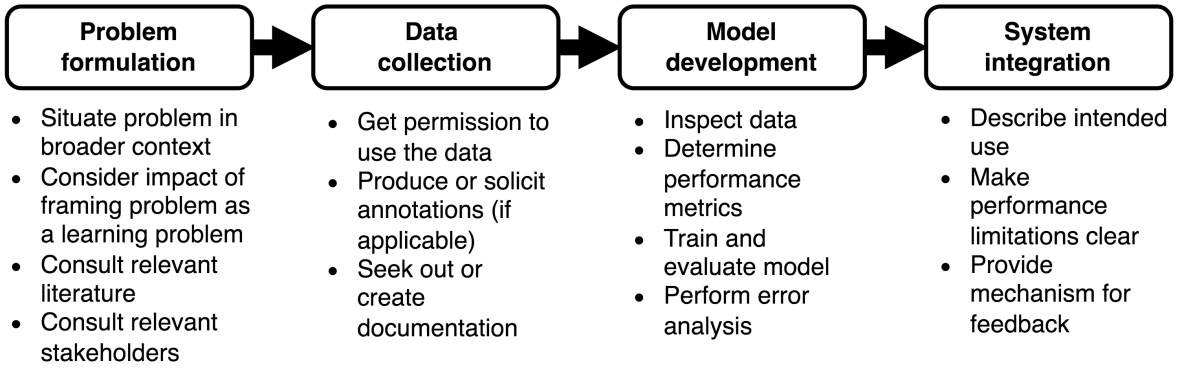


Figure 3.1: Idealized pipeline for a data-driven machine learning system

3.10.1 For formulating problems

We echo Costanza-Chock (2020) in that we should use design to liberate people from systems of oppression, not replicate them, and should thus actively seek out insight from the

most impacted stakeholders in the design of systems. We should become comfortable with the possibility that our interventions will be rejected. We also argue that faulty problem formulation negates any downstream concerns about data quality. To use a particularly perverse example, the DeepNude application applies a generative adversarial network trained on a dataset of nude images of women and takes as input 'a photo of a clothed person and creates a new, naked image of that same person. It swaps clothes for naked breasts and a vulva' (Cole, 2019). Whether the dataset used to train DeepNude was 'diverse' or not is irrelevant in considering the harms of such an application, because the ensuing product is a shameful and dehumanizing invasion of individual privacy regardless of the data distribution it was trained on. Additionally, the design of such an application relies on the collection of tens of thousands of images of nude women. This is incredibly sensitive data, and the developer did not disclose where or how they acquired the images. Although I can't imagine a socially beneficial reason to collect such data, for the sake of argument, the collection of such sensitive personal data should have a clear rationale justifying its collection *before* it is even collected. The DeepNude application is, arguably, not a justifiable rationale.

3.10.2 For collecting, distributing and using data

The massive sizes of contemporary machine learning datasets make it intractable to thoroughly scrutinize their contents, and thus it is hard to know where to begin looking for the kinds of representational and statistical biases outlined above. Indeed, a culture characterized by a desire to harness large datasets without questioning what is in them or how it got there, no matter how unsavory the details might be, produces what Vinay Prabhu calls the 'abattoir effect' (Raji et al., 2020). While many of the dysfunctional contents discovered in datasets were found by using intuition and domain expertise to construct well-designed dataset ablations and audits, some of the most disturbing were found by manually combing through the data.

As data solicitors often commission data labels at scale from crowdworkers, more mechanisms for feedback should be made available to data labelers to contest particular instances or raise red flags. Data labelers are arguably the most intimately familiar with the dataset and thus have valuable perspectives on dataset contents, and their input should be taken into account, rather than dismissed.

Downstream users of datasets should seek out documentation, and if it is not available, they should carefully consider the pitfalls of using under-documented data — for example, it may make error analysis difficult during the model development stage. Researchers have taken up the helm of retrospectively combing through data to document it retrospectively, which is a worthy pursuit, but I argue that we should not rely solely on this practice — we should also be more careful about what we collect in the first place, and document it prospectively.

We briefly summarized some proposals for procedural dataset modifications and bias mitigation techniques that can help in making systems more robust. We recommend the use of these tools, but emphasize that these methods are only useful insofar as the dataset in question itself represents a well-designed task. In other words, when making lemonade from lemons, we must ensure the lemons are not ill-gotten or poorly formed.

When distributing datasets, it is useful to host them on platforms like Zenodo, which provides stable identifiers and download metrics for datasets so that authors can revise, update, or remove data and have this reflected at a stable location. Although we cannot track every illicit instance of a dataset, we can at least promote better practices for sharing data.

When collecting data from the web, we should favor content that is available under a permissive license. While legal tools are just one way to promote standards for data collection and use, the average social media user does not carefully read terms of service before sharing data and content online (Obar and Oeldorf-Hirsch, 2020). Public backlash against social media platforms and other companies that harvest data, particularly biometric data such as pictures of people, shows how certain data collections, even those that are techni-

cally legal, can still be considered wrong and erode public trust (O'Sullivan, 2020). Clearly there is a disconnect between how people think their content will be used, and how it is actually used, and of course, just because something is technically legal does not mean it is moral. Thus, we should err on the side of caution when collecting data, and consider that many people who post content online have not provided their informed consent to the collection of their data.

3.10.3 For model development

In addition to carefully documenting and reporting details of model development prior to system release, developers should also make use of the data exploration techniques described above during model evaluation, to better surface edge cases and to inform error analyses. We echo proposals to minimize the environmental impact of model training.

3.10.4 For system integration

We echo Heuer and Buschek (2021) in advocating for more interdisciplinary collaboration between researchers in HCI and NLP to inform more user-centric, participatory modes of developing and releasing NLP applications. Because this stage involves complex real-world contexts, it is difficult to make general recommendations — a case-by-case analysis seems more appropriate.

In the following chapters, we will look at three case studies in NLP applications and tailor these recommendations to particular contexts.

Chapter 4

VECTOR SPACE MODELS OF TEXT

The purpose of this chapter is twofold: to summarize efforts to produce computational representations of text that can be used for language processing tasks, as background for the remaining chapters, and to provide a case study in the concerns implicated in the development of these models. In particular, we explore how the choice of training data has repercussions for downstream applications. We summarize some examples of the architectures used in such systems and the data sources used in their development.

4.1 Representing meaning

Suppose one wanted to find all articles about cats in a collection of documents. The Ctrl+F function, available on most modern computers and familiar to most users, can be used to identify exact string matches for the word 'cats' in a particular document. It would be inefficient, however, to manually perform this search over every document, and moreover, the process would need to be replicated with alternate expressions like the singular 'cat' or the synonym 'feline.' A more expressive, sophisticated approach is the use of regular expressions — for example, the regular expression '`cat(s)|feline(s)`' will capture matches for the strings 'cat', 'cats', 'feline', and 'felines', but this is still ultimately based on matching characters at the string level. There are several shortcomings to the string-match-based approach to searching for documents on a topic — we might turn up matches that look identical but are unrelated to our topic of interest, such as 'CAT scan' or 'CAT construction equipment' or idioms like 'someone let the cat out of the bag.' We are also limited by how many different ways we can come up with for expressing terms related to cats, and might miss relevant documents that use a term we didn't think of.

Going beyond simple string matching methods, if we want to find all documents that *conceptually* relate to cats: there are systems of cataloguing documents based on keywords or other metadata, relying on human categorization and taxonomies. However, these categories are not very flexible, and as language and culture change over time, so do our conceptions of what constitutes salient categories, and what names are appropriate for these categories (Buckland, 2017).

Rather than relying on exact character matches or available metadata and cataloguing systems for finding information, we can rely on *statistical properties* of language to produce useful representations of terms and concepts. Indeed, early work proposing a vector space model for documents based on their contents was addressing the problem of information retrieval (Salton et al., 1975). These days, vector space models of text have come to form the basis for a wide variety of tasks in natural language processing, and numerous techniques for building such models have been introduced.

Distributed methods are so named because they distribute values across a vector, contrasting with the one-hot approach. The distributed model is more efficient, and provides some interesting affordances in terms of geometric properties that correspond to conceptual relatedness. Within the distributed model, there are count-based methods for deriving vectors, which rely on word frequencies, and predictive models, which rely on predicting context for a word. Baroni et al. (2014) show that predictive models outperform count-based models (trained on the same data) on semantic similarity tasks and analogy tasks.

The modern NLP pipeline typically uses distributed, continuous vector representations of text, known as 'word embeddings,' as input to models. These embeddings have been shown to be effective as input to downstream tasks, but they have also been used for exploring associations in text.

Many of these approaches, and those that derive from them, rely on the distributional hypothesis (Harris, 1954; Firth, 1957), namely, that a word's meaning can be understood as corresponding to the contexts in which it appears. For an overview of word embeddings in NLP, refer to Pilehvar and Camacho-Collados (2020). Turney and Pantel (2010) also

provide a survey of vector space models in natural language processing.

4.2 Examples

Here we introduce a selection of architectures for learning vector space models.

4.2.1 Skip-gram with negative sampling

Mikolov et al. (2013a) introduce the skip-gram architecture. The skip-gram with negative sampling algorithm involves training a neural network to estimate the probability of a term t_c appearing within a sliding context window centered on an observed term, t_o. The training objective involves maximizing this probability for true context terms $P(t_c|t_o)$, and minimizing the probability $P(t_{\neg c}|t_o)$ for randomly drawn terms $t_{\neg c}$ that do not appear in such a context window, with the probability estimated as the sigmoid function of the scalar product between the input weight vector for the observed term and the output weight vector of the context term, $\sigma(\overrightarrow{t_o}.\overrightarrow{t_{c|\neg c}})$.

4.2.2 Embedding of Semantic Predications

Embedding of Semantic Predications (ESP) is introduced in Cohen and Widdows (2017). ESP is trained with a neural network architecture to estimate the probability of encountering the object, o, of a subject-predicate-object triple sPo. The training objective involves maximizing this probability for true objects $P(o|s, P)$ and minimizing it for randomly drawn counterexamples, $\neg o$, $P(\neg o|s, P)$. The non-negative normalized Hamming distance (NNHD) shown in Equation 4.1 to estimate the similarity between them. $P(o|s, P)$ is estimated as NNHD$(o, s \otimes P)$, where $\otimes$ represents the use of pairwise exclusive OR as a *binding operator*, in accordance with the Binary Spatter Code (Kanerva, 1996).

$$\text{NNHD} = \max\left(0, 1 - \frac{2 \times \text{hamming distance}}{\text{dimensionality}}\right) \tag{4.1}$$

4.2.3 Contextualized language models

The last few years have seen a surge in influence from so-called contextualized language models, which result not in static vectors for each word, but representations that are computed based on context. The ELMo model of Peters et al. (2018b) spurred interest in this innovation, and shortly thereafter, BERT (Devlin et al., 2019) combined the Transformer architecture of Vaswani et al. (2017) and a bidirectional masked language model objective to build on this work. BERT notably beat the GLUE benchmark human baselines, spurring new research in model interpretability (Rogers et al., 2020).

4.3 Affordances of statistical models of text

Here we discuss two key affordances of continuous models of text: the ability to surface statistical correlations between linguistic units, and to build on that affordance for finding relationships between concepts.

There has been much debate over the extent to which such models of text afford a system with a language facility on par with that of a human, since they apparently do well on GLUE. Whether or not these models absorb information that corresponds to linguistic theories of the underlying structure of language is still under active investigation. In recent work, we probed multilingual BERT for evidence that the model has encoded information that corresponds to theoretical constructs of linguistic structure (Shapiro et al., 2021).

4.3.1 Associations

By their nature as distributional models of text, the word associations represented in vector space models reflect conceptual correlations evinced in the data they are trained on. This makes such models useful tools for studying public discourse: for example, Rodman (2020) find that word associations in a corpus of historical news documents align with human assessments of the shifting dialogues around equality in the United States. On the

other hand, it is no surprise there are negative connotations and harmful stereotypes reflected in the word associations in these models, both in generic text (Caliskan et al., 2017) and clinical text (Zhang et al., 2020). Long (2021) writes, 'Indeed, their ability to detect bias is of value to scholars who wish to interrogate normative structures of stereotype, even as it warrants critical suspicion of their deployment in social media platforms and search engines under the pretense of algorithmic objectivity' (p. 218). However, as Thorp (2021) points out, it's not just that the tool is for a 'map of language, a playground for the linguistically curious' — these systems are used to 'make decisions – specifially classifications based on data' (p. 43).

Christian (2020) advises using these models *des*criptively rather than *pres*criptively as a way to surface the latent societal biases they model.

4.3.2 Analogies

Analogical reasoning has long been thought to be a critical component of generalizable intelligence. In an influential discussion in 1952, Alan Turing proposed analogies as a challenging example of intelligent, human-like reasoning for a computer (Copeland, 2004, p.499). Here, we discuss how vector space models of text have been tested for analogical reasoning capabilities.

Turney and Littman (2005) provided an early demonstration of the ability for algorithms to leverage vector space representations, inspired by the information retrieval work of Salton et al. (1975), to learn analogies from text in an unsupervised fashion and to evaluate this approach with arithmetic over vectors. Mikolov et al. (2013b) demonstrated how word2vec can be used to solve proportional analogy problems using simple geometric operators over vectors. This work demonstrated the use of a continuous vector space model of language that could be used for analogical reasoning when vector offset methods are applied, providing the following canonical example: if x_{w} is the vector corresponding to word w, x_{king} - x_{man} + x_{woman} yields a vector that is close in proximity to x_{queen}. This result

suggests that the model has encoded something about semantic gender. They identified some other linguistic patterns recoverable from the vector space model, such as pluralization: x_{apple} - $x_{\text{apples}} \approx x_{\text{car}}$ - x_{cars}. This model was trained on transcriptions of English Broadcast News.

However, work soon followed that pointed out some of the shortcomings of attributing these results to the models' analogical reasoning capacity. For example, Linzen (2016) showed that the vector for 'queen' is itself one of the nearest neighbors to the vector for 'woman,' and so it can be argued that the model does not actually learn relational information that can be applied to analogical reasoning, but rather, can rely on the direct similarity between the target terms in the analogy to produce desirable results. Furthermore, Gladkova et al. (2016) introduced the Better Analogy Test Set (BATS) to provide a more challenging evaluation set for analogical reasoning that includes a broader set of semantic and syntactic relationships between words, and found that models such as skip-gram performed underwhelmingly for certain relationship types.

As for contextualized word representations from language models such as BERT and ELMo, Peters et al. (2018a) showed (at least in the case of ELMo) they underperform other styles of embeddings in relational analogy tasks. One promising result comes from Jin et al. (2019), who show that BioELMo embeddings trained on biomedical text encode relational information better than BioBERT. Jin et al. (2019) probe biomedical term embeddings from language models also using nearest neighbor analysis. Researchers have proposed using prompts for eliciting knowledge from language model parameters — for a perspective on this problem in the biomedical domain, refer to Nadkarni et al. (2021).

4.4 Natural language from recycled materials

The Internet has both invited the production of a lot of user-generated text and facilitated the dissemination of it. Wikipedia and Common Crawl are some of the most common sources of 'pre-training' data for large language models. There are issues of provenance,

credit, and quality inherent to the scraping paradigm, which we discuss in more detail in Ch. 3. Tensions arise between the values of the free culture movement and intellectual property rights, and there is also a pervasive issue of power — who gets to collect and store data, who is given the informed ability to opt in or out of having their data included. Fair Use is determined on a case by case basis, and the legal status of large language models that are trained on large collections of text scraped from the web is under active investigation. Meanwhile, recent research has highlighted some of the ethical pitfalls of large language models, most notably Bender et al. (2021). For example, the word co-occurrences in datasets used to train these models frequently reflect social biases and stereotypes relating to race, gender, (dis)ability, and more (Caliskan et al., 2017; Garg et al., 2018; Hutchinson et al., 2020). This leads to issues when these models are used as the basis for a number of applications. For example, Speer (2017) discovered that a simple sentiment classifier, seeded with an existing sentiment lexicon and using trained GloVe vectors (trained on Common Crawl) as its base, rated the sentence 'Let's go get Mexican food' as having less positive sentiment than 'Let's go get Italian food,' where the only difference between the sentences is in the ethnicity named.

The quality of the training data and the privacy issues it poses have also come under investigation. Carlini et al. (2020) illustrate how sensitive, personally identifying information can be extracted from the training data of large language models. Caswell et al. (2021) show the value of manual audits of multilingual corpora to highlight the dubious quality of many datasets used for language model training. Their team of human volunteers, with proficiency in about 70 languages altogether, found that several corpora scraped from the web are rife with examples of mistranslated text and mislabeled linguistic content (i.e., content in a particular language labeled erroneously as belonging to another language).

Levendowski (2018) has argued that copyright is actually a useful tool for battling algorithmic bias by offering a larger pool of works from which machine learning practitioners can draw. She argues that, given that pre-trained representations like `word2vec` and other word embeddings suffer from gender and racial bias (Caliskan et al., 2017; Packer et al.,

2018), and other public domain datasets are older or obtained through means likely to result in amplified representation of stereotypes and other biases in the data (e.g. the Enron text dataset), that using copyrighted data can battle biased datasets and their use would fall under copyright's fair use exception.

4.5 Discussion

In order to ensure public trust, builders of language models should strive to incorporate data that is unequivocally, enthusiastically made available, e.g. with a Free Cultural License that permits radical reuse. Text sources like personal blogs constitute a morally dubious area for language modeling — while the websites that host blogs make these sources technically publicly available, internet users may be underage or otherwise unable to consent to having their data collected, and may be unaware that their personal thoughts are being scraped, read, and reused by researchers.

Many of the issues highlighted with language models has been due to the contexts in which they are used. Developers should clarify the intended downstream use case and ensure the model is adequately calibrated for this, or at least make shortcomings apparent.

These models have also been employed in generative art because the probabilities can be decoded to generate text from the model's parameters. Stochastic processes and remixing are well-established in the art world, as in the Dadaist movement, collage, etc. Rather than imbuing these models with any sort of divinatory power, I see them as a mirror of a particular set of ideas reflected in their training data, and as a generative process with which to build upon a lineage of human thought, similar to the Tarot or the I Ching. Artists have proposed ethical guidelines for producing generative art from neural networks that include critical reflection on data sources, the creator's relationship to those data sources, and the legal, ethical, and representational issues reflected therein (Leibowicz et al., 2021).

Chapter 5

MACHINE TRANSLATION IN CONTEXT

> "The great cultural barrier imposed by a separate language is perhaps the most effective guarantee that a social world, easily accessible to insiders, will remain opaque to outsiders... a unique language represents a formidable obstacle to state knowledge, let alone colonization, control, manipulation, instruction, or propaganda."
>
> — James C. Scott, 'Seeing Like a State'[1]

The practice of translation between human languages has long been shaped by power asymmetries (Gal, 2015). The boundaries applied to a continuum of linguistic communities and practices on the African continent to produce European understandings of discrete language objects, and indeed the very names applied to these objects, were imposed by European colonizers as the basis for the creation of language documentation and translation materials that undergirded colonializing efforts centuries ago (Errington, 2001). Some of the first grammar resources for previously unwritten languages were created by Christian missionaries in order to translate the Bible and proselytize to indigenous peoples worldwide (Pennycook, 2005). Throughout history are examples of colonial subjects who were forced to learn the languages of the colonizers, often facing punishment for speaking in their native languages (Bear Nicholas, 2011; Pak and Hwang, 2011). In many cases, this linguistic oppression has contributed to the decline in native speakers of indigenous languages, and the demand for colonial subjects to render themselves legible through obligatory translation further deepened their subjugation.

---

[1] Scott (2008, p. 72)

In 2019, the United States Department of Homeland Security (DHS) announced its plans to collect social media usernames of foreign individuals seeking entry into the United States, whether as travelers or immigrants, as part of a new "extreme vetting" process to determine admissibility into the country[2]. For those whose online activity is mediated primarily in a language other than English, an official manual distributed by US Citizenship and Immigration Services instructs officers to use Google Translate to translate their social media posts into English. This practice continued in spite of Google's warning that its translation services are not intended to be used in place of human interpreters (Patel et al., 2019; Torbani, 2019).

As the DHS social vetting protocol makes evident, the deployment of machine translation technology extends a tradition of surveillance and discipline of subordinate groups by forcibly rendering their speech legible. In this manner, language technologies enable a new kind of linguistic surveillance. In fact, such interests are precisely what fostered the development of machine translation technology in the mid-20th century.

In this chapter, we illustrate how the sociopolitical context in which machine translation was first developed has shaped the core goals and assumptions of the project, and how its continued development and use not only facilitate but require a consolidation of resources and power at an increasingly large scale. We explore how technology complicates the concept of language ownership and how linguistic communities worldwide have fought to reserve control over how their languages are used.

5.1 History of Machine Translation

We begin with a brief overview of the development of machine translation technology, focusing on text-based translation in the United States from the Cold War to today. Early rule-based systems were largely developed with funding from and for use by the military and

[2] https://www.federalregister.gov/documents/2019/09/04/2019-19021/agency-information-collection-activities-generic-clearance-for-the-collection-of-social-media

other federal agencies, often relying on interdisciplinary collaboration between engineers and linguists. After a period of steady research largely dominated by government-funded academic work, machine translation became widely available to the general public during the personal computing revolution of the 1990s, with the advent of commercially available translation software. By the 2000s, Google's tremendous index of web content and its swaths of capital enabled the enrichment and application of statistical (and later, neural) machine translation techniques, leading to the deployment of freely available translation services on the web as they are commonly used today.

5.1.1 Roots of Machine Translation: 1949-1997

> "[O]ne naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say, 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."
>
> – Warren Weaver, in correspondence to Norbert Wiener, 1947[3]

Modern machine translation traces its roots to work in cryptography and codebreaking during World War II. American scientist Warren Weaver, who had collaborated with pioneering information theorist Claude Shannon, had an interest in the application of information theory to the translation of human languages. In 1949, Weaver, then director of the Natural Sciences Division at the at the Rockefeller Institute, distributed a highly influential memo, entitled *Translation*, to a handful of linguists and engineers in which he laid out a call to action for the application of computers to the translation of human languages. Weaver's memo spurred a proliferation of research efforts in machine translation at a variety of institutions in academia and industry, including the University of Washington, Georgetown University, IBM, and the RAND corporation (Hutchins, 2000).

[3]Quoted in Hutchins (1997)

The mere decision of which languages to target for the first efforts in automatic translation was a political one, shaped at the time by Cold War rivalries between the United States and the Soviet Union, and spurred in particular by a desire to increase monitoring of scientific literature in Russian. Anthony Oettinger, then an undergraduate at Harvard University, recalls being recruited to collaborate with computer scientist Howard Aiken, one of the recipients of Weaver's memo, specifically because he was a student of Russian (Hutchins, 2000).

Research continued at a steady pace, and despite a promising system demonstration of Russian-English translation by the Georgetown-IBM team, funding dwindled in the 1960s in the wake of the damning ALPAC report that bemoaned the poor quality of machine translation. Then, in the late 1980s, a team at IBM applied the data-driven statistical methods they had been using for the task of speech recognition to the task of translation between English and French, largely on a whim[4]. The ensuing publication, 'A statistical approach to machine translation' (Brown et al., 1990), was initially met with skepticism, but proved to be incredibly influential; by some accounts, it was the impetus for the subsequent rise of statistical approaches to many more areas of NLP (Li, 2017).

Meanwhile, the United States government remained a faithful consumer of machine translation technology; in Tom Pedtke's 1997 keynote address at the sixth Machine Translation Summit, he reflects on several key developments in the 1990s fostered by government demand. For example, the Drug Enforcement Agency was devoting resources to the improvement of Spanish-English translation in 1991, while projects in Chinese-English and Korean-English translation were championed by the NSA, the FBI, DARPA, and the Navy (Locke and Booth, 1998). The end of the 1990s saw a shift in the key players in (and consumers of) machine translation.

---

[4]Li (2017) recounts how Brown et al. performed these experiments while their director, the famed Fred Jelinek, was on vacation, due to concerns that their proposal would be shot down on theoretical grounds (p. 200).

5.1.2 Data-driven Translation: 1997 to Now

> "The most important thing happening in Silicon Valley right now is not disruption. Rather, it's institution-building — and the consolidation of power — on a scale and at a pace that are both probably unprecedented in human history."
>
> – Gideon Lewis-Kraus, 'The Great A.I. Awakening'
> *New York Times Magazine*, Dec. 14, 2016 [5]

By the mid to late 1990s, advances in computational processing power and the personal computing revolution had enabled the development of translation tools for use by civilians. SYSTRAN, which had been developed out of Georgetown's original machine translation program, teamed up with hardware powerhouse Digital Equipment Corporation to launch AltaVista, the first free web-based translation service, in 1997. Originally limited to translation between English and a handful of Romance languages, it was widely applauded; user studies revealed heartwarming anecdotes of how the service enabled communication with beloved monolingual family members and provided a unique source of amusement when translations went awry (Yang and Lange, 1998).

The following year, in September of 1998, Larry Page and Sergey Brin incorporated the web search company known as Google. As graduate students at Stanford University, Page and Brin had begun work on building a massive index of the contents of the nascent World Wide Web, as part of the Digital Libraries Project funded jointly by DARPA, NSF, and NASA; this work would come to form the basis for the Google search engine (O'Mara, 2019). By 2004, Google was an enormously valued, publicly traded company that had earned the praise of web surfers worldwide. Brin claims that it was a message from a South Korean fan, mis-translated to 'The sliced raw fish shoes it wishes. Google green onion thing!' by the SYSTRAN software Google had been licensing, that spurred the decision to expand Google's capabilities to include the translation of languages (Helft, 2010). After all, in

[5] Lewis-Kraus (2016)

Google's quest to index all of the web, it would need to be able to include those parts of the internet that were not in English.

That year, Page reached out to Franz Och, then a research scientist at the Information Science Institute at the University of Southern California, to hire him to build what would become Google Translate. Och was skeptical at first, and bewildered as to why a search engine company would want to dive into the domain of translation, but was enticed by the fact that Google had unprecedented computational resources with which to push the frontiers of statistical machine translation, made newly tractable by the sheer quantity of text data at Google's disposal (Sarno, 2010b).

Over the next few years, under Och's direction, Google Translate vastly edged out other machine translation efforts by university research groups, developing efficient systems for dozens of languages. Mark Przybocki, who oversaw machine translation evaluation contests at the National Institutes of Standards and Technology in 2010, likened Google's competitive advantage to "going up against someone with access to a football field worth of processors to collect data." (Sarno, 2010a) Today, Google Translate boasts the ability to translate texts between over a hundred languages, and other tech giants like Microsoft and Facebook have also ventured into machine translation research.[6]

5.2 Consequences of deployment

A key driving force behind machine translation has been the quest for an exhaustive collection of knowledge that transcends local contexts. The earliest efforts in American machine translation were intended to decipher Cold War-era Russian communications and scientific papers, and now, Google has deployed its state-of-the-art machine translation tools to build its massive database of the world's online content. While the casual user of Google Translate ostensibly benefits from access to this resource, these free tools may be understood as 'hooks' that ensnare users further into the extractive relationship of surveillance capitalism

[6] https://translate.google.com/intl/en/about/languages/

(Zuboff, 2015) and 'shifts economic activity to a handful of tech giants as providers of translation' (Larsonneur, 2019).

While the key government benefactors of machine translation technology emphasized its utility for 'peacekeeping' via mutual understanding (Locke and Booth, 1998), Google advertises its translation service as a tool that 'break[s] language barriers and ... [makes] the world more accessible'[7]. This imagery of language as a "barrier" is often invoked in discussions of machine translation, offering a utopian view of universal understanding when these barriers are broken. Ironically, as the Department of Homeland Security's social media vetting process shows, translation software is deployed specifically to uphold cultural barriers, merely adding to the arsenal of technological tools for demarcating 'in' and 'out' groups (Torpey, 2018).

Further complicating the matter is that the apparent fluency of neural machine translation output for many language pairs can disguise the fact that systems still struggle to produce adequate translations, can amplify social biases, and are prone to inaccuracy in conveying important aspects of meaning like negation (Martindale et al., 2019; Stanovsky et al., 2019; Hossain et al., 2020; Prates et al., 2020). This is particularly dangerous when considering the high-stakes scenarios in which machine translation technology is frequently used and relied upon, such as in encounters between police and civilians, where a misunderstanding can be fatal (Liebling et al., 2020). At the same time, we must also be attentive to the conditions that make scenarios like police-civilian interactions so high-stakes in the first place. More accurate translation systems will not meaningfully disrupt stark power imbalances in society, and we should not pretend that they will.

We must be vigilant when applying probabilistic tools in an attempt to render legible that which has been obscured or distorted. At worst, the act of enlisting technology to attempt to translate a text in a language one does not speak may fit perversely alongside a broader trend of machine learning tools meant to forcibly reveal that which is obscured,

[7]https://ai.googleblog.com/2006/04/statistical-machine-translation-live.html

such as the DeepNude app described in Chapter 3, 'uncropping' algorithms for inferring obscured information in photographs, and 'de-pixelizing' algorithms for upsampling blurred images of human faces – thereby attempting to infer a 'ground truth' where it is otherwise humanly impossible to do so. In spite of the boundaries drawn by the distributors of Google Translate about the caveats of the system, the availability of the tool makes it irresistible.

As this chapter has been drafted in the midst of the global COVID-19 pandemic, we would be remiss to overlook the critical role that translation has played in the exchange and spread of vital information on best practices for prevention, testing, and the search for a treatment. The increasing reliance on automatic translation for gleaning insights from the international ecosystem of scientific knowledge has prompted calls for scholars to develop a 'machine translation literacy' toward an understanding of the shortcomings of automatically translating scholarly texts (Bowker and Ciro, 2019). The limitations of machine translation must be considered by technologists, policymakers, and affected stakeholders in delineating appropriate uses for it.

5.3 Rethinking and reshaping machine translation

> "The fact that language is not a tangible object that can be located or re-located makes issues of cultural ownership more subtle but also more urgent than for concrete pieces of art or other cultural objects."
>
> – Margaret Speas, 'Language Ownership and Language Ideologies'[8]

The training and evaluation of state-of-the-art neural machine translation techniques tends to rely on large, parallel collections of data produced by human translators, a practice informed by the information-theoretic roots of the paradigm. Weaver's characterization of translation between languages as mere decryption of encoded messages may

[8] Speas (2013)

seem crude to translation scholars and literary critics, some of whom have reservations about the possibility of faithful translation (particularly of literature and poetry; Weaver himself concedes this limitation (1955)). Indeed, the concept of 'equivalence' between texts is vigorously debated within translation studies (Panou, 2013). This is not to say that machine translation is epistemologically bereft; the parallel text basis of contemporary machine translation paradigms aligns with Quine (2013)'s pragmatic, behaviorist approach to translation (Piperidis, 2009). Whether one finds this framing compelling or not, it is nonetheless important to recognize that the data treated as gold standard translations embeds the situated and subjective positions of the people who wrote them, which impacts the ensuing associations embedded in automated systems.

The success of contemporary neural machine translation systems is largely due to a reliance on massive collections of linguistic data from the web. There are thousands of so-called 'low-resourced' languages (and minoritized dialects of dominant languages) for which there exist neither political nor financial incentives for industry powerhouses to develop translation tools, nor the sheer volume of digitized resources required for the successful application of neural machine translation. In this regard, there may be space for linguistic communities to be selective about whether — and if so, to whom — to submit their knowledge and culture for observation[9].

In 2005, the leaders of the Mapuche people issued an ultimately unsuccessful lawsuit against Microsoft, accusing them of 'intellectual piracy,' when the software company attempted to release a version of the Windows operating system in Mapudungun, the language of the Mapuche (Speas, 2013). Microsoft had not consulted with the Mapuche or sought their consent to use their language, instead working with the Chilean government to develop the resource, and yet the lawsuit came as a surprise. Technology has complicated the question of whether one can really 'own' a language; is a corpus of a thousand sentences scraped from the web enough to extract sufficient morphosyntactic features for

[9]Kilito (2008) explores the ethics of translation in the provocatively titled 'Thou Shalt Not Speak My Language' — a text that, ironically, this author could only encounter and enjoy in translation.

downstream processing and translation? What recourse does a linguistic community have if they do not wish to entrust software companies with the development of tools in their language?

Western discourses of language endangerment uncritically treat the development of technologies for low-resource languages as a social good, and indeed, the very framing of the 'low-resource' denomination implicitly prioritizes the gaze of the data collector, when speakers of a language have plenty of resources unto themselves in the form of idioms, jokes, fables and oral histories. On the other hand, forced assimilation and colonization led to stark decreases in numbers of native speakers of countless indigenous languages, and documentation and revitalization efforts for languages like Māori and Yup'ik become the focus of urgent attention. Efforts such as the recent First Workshop on NLP for Indigenous Languages of the Americas also encourage work in this direction.

In light of this, we consider alternative data collection and software development practices that subvert conventional paradigms. Adopting a participatory approach to addressing the paucity of technological resources for low-resourced languages, the Masakhane project proposes the creation of African language technologies by and for Africans, thereby involving the most impacted stakeholders in guiding the research direction and the curation of data from the very beginning of the project. Masakhane creates ways for participants without formal training in computational methods to participate directly and meaningfully, and represents a promising step toward using translation technology to empower native and heritage speakers of African languages (∀ et al., 2020).

In this chapter, we have considered how the creation, development, and deployment of machine translation technology is historically entangled with practices of surveillance and governance. Translation remains a political act, and data-driven machine translation developments, largely centered in industry, complicate the mechanisms by which translation shifts power. An awareness of the shortcomings of machine translation as a tool and as a paradigm are necessary for articulating appropriate contexts for its use.

Chapter 6

LEARNING FROM THE BIOMEDICAL LITERATURE

In this chapter, we briefly introduce the problem of extracting information from publications in biomedicine, and in particular, the task of literature-based discovery. We describe approaches to the problem and consider stakeholders in the process.

6.1 Background

Hundreds of thousands of research papers are published annually in biomedicine, and the pace is ever increasing. In 2020, almost 100,000 papers were published about COVID-19 alone (Wang and Lo, 2021). However, most researchers only read about an average of 250 papers a year (Tenopir et al., 2009), and it is increasingly difficult to keep up with the latest research. Swanson (1960) foresaw this 'crisis of inundation' and proposed that computational methods for retrieving information by searching over natural language text were a promising way forward. In the decades since, the rise of large digitized data resources has fostered the development of algorithmic systems in response to this challenge, with a proliferation of researcher-directed tools that leverage computational representations of the vast scientific literature to partially automate literature review, automatically extract and synthesize information, and recommend papers to read (e.g. Semantic Scholar[1], Trialstreamer[2]). Additionally, consumer-facing tools such as SUPP.AI (Wang et al., 2020b) facilitate access to information drawn from the scientific literature for non-experts.

The field of biomedical text mining has come to encompass a variety of tasks and approaches for tackling the problem first identified by Swanson (1960). Cohen and Demner-

[1] https://www.semanticscholar.org/

[2] https://trialstreamer.robotreviewer.net/

Fushman (2014) provide an in-depth introduction to the subject. In this chapter, we provide a brief overview of literature-based discovery in the biomedical domain. We thus focus on the mining of peer-reviewed literature largely drawn from journal publications and clinical trial reports, and not clinical text mining, which is based on text produced in a clinical setting, such as electronic health records and doctor's notes. For more on clinical text mining, refer to Cohen and Demner-Fushman (2014), Chapter 2.3, and Percha (2021).

6.2 Literature-based Discovery

Literature-based discovery (LBD) is a research paradigm that attempts to draw connections across disparate portions of published research to lead to novel insights or to assist in the generation of hypotheses about implicit connections between known concepts. It was originally proposed in a pair of studies (Swanson, 1986a, 1988) that applied manual analysis of article titles to identify common concepts across disjoint literatures. Thus, Swanson was able to hypothesize that *fish oil* might be a promising treatment for *Raynaud's disease* by observing, from paper titles, that patients with Raynaud's disease had high blood viscosity, and that dietary fish oil reduces blood viscosity. In spite of this logical connection, he found, through manual analysis of citations, that the literatures on Raynaud's disease and fish oil did not reference each other at that time (Swanson, 1986b). He applied a similar process to discovering a previously unpublished connection between *migraines* and *magnesium*.[3] These findings were validated by subsequent clinical research (Swanson, 1993). A closely related area of research, 'knowledge base completion' (KBC)[4], involves discovering information and structuring it such that it is amenable to augmenting databases of relational facts, often expressed as triples. In this fashion, literature-based discovery can

[3]In a biographical retrospective on Don Swanson, Neil Smalheiser notes that Swanson was driven to research these problems because he had Raynaud's and suffered migraines, himself (Smalheiser, 2017).

[4]Sometimes interchangeably referred to as 'knowledge graph completion' — refer to Ehrlinger and Wöß (2016) for commentary on this distinction or relative lack thereof.

be considered as a method that can assist in knowledge base completion, but literature-based discovery is not always applied to the problem of knowledge base completion, nor is knowledge base completion always achieved by literature-based discovery. In one recent example, Zhang et al. (2021) apply literature-based discovery to the problem of drug repurposing for the treatment of COVID-19, casting the problem as a knowledge graph completion problem using a literature-derived knowledge graph. Recent work has also examined the use of large language models trained on biomedical corpora as a foundation for knowledge-graph completion (Nadkarni et al., 2021).

In the last few decades, supported by the rise of large computational resources of data, several systems have emerged for leveraging computational representations of literature for doing literature-based discovery. LBD applications have largely remained within the biomedical domain, although some papers investigate LBD applications to domains such as climate science and water purification techniques (see Thilakaratne et al. (2019) for a review). Evaluation of LBD systems remains a challenge, largely in part due to the difficulty of curating validation data and the problem of defining what counts as a discovery; additionally, providing rationales for hypotheses remains an under-explored frontier.

In the following sections, we summarize some of the approaches to literature-based discovery, focusing on approaches based in the techniques of natural language processing.

6.2.1 Task formulations

In this section, we review the variety of ways of framing LBD as a problem. Although LBD has been structured as both a bibliometric analysis problem (i.e. using citation graphs) and a linguistic analysis problem (i.e. using models of the text of articles) (Thilakaratne et al., 2019), we focus on the latter.

Typically, 'literature' in LBD refers to peer-reviewed, published scientific literature, and 'discovery' refers to the general idea that something new has been found, typically a connection between concepts that exist already. Literature-based discovery can be considered

a particular task within the broader area of text mining. While text mining involves a breadth of tasks, LBD is specifically about 'discovery' i.e. establishing novel information. For surveys on the breadth of LBD applications and methods, including a book, refer to Bruza and Weeber (2008); Henry and McInnes (2017); Thilakaratne et al. (2019).

Swanson's original model is known as the 'A-B-C' model for literature-based discovery, in which concepts A and B are known to be connected, and concepts B and C are known to be connected, but A and C have not yet been explicitly connected — this is exemplified by the Raynaud's/fish oil and magnesium/migraine examples previously described. In the decades since, a majority of work in LBD has followed this paradigm, although other ways of framing the problem have also emerged. For example, Cohen et al. (2011, 2012) demonstrate the utility of casting the problem of literature-based discovery as an analogical retrieval task.

In the LBD literature, analysis is typically performed at the concept or term level. For example, in the A-B-C model, A can be a drug (either a name for a drug or a concept identifier from a medical vocabulary). In this fashion, 'knowledge' consists of a body of concepts, typically represented by tokens from processed text or standardized terms.

The task of 'discovery' in LBD, then, typically consists of the following: Given a discrete set of known entities E and relationships R between them, identify novel mappings, e.g. $< e_1, r, e_2 >$ such that $e_1, e_2 \in E$ and $r \in R$, that are implicit given the current state of the literature. In other words, LBD does not typically consist of the discovery of completely novel concepts or novel kinds of relationships (E and R remain fixed), but rather novel mappings between known entities and relationship types. Thus, it typically relies on resources, such as knowledge bases or vocabularies, that instantiate concepts and relationships between them. Weeber et al. (2001) distinguishes between open discovery, which involves exploring related concepts for a given entity in an open-ended fashion for hypothesis generation, and closed discovery, which involves prompting an LBD system with two entities of interest and finding related terms that may describe a relationship between as a way to test the hypothesis that the entities are related.

6.2.2 Evaluation approaches

Naturally, it is hard to validate a 'discovery' if it is truly novel. Clinical trials are the most favorable way to test these hypotheses, but very costly in both time and resources. It would be intractable to validate every hypothesis in this fashion. To that end, several proxy studies have been proposed to evaluate LBD systems. For example, Yetisgen-Yildiz and Pratt (2009) addressed a gap when there was no standard way to compare across LBD systems by proposing 'time-sliced evaluation' and also proposed at semantic hierarchy-based filtering to rule out certain generated hypotheses that were literally impossible or nonsensical. This method involves training a system on a subset of the literature up until a cut-off date and then seeing whether such a system can 're-discover' knowledge that has been established since that cut-off date. Bekhuis (2006) also advocates this approach. Other systems, including our own previous work (Paullada et al., 2020), use the retrieval of known information as an indicator of system quality as a proxy for the trustworthiness of novel hypotheses. We also used manual analysis of the existing literature to judge selected system hypotheses to the best of our ability, which is also a commonly used practice (e.g., Zhang et al. (2021)).

6.2.3 Stakeholders

Aside from the developers of such systems, the direct and indirect stakeholders in biomedical LBD systems include the imagined users of said systems (i.e. researchers and/or the general public who may be curious in examining the literature)and the research subjects whose data are represented in the literature. Because such systems have been applied to problems in the medical domain, virtually anyone who participates in health care systems stands to incur benefits or harms from the manner in which LBD findings are established and acted upon, whether directly or indirectly (e.g., through exclusion).

Identifying how users wish to use such systems is important, as well. Prior user research for LBD systems by Cohen et al. (2010) identified that users prefer using such systems for

discovering connections that are new to them, but not necessarily new to science. Bekhuis (2006) notes that such systems are 'exploratory and therefore useful in early phases of research programs or in proof-of-concept studies' and thus could be more aptly referred to as 'exploratory mining' (Bekhuis, 2006, p. 5). Workman et al. (2014, 2016) propose applying information-foraging theory to the design of a literature exploration system that promotes 'serendipitous knowledge discovery.'

6.3 Summary

We have briefly introduced the task of literature-based discovery, including a selection of current approaches and applications. In the following chapter, we describe our approach to building and validating a system for learning representations of biomedical relationships, which can be applied to the problem of literature-based discovery.

Chapter 7

EVALUATING LINGUISTIC REPRESENTATIONS OF BIOMEDICAL RELATIONSHIPS

7.1 Introduction

A vast amount of knowledge on biomedical relationships of interest, such as therapeutic relationships, drug-drug interactions, and adverse drug events, exists in largely human-curated knowledge bases (Zhu et al., 2019). However, the rate at which new papers are published means new relationships are being discovered faster than human curators can manually update the knowledge bases. Natural language processing tools have shown promise as an aid for curators. For example, Alex et al. (2008) provided evidence that human curators were able to speed up their curation process threefold with an NLP-driven protein-protein interaction discovery assistant.

Such an NLP pipeline can be supported by a system that produces linguistically-informed representations of biomedical relationships. In this work, we ask: How do different methods of processing biomedical literature, involving varying representations of linguistic structure, impact the ensuing representations of biomedical relationships? We validate different pipelines for processing biomedical text and building vector space models of this text with two extrinsic evaluation tasks: a relationship retrieval task that tests the models' ability to represent known relationships in the literature, and a task based on the literature-based discovery paradigm, introduced in Chapter 6, to test models' ability to infer relationships between entities that have not been stated explicitly in the literature.

This chapter is organized as follows. In Sections 7.2 and 7.3, we describe the linguistic resources that form the basis for this work. In Section 7.4, we describe our methods for producing models of text based on those linguistic resources. In Section 7.5, we describe

the extrinsic evaluation paradigm, based on an analogical ranked retrieval task, that we use to evaluate models of text. In Section 7.6, we describe the various human-curated knowledge bases that serve as validation data resources for our system, and our method of producing evaluation data sets from these sources. In Section 7.7, we describe the metrics we use to measure model performance. In Section 7.8, we describe our experimental setup, and in section 7.9 we present and analyze the results of these experiments.

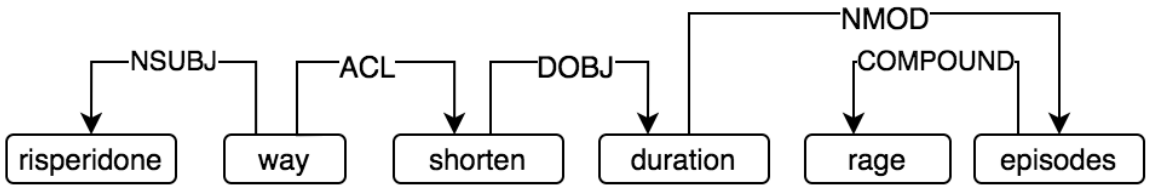
We describe our methods for producing vector space representations of relationships between entities, sourced from large collections of research papers in biomedical literature, and our pipeline for producing evaluation data sets from a variety of biomedical knowledge bases (Section 7.6).

7.2 Linguistic representations

Here, we discuss the different linguistic representations used in the present work.

7.2.1 Universal Dependencies

We use a corpus of dependency paths provided by Percha and Altman (2018), which are in turn derived from dependency parse trees from the Universal Dependencies formalism (De Marneffe and Manning, 2008). Percha and Altman (2018) user PubTator named entity recognition to identify mentions of Chemicals, Genes, and Diseases in the parsed sentences, and then prune the full parse trees to extract paths of dependency relations that connect entities of interest. Here, Chemical includes drugs, and Disease includes side effects and other conditions. Figure 7.1 shows an example of a path between the Chemical *risperidone* and the Disease *rage*.



"Liquid **risperidone** may be a safe and effective way to shorten the duration of **rage** episodes."

```
way_nsubj_START_ENTITY way_acl_shorten shorten_dobj_duration
    duration_nmod_episodes episodes_compound_END_ENTITY
```

Figure 7.1: Example of a dependency path derived from a sentence.

7.2.2 SemRep

SemRep is a multi-step NLP system, based on the Unified Medical Language System (UMLS), that extracts semantic triples from sentences in PubMed abstracts (Kilicoglu et al., 2020). The pipeline includes entity normalization via MetaMap, hypernym resolution, and negation processing. The SemRep ontology uses a subset of UMLS Semantic Network relations — 25 in total, as well as negated versions of the predicates. We use subject-predicate-object triples as input to train an Embedding of Semantic Predications (Cohen and Widdows, 2017) model (see later section for details). Prior work that uses SemRep triples as the foundation for an LBD system includes Hristovski et al. (2006) and Preiss (2014).

The following example of a sentence and the ensuing SemRep triple is drawn from Kilicoglu et al. (2020). Entities are italicized, while the predicate is bolded:

Input: 'Overnight incubation with 1 microM *safrole* did not **alter** *cell proliferation*'

Output: `Safrole-NEG_AFFECTS-Cell Proliferation`

7.3 Corpora

Here we describe the different corpora used as the foundation for training vector space models.

Each corpus consists of annotated abstracts from PubMed, each identified by a unique PubMed ID (PMID). We filter these such that the abstracts must also appear in MEDLINE and use December 2019 as a cutoff publication date across all abstracts, as this is the most recent date for which we have annotations from all three corpora. Figure 7.2 shows how each of these resources are related to one another. Table 7.3 shows some of the details for each corpus.

corpus	version	Abstracts processed	Final abstract count	Total sentences / predications	Vocabulary
PubTator	Feb. 15 2020	30,022,731	30,022,731	181,749,533 sentences	various vocabularies
SemMedDB	v. 40	29,115,337	18,184,627	97,972,561 predications	CUI from UMLS Metathesaurus
GNBR	v. 7	(unknown)	6,941,586	72,053,427 dependency paths	based on PubTator NER

Table 7.1: Details about corpora.

The MEDLINE 2019 Baseline consists of 28,111,922 total abstracts, represented by the rectangular region. These are the abstracts for which we have publication dates from MEDLINE. Of these, 27,890,839 appear in at least one of the three corpora used (represented by the three circles). PubTator, as of February 2020, consists of 30,022,731 abstracts, represented by the white circular region. 27,823,466 of these abstracts are included in the MEDLINE 2019 Baseline. Version 7 of GNBR consists of a set of 6,512,768 abstracts that are a proper subset of the PubTator abstracts. Version 40 of SemMedDB consists of 17,477,593 total abstracts. For all of PubTator, GNBR, and SemMedDB, there is a negligible number of abstracts that do not appear in the MEDLINE 2019 Baseline.

The darkest shaded region in Figure 7.2 represents the abstracts for which each pipeline (SemRep, GNBR, and PubTator) yielded at least one processed sentence. This intersection consists of 6,028,764 abstracts in common between all three corpora and included in the MEDLINE Baseline. Figure 7.3 shows an example of an abstract from this intersection and the resulting extractions from each pipeline. As the figure shows, neither SemRep nor GNBR produced a representation for *every* sentence in the abstract.

7.3.1 SemMedDB

We use Version 40 of the SemMedDB corpus (Kilicoglu et al., 2020), which is based on the MEDLINE BASELINE 2019 and uses SemRep version 1.8 to to produce predicate triples. This version of SemMedDB consists of 18,184,628 unique PubMed IDs.

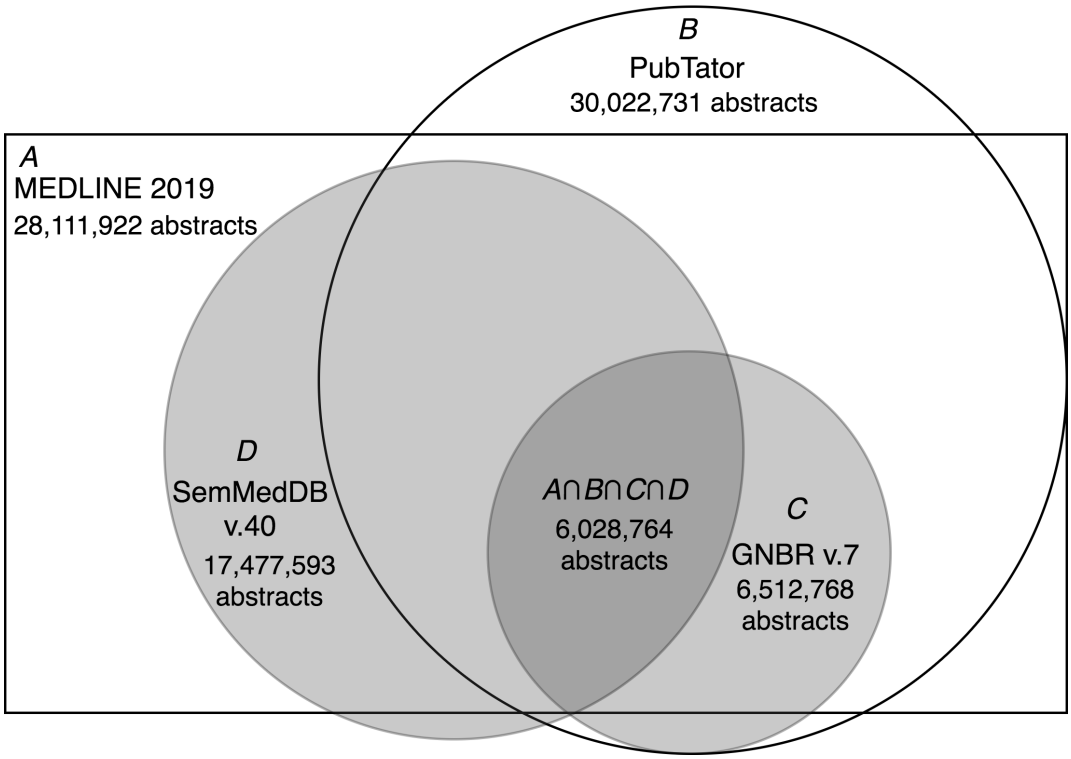


Figure 7.2: Illustration of overlaps between the corpora.

7.3.2 PubTator

We retrieve the February 2020 version of PubTator, which provides entity-level annotations and vocabulary mapping on 30,022,731 total PubMed abstracts (181,749,533 total sentences; average 6 sentences per abstract). We retain the full sentences and post-process

the text to normalize the vocabulary using the UMLS[1]. Figure 7.3 shows an example of the result of this pipeline on an abstract. In the original sentence "The therapeutic use of botulinum toxin was discovered in the 1970s and has since been used to treat patients with a broad range of medical complaints," the term "patients" is mapped to "man," referring to the species.

7.3.3 A Global Network of Biomedical Relationships

We use Version 7 of A Global Network of Biomedical Relationships (GNBR)[2] by Percha and Altman (2018), which is in turn based on a subset of the PubTator corpus. Version 7 of GNBR is based on the September 15, 2019 version of PubTator. GNBR consists of dependency paths that connect Genes, Chemicals, and Diseases as annotated in PubTator. We construct subject-predicate-object triples based on these dependency paths as input to train an Embedding of Semantic Predications (Cohen and Widdows, 2017) model.

GNBR (unfiltered) consists of 6,941,586 abstracts in total, in which there are approximately 72 million predications involving two of the entities of interest (Genes, Chemicals, and Diseases).

7.4 Embedding methods

After finalizing the set of abstracts to use from each corpus as described in the previous section, we process the data for training different embedding models, as shown in Figure 7.4.

---

[1]Oliver Li produced the original code for this process, which is currently unpublished.

[2]Available at `https://zenodo.org/record/3459420`

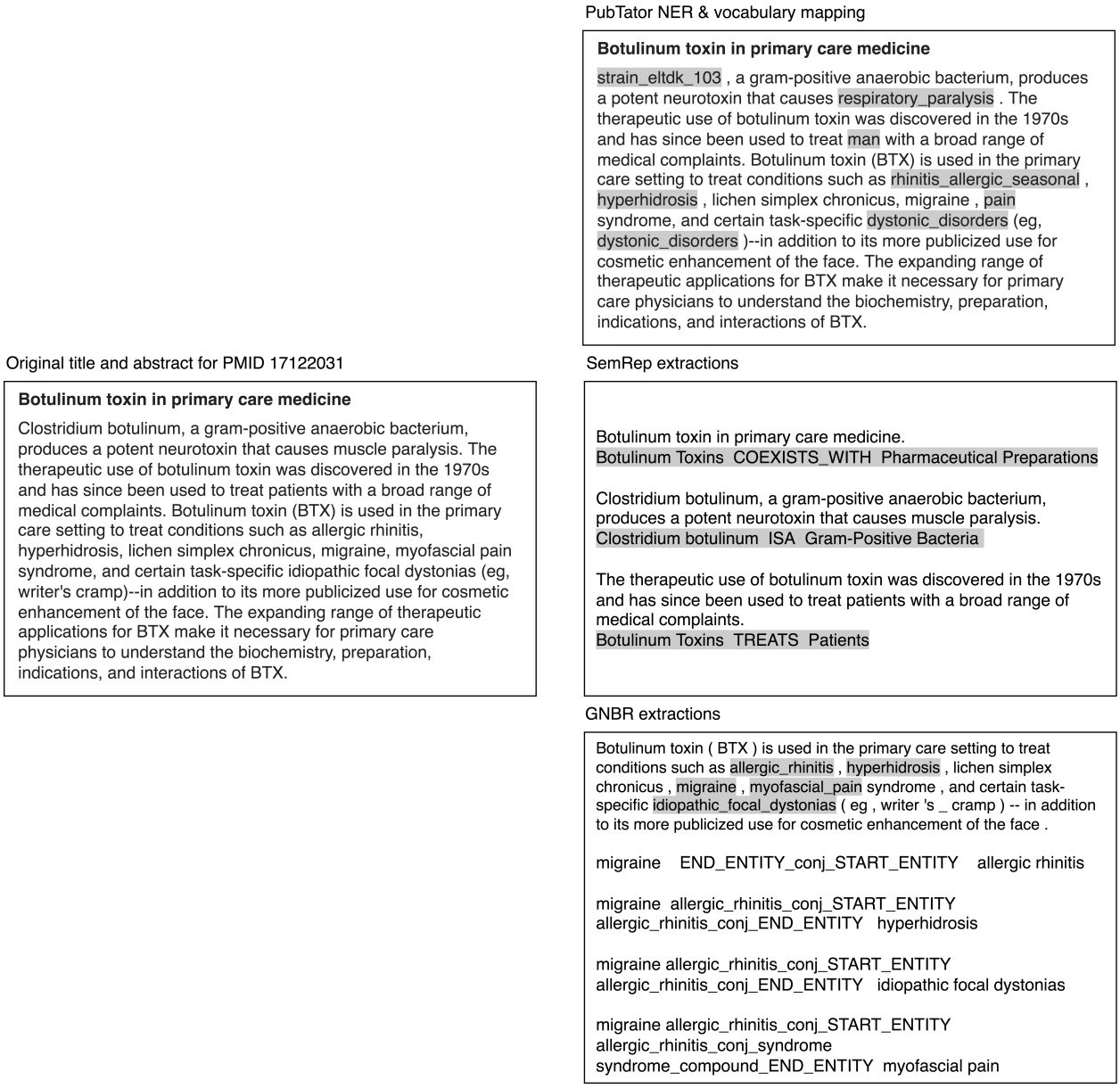


Figure 7.3: Comparison of extractions for abstract from Felber (2006). Entity recognition for all three pipelines and predicate triples for SemRep are shown highlighted in grey.

Skip-gram with Negative Sampling

We described the Skip-gram with Negative Sampling (SGNS) algorithm in chapter 4. We used the `Semantic Vectors`[3] implementation of SGNS to train 250-dimensional embed-

[3]https://github.com/semanticvectors/semanticvectors

dings, with a sliding context window radius of 2. We train the model on the full sentences processed by PubTator as described in Section 7.3.2. Our method is similar to Chen et al. (2020b).

Embedding of Structural Dependencies

We adapt the Embedding of Semantic Predications (Cohen and Widdows, 2017) algorithm to learn embeddings of dependency paths rather than single term predicates. Here, as in the original ESP work, we estimate the probability of encountering the object, o, of a subject-predicate-object triple sPo, but instead of SemRep predicates, the predicate is a concatenated dependency path (see Figure 7.1). We concatenate the dependency relations (the underscored parts in Figure 7.1) into a single predicate token for which a vector is learned. We used the `Semantic Vectors`[3] implementation of ESP, with binary vectors as representational basis (Widdows and Cohen, 2012). For the current work, we set the dimensionality at 8,000 bits (as this is equivalent in representational capacity to 250-dimensional single precision real vectors).

We have now described the corpora and the models. Figure 7.4 shows which corpora are used for which models.

7.5 Task formulation

We frame the task of evaluating models for their ability to represent biomedical relationships as an analogical retrieval task, using pairs of entities known to relate in particular ways. In our work, we follow the pipeline of Percha and Altman (2015) for assembling evaluation data because it consists of expert-curated structured data.

We follow prior work in using proportional analogies as a test of relationship representation in the general domain (Chapter 4) with existing studies on vector space models trained on generic English. Our biomedical data is largely in English, and we constrain our evaluation to specific biomedical concepts and relationships as we apply and extend

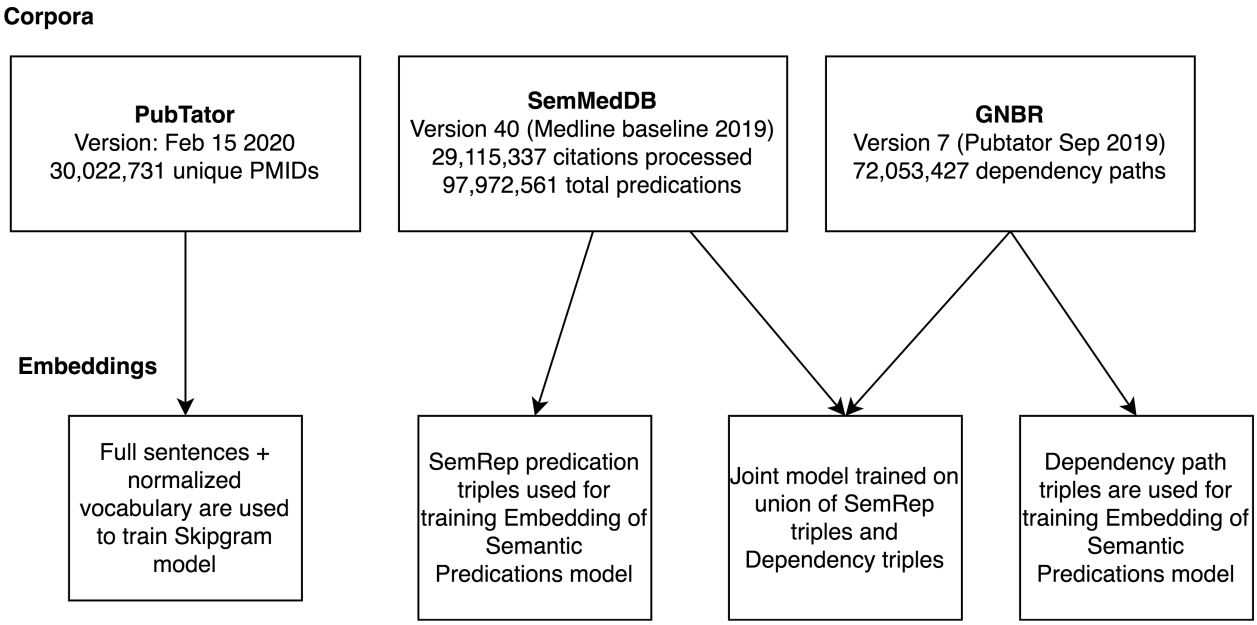


Figure 7.4: Details of corpora and the embeddings derived from them.

established methods.

7.5.1 Analogical ranked retrieval

As described in Chapter 4, a useful property of some distributed models of text is their ability to support the expression of proportional analogies using simple vector arithmetic. Here, we describe an approach to evaluating models for relational representations using an analogical retrieval task.

From a set of (X, Y) relationship pairs from a knowledge base, we select a number of (A, B) cue pairs and (C, D) targets, with the C held constant. The goal is to record the ranks of the targets D for a given C in a ranked retrieval task. The vectors for the cues and the target C term are summed as shown in Figure 7.5. A K-nearest neighbor search is performed (using cosine distance for SGNS, NNHD for ESP) over the search space and we record the ranks for each D target. For example, if the (C, D) pairs (*citalopram*, *de-*

pressive_disorder) and (*citalopram*, *dysthymic_disorder*) exemplify a treatment relationship, we draw a number of (A, B) cue pairs that also exemplify a treatment relationship, e.g. (*benzoyl_peroxide*, *acne*) and (*fluconazole*, *candidiasis*). Crucially, there is no overlap between (A, B) and (C, D) terms. That is, the pairs (*citalopram*, *anxiety*) and (*fluoxetine*, *dysthymic_disorder*), while valid exemplifications of treatment relationships, would be invalid cues for the (C, D) pairs above, because the vector similarity between the overlapping terms would confound the retrieval results.

We use an analogical ranked retrieval task for both the RR and LBD tasks. Figure 7.5 visualizes this process. From a set of (X, Y) entity pairs from a knowledge base, given a term C and all terms D such that (C, D) is a pair in the set, we select n random (A, B) cue pairs from a disjoint set of pairs. In prior work, Cohen et al. (2012) found that using a composite of cues, rather than a single cue, led to improved recovery. We refer to (C, D) pairs as 'target pairs,' correct D completions as 'targets,' and (A, B) pairs as 'cues.' The vectors for the cue terms (A, B) and the term C are summed in the following fashion to produce the resulting vector *v*. Given an analogical pair *A:B::C:D*, where *A* and *C*, *B* and *D* are of the same semantic type, respectively, we develop cue vectors for the target *D* in each model as follows:

$$\begin{aligned} \text{SGNS} : v = &\quad \overrightarrow{B} - \overrightarrow{A} + \overrightarrow{C} \\ \text{ESP} : v = &\quad \overrightarrow{I(A)} \otimes \overrightarrow{O(B)} \otimes \overrightarrow{I(C)} \end{aligned}$$

where I and O represent the input and output weight vectors of the ESP model, respectively. The SGNS method is the same as the 3COSADD method as described in Levy and Goldberg (2014). We sum the cue vectors and normalize the result.

A K-nearest neighbor search is performed for v (using cosine distance for SGNS, NNHD for ESP) over the search space, and we record the ranks for each correct D target. The search space is constrained such that it consists of those terms from our training corpus that have a vector in all three trained spaces, a total of about 300,000 terms overall. For ESP, this space consists of the output weight vectors for each concept. For the proportional

analogy task using K-nearest neighbors to rank completions to the analogy, the desired outcome is for the correct targets to be highly similar to the analogy cue vector v, such that the highest ranks are assigned to the correct target terms D in a search over the entire vector space. In this fashion, we perform this KNN search for every (X, Y) pair in the knowledge base and record the ranks for correct targets. We then compare the median ranks of terms D across both vector spaces; the higher the ranks, the better the model is at capturing relational similarity. Section 7.7 describes the metrics used for comparing system performance.

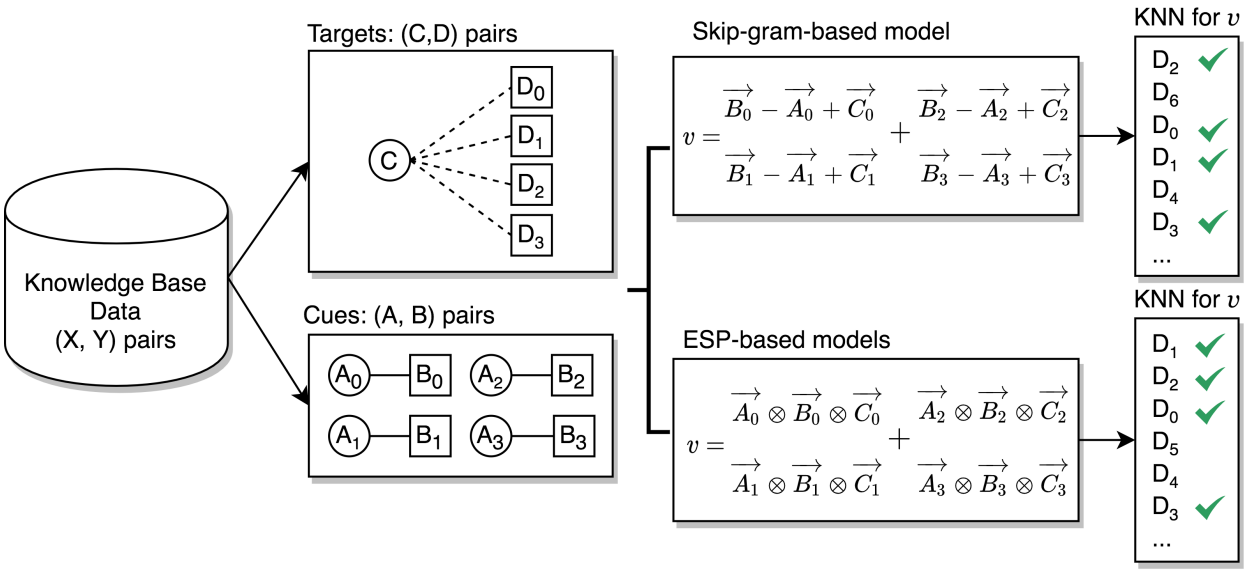


Figure 7.5: Simplified overview of analogical ranked retrieval paradigm.

7.5.2 Relationship retrieval (RR)

This problem asks: Given the literature, can we reliably recover explicitly stated information? This is a general assessment of the model's ability to represent relational information. Here, the AB and CD pairs are those that are drawn from knowledge bases and

are connected by a sentence or a path in our corpora, i.e., those pairs for which we have supervision.

7.5.3 Literature-based discovery (LBD)

This problem asks: Can we identify relationships between entities that weren't explicitly stated in the literature, but that we have reason to believe are related to one another? For this task, we use silver standard proxies, discussed in detail in section 7.6. Here, the AB pairs are still drawn from known relationships (specifically, they co-occurred in the corpora), while the CD targets are those that do not co-occur in the corpora, yet are drawn from an expert-curated knowledge base.

7.6 Evaluation data

In this section we describe the sources of data used for evaluating the models just described, and the manner in which we construct evaluation data sets from these data sources.[4]

7.6.1 Data sources

We distinguish between a *database*, which consists of a collection of facts that may be used for a variety of purposes and is typically used for reference, and a *data set*, which is curated as training and/or evaluation data for a specific task. We outline the process of extracting data from a *database* toward the creation of a *data set*.

Following Percha and Altman (2018), our evaluation pipeline operates on pairs of entities from the following databases: **DrugBank** (Wishart et al., 2018), **Online Mendelian Inheritance in Man (OMIM)** (Hamosh et al., 2005), **PharmGKB (PGKB)** (Whirl-Carrillo

[4]Bethany Percha wrote the code for extracting the data from various knowledge base APIs; this work is not yet published.

et al., 2012), **Reactome** (Fabregat et al., 2016), **Side Effect Resource (SIDER)** (Kuhn et al., 2016), and **Therapeutic Target Database (TTD)** (Wang et al., 2020c). Each of these resources uses different strategies for naming and normalizing concepts, which are not standard across each set, and have similarities and differences with the normalization strategies used in the processing pipelines described above.

The ensuing evaluation data sets from Percha and Altman (2018) consist of pairs of entities that relate in a specific way. For example, SIDER Side Effects consists of *chemical-disease*-typed pairs such that the chemical is known to have the disease as a side effect, e.g. (*sertraline*, *insomnia*). Meanwhile, another *chemical-disease* pair from a different database, Therapeutic Target Database (TTD) indications, is such that the chemical is indicated as a treatment for the disease, e.g. (*carphenazine*, *schizophrenia*).

In this section, we apply and adapt the Datasheets for Datasets paradigm by Gebru et al. (2021) to assess both the original databases and the ensuing data sets produced by Percha and Altman (2018) after post-processing. In particular, we are interested in assessing the quality of the data, documenting how it was curated, the license it was released with (if applicable), and providing examples from the literature that exemplify each relationship, where possible.

We also show an example sentence that exemplifies each relationship, using pairs of entities drawn from the respective knowledge bases.

DrugBank

DrugBank (Wishart et al., 2018) contains information about drugs and their disease targets as well as their genetic targets, or those targets to which a drug binds in order to have the intended therapeutic effect. DrugBank contains information about enzymes (which enzymes catalyze reactions with a particular drug as the substrate), carriers (proteins that bind to drugs for targeted delivery), and transporters (proteins that carry drug molecules in and out of cells).

Resource	Reference	How curated?	Sources?	License
DrugBank	Wishart et al. (2018)	expert curators	peer-reviewed biomedical literature	Creative Commons (CC BY-NC 4.0)
Online Mendelian Inheritance in Man (OMIM)	Amberger et al. (2015)	expert curators	peer-reviewed biomedical literature	OMIM Use Agreement `https://www.omim.org/help/agreement`
Pharmacogenomics Knowledge Base (PGKB)	Whirl-Carrillo et al. (2012)	expert curators	peer-reviewed biomedical literature and FDA data	Creative Commons (CC BY-SA 4.0)
Reactome	Fabregat et al. (2016)	expert curators	peer-reviewed biomedical literature	Creative Commons Public Domain
Side Effect Resource (SIDER)	Kuhn et al. (2016)	text-mining and manual review	drug labels from national registries	Creative Commons (CC BY-NC 4.0)
Therapeutic Target Database (TTD)	Wang et al. (2020c)	expert curators	literature and patents	*no license*

Table 7.2: Overview of information on knowledge bases used for evaluation.

Gene-Target example sentence: "AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to **losartan**, an **AGTR1** antagonist.' (Rhodes et al., 2009)

Enzyme example sentence: '**Losartan**, an antihypertensive agent, is also a substrate for **CYP2C9**.' (Fischer et al., 2002)

Carrier example sentence: 'Sex Hormone-Binding Globulin (**SHBG**), the plasma carrier for both **estradiol** and androgens, inhibits the estradiol-induced growth of MCF-7 cells (estrogen-dependent breast cancer cells), through its membrane receptor (SHBG-R), cAMP and PKA.' (Fazzari et al., 2001)

Transporter example sentence: '**Organic cation transporter 1 (OCT1, SLC22A1)** is a membrane transporter that is important for therapeutic effect of the antidiabetic drug **metformin**.' (Rulcova et al., 2013)

Online Mendelian Inheritance in Man

Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005; Amberger et al., 2015) is a knowledge base of causal/pathogenic mutation relationships between phenotype and genotype, curated by experts from the biomedical literature. The website provides descriptions and connections to relevant citations.

Example sentence: 'Mutations in fibrillin-1 (**FBN1**) cause a wide spectrum of disorders, including **Marfan syndrome**, which have in common defects in fibrillin-1 microfibrils.' (Kuchtey et al., 2013)

Pharmacogenomics Knowledge Base

The Pharmacogenomics Knowledgebase (PharmGKB) (Whirl-Carrillo et al., 2012) consists of manually curated drug-gene, gene-gene and gene-phenotype relationships, with the goal of representing the effects of genetic variation on reactions to drugs. It is available under a Creative Commons license. These relationships are extracted both via manual curation and with natural language processing techniques. Patient responses to the same drug can

vary due to genetic differences. Thus, 'personalized medicine' aims to use an individual's genotype to predict how they will respond to a drug.

Example sentence: 'Our study demonstrated an association of IL-10-1082 polymorphism and psoriasis and between **TNF** alpha-308 polymorphism and **psoriasis** disease and severity.' (Karam et al., 2014)

Reactome

Reactome (Fabregat et al., 2016) is a knowledge base of biomolecular pathways. Our evaluation data sourced from Reactome consists of pairs of genes that are known to react with each other and to form complexes with each other.

Example Complex sentence: 'In this study we demonstrate that **ERp57** forms discrete complexes with the ER lectins, **calnexin** and **calreticulin**.' (Oliver et al., 1999)

Thus our data has both (erp57, calnexin) and (erp57, calreticulin) as evaluation pairs.

Example Reaction sentence: 'The failure of mitochondrial reduction-oxidation (redox) homeostasis and the formation of excessive free radicals are tightly linked to dysregulation in the Renin Angiotensin System (RAS). A main rate-controlling step in RAS is **renin**, an enzyme that hydrolyzes **angiotensinogen** to generate angiotensin I.' (Vajapey et al., 2014)

Side Effect Resource

Side Effect Resource (SIDER) (Kuhn et al., 2016) is a set of adverse drug reactions (ADRs) and drug indications that are curated through applying an NLP pipeline to drug labels. It has been used previously as a benchmark for ADR extraction from text.

Example side effect sentence: '**Nausea**, dizziness and somnolence were the most commonly reported adverse events and were reported at a higher incidence by patients receiving **ropinirole** than by those receiving placebo.' (Matheson and Spencer, 2000)

Example indication sentence: 'Non-ergot-type dopamine receptor agonists such as **ropinirole** are used for the treatment of **Parkinson** disease, but they occasionally show serious

side effects including sleep attacks and daytime sleepiness' (Michinaga et al., 2010)

Therapeutic Target Database

The Therapeutic Target Database (TTD) Wang et al. (2020c) consists of information pertaining to relationships between drugs, diseases, disease biomarkers, and therapeutic targets for diseases, including proteins and nucleic acids. TTD indication relations express the same relationship as SIDER indications, although in our evaluation set, they have 83 pairs in common. This is likely due to different naming conventions for drugs and diseases in each resource.

7.6.2 Evaluation data set construction

We construct evaluation data sets using pairs from the knowledge bases described in Section 7.6. Table 7.3 shows, for each of the knowledge bases we use, the corresponding entity pair types, as well as the number of (X, Y) pairs from each that are used in our evaluation data. We also show the mean and maximum number of Y terms per X. While many X terms have multiple Y terms they relate to, for the *relationship retrieval* task, the most common number of Y terms per X is 1. This is also true for most of the evaluation sets for *LBD*.

The procedure for constructing an evaluation set for a given knowledge base below can be described as follows: given all (X,Y) pairs, let C be an X term and all its corresponding Y terms be D. Select n (A,B) cue pairs such that A $\neq$ C and B $\notin$ D. Algorithm 1 illustrates this process.

Data: Set K of knowledge base pairs (X, Y)

Result: Evaluation data set

$eval\_set \leftarrow$ list[];

for *all terms X* **do**

  $C \leftarrow$ X;

  $D \leftarrow$ list of all terms Y such that (X,Y) $\in$ K;

  $cues \leftarrow$ all pairs (X,Y) such that X $\neq C$ and Y $\notin D$;

  $cue\_sample \leftarrow n$ random pairs from $cues$;

  $quad \leftarrow$ list[C, D, cue_sample];

  $eval\_set$.append($quad$);

end

return $eval\_set$

Algorithm 1: Construction of evaluation set for a given knowledge base

The *relationship retrieval* data consists of knowledge base pairs that appear in our training corpora connected by a dependency path, predicate, or context window at least once in each respective corpus, while the *literature-based discovery* targets are those knowledge base pairs that do not appear connected by a dependency path, predicate, or context window in any training corpora. Mixed cases are excluded from evaluation. For example, a pair of terms that are connected by a dependency path in GNBR but do not appear in a SemRep triple nor within a size 2 context window of each other in the PubTator corpus is excluded from evaluation.

It should be noted that a term pair that appears in a human-curated knowledge base but does not co-occur in the corpus may consist of term synonyms that are expressed in different ways across these resources. For example, the pair (*dextroamphetamine*, *narcolepsy*), drawn from the TTD Indications database, co-occur in the corpora, but the pair (*dexedrine*, *narcolepsy*), also drawn from TTD Indications, do not co-occur. Dexedrine is a trade name

for dextroamphetamine.

7.7 Ranking and scoring

To evaluate model performance, we draw from the literature on information retrieval for evaluation metrics (e.g. (Berger, 2001)). For each model, we compute a composite performance score for each knowledge base in our evaluation data, using the following steps. First, during model evaluation, we record a list of the raw ranks of each D target term in the K nearest neighbor ranking for (C, D) pairs. Then, we compute the median of this list of ranks. This is a macro median, in the sense that we compute the median rank for all individual (C, D) instances. We then normalize the median rank, described as follows:

Normalized median rank Given a median rank as calculated above, we normalize it by the size of the search space and subtract from 1. This gives us a normalized median rank score where 1 is the highest possible score, and 0 is the lowest, making it easier and more intuitive to compare model performance across evaluation sets.

$$NMR = 1 - \frac{\text{median rank}}{|\text{search space}|} \tag{7.1}$$

We ran a simulation in which the entire search space was shuffled randomly 100 times, and recorded the median ranks of multiple target D terms, given some C. We found that the median rank for D terms in a randomly shuffled space tended toward the middle of the ranked list. Thus, the baseline score for this metric is established as 0.5; any score lower than this means the model performed worse than a random shuffle at retrieving target terms.

7.8 Experiments

We consider two pipelines for processing biomedical literature, as illustrated in Figure 7.2. Our full set of corpora consists of MEDLINE abstracts such that they all have the same publication cutoff date of December 2019. Pipeline one ('intersect') involves those abstracts at

Table 7.3: Pairs per knowledge base. Co-occurrence is computed with respect to the intersected corpus.

Dataset	Total terms		Pairs	
	X	Y	Co-occur	Do not co-occur
Chemical-Gene				
Enzymes (DrugBank)	932	245	878	1142
Targets (DrugBank)	1729	1944	1020	5304
PharmGKB	664	822	348	2953
Drug-Target (TTD)	482	132	170	169
Drug-Inhibitor (TTD)	231	75	97	84
Chemical-Disease				
Side Effect (SIDER)	299	111	289	436
Indication (SIDER)	608	354	677	419
Biomarker-Disease (TTD)	83	29	60	4
Indication (TTD)	643	140	373	110
Target-Disease (TTD)	205	126	162	61
Gene-Disease				
OMIM	805	338	438	110
PharmGKB	452	157	133	370
Gene-Gene				
Carriers (DrugBank)	222	51	49	185
Transporters (DrugBank)	578	149	258	954
PharmGKB	395	395	424	1410
Complex (Reactome)	182	175	133	23
Reaction (Reactome)	103	103	82	23

the intersection (darkest shaded region) of all three corpora. Because this set of abstracts is shared across all the models, we expect that the models that perform best compared to other models in this pipeline derive their advantages from the linguistic representations used, since no model has an advantage of having more abstracts. However, as shown in Figure 7.3, the predication-based models may not produce a representation for every sentence in the abstract. Pipeline two ('full abstracts') allows each model access to have the full extent of what its respective corresponding pipeline has been able to process. Here, we are interested in comparing performance by a model with a relatively small degree of preprocessing on a larger set of documents (as in our Skipgram model) with that of models with more linguistically informed and domain specific approaches to preprocessing, albeit with fewer processed abstracts as a result (our ESP-based models). In other words, we are looking at the impact of a tradeoff in precision versus generality in the processing pipeline.

In addition, we are interested in exploring (1) the impact of retrieval order: the relationships under consideration should be bidirectional, e.g. if drug A treats disease B, then disease B can be treated by drug A. We conduct experiments in which the cue pairs are issued in a canonical order (e.g., drug-disease) and a reversed order (e.g., disease-drug). Are any models 'biased' as it were for one ordering vs. another? and (2) the impact of using complete AB:CD analogies as our retrieval setup versus a directly similarity-based retrieval (B:D). This is to understand whether the models are relying on term similarity alone to perform rankings, or whether they appear to be utilizing relational information.

7.9 Results

In the following sections, we report the models' performance on the relationship retrieval and literature-based discovery tasks. We present results for experiments using the canonical retrieval order, because reverse order performance patterns were virtually identical to the canonical order. Appendix A shows these results for interested parties. We report overall system performance in Section 7.9.1, and examine the analogical retrieval abilities

of each model in 7.9.2. Finally, we perform a qualitative analysis for an example query in section 7.9.3.

7.9.1 Overall performance

In previous work (Paullada et al., 2020), we compared Skip-gram with Negative Sampling (SGNS) (Mikolov et al., 2013b,a) and Embedding of Semantic Predications (ESP); Cohen and Widdows (2017) using dependency paths as predicates. Here we found that ESP generally outperformed SGNS in both the relationship retrieval and LBD tasks. In these newer experiments, SGNS performs better in the relationship retrieval task, perhaps because the model had access to vastly more data (in the previous work, the SGNS model was trained only on those sentences for which GNBR had extracted a dependency path; this time around, SGNS has access to each sentence in its pool of abstracts) and a normalized concept vocabulary as compared with the setup from our previously reported work.

Contrary to our expectations, performance on relationship retrieval is not predictive of literature-based discovery performance: while SGNS has a comparative advantage on the former, the ESP-based models perform best on the latter. Also contrary to our expectations, restricting the models to using the intersection of processed abstracts (as opposed to the full set of available texts) does not change the performance patterns.

This is a rather low-precision system. Each C-term only has about 1-4 'known' relevant completions, and the search space is huge by comparison. However, because we are trying to promote 'novelty' and extrapolation from known relationships to unseen ones, precision is not a metric that reflects the desiderata for this system. Ideally, we would have human evaluation by an expert to determine whether the suggested novel relationships, as in the top ranked terms for a given query, are legitimate and interesting to explore. Qualitatively, this work so far serves primarily as a proving ground for different methods of representing text. It can be an interesting source of inspiration for experts to examine proposed associations between entities of interest, but probably should not be used in a downstream system

Table 7.4: Results for relationship retrieval task: Normalized macro-median ranks, full analogy and (B:D), full set of abstracts, 25 cues per target

Dataset	Avg. Targets	SGNS	ESP		
		+PubTator	+GNBR	+SemRep	+GNBR +SemRep
Chemical-Gene					
Enzymes (DrugBank)	2	**1.00** (1.00)	**1.00** (1.00)	0.64 (0.40)	**1.00** (0.83)
Targets (DrugBank)	2	**0.99** (0.66)	0.80 (0.76)	0.83 (0.47)	0.83 (0.52)
PharmGKB	2	**0.98** (0.82)	0.88 (0.97)	0.52 (0.49)	0.91 (0.62)
Drug-Target (TTD)	1	**1.00** (0.64)	0.99 (0.71)	0.99 (0.39)	0.99 (0.48)
Drug-Inhibitor (TTD)	1	**1.00** (0.70)	0.99 (0.66)	0.99 (0.36)	0.99 (0.53)
Chemical-Disease					
Side Effect (SIDER)	3	**1.00** (1.00)	**1.00** (1.00)	0.99 (0.22)	0.99 (0.57)
Indication (SIDER)	2	**1.00** (0.97)	**1.00** (0.99)	**1.00** (0.28)	0.99 (0.41)
Biomarker-Disease (TTD)	1	**0.99** (1.00)	0.95 (0.99)	**0.99** (0.09)	**0.99** (0.36)
Indication (TTD)	1	**1.00** (0.96)	**1.00** (0.99)	**1.00** (0.39)	**1.00** (0.45)
Target-Disease (TTD)	1	**0.99** (0.99)	0.86 (0.99)	**0.99** (0.20)	0.98 (0.62)
Gene-Disease					
OMIM	1	**1.00** (0.95)	0.98 (0.96)	0.99 (0.21)	0.99 (0.44)
PharmGKB	1	**1.00** (1.00)	0.78 (0.99)	0.99 (0.37)	0.99 (0.57)
Gene-Gene					
Carriers (DrugBank)	2	**0.86** (0.77)	0.84 (0.88)	0.33 (0.64)	0.77 (0.34)
Transporters (DrugBank)	2	**1.00** (1.00)	**1.00** (1.00)	0.50 (0.74)	**1.00** (0.48)
PharmGKB	3	**0.98** (0.80)	0.97 (0.99)	0.33 (0.71)	0.87 (0.57)
Complex (Reactome)	1	**1.00** (0.62)	**1.00** (0.74)	0.91 (0.45)	0.99 (0.48)
Reaction (Reactome)	1	**1.00** (0.65)	0.98 (0.78)	0.95 (0.54)	0.99 (0.53)

Table 7.5: Results for relationship retrieval task: Normalized macro-median ranks, full analogy and (B:D), intersection of abstracts, 25 cues per target

Dataset	Avg. Targets	SGNS	ESP		
		+PubTator	+GNBR	+SemRep	+GNBR +SemRep
Chemical-Gene					
Enzymes (DrugBank)	2	**1.00** (1.00)	0.99 (1.00)	0.69 (0.46)	**1.00** (0.86)
Targets (DrugBank)	2	**0.99** (0.62)	0.67 (0.72)	0.80 (0.48)	0.81 (0.55)
PharmGKB	2	**0.98** (0.78)	0.75 (0.97)	0.45 (0.52)	0.88 (0.54)
Drug-Target (TTD)	1	**1.00** (0.60)	**1.00** (0.63)	0.99 (0.47)	0.98 (0.39)
Drug-Inhibitor (TTD)	1	**1.00** (0.72)	**1.00** (0.63)	0.99 (0.59)	0.98 (0.36)
Chemical-Disease					
Side Effect (SIDER)	3	**1.00** (1.00)	0.96 (1.00)	0.99 (0.41)	0.99 (0.07)
Indication (SIDER)	2	**1.00** (0.95)	**1.00** (0.99)	**1.00** (0.22)	**1.00** (0.12)
Biomarker-Disease (TTD)	1	**0.99** (0.99)	0.60 (0.99)	**0.99** (0.03)	0.98 (0.55)
Indication (TTD)	1	**1.00** (0.94)	**1.00** (0.99)	**1.00** (0.16)	**1.00** (0.20)
Target-Disease (TTD)	1	**0.99** (0.96)	0.58 (0.99)	**0.99** (0.07)	0.98 (0.37)
Gene-Disease					
OMIM	1	**1.00** (0.91)	0.93 (0.88)	0.99 (0.32)	0.99 (0.32)
PharmGKB	1	**1.00** (0.99)	0.58 (1.00)	0.99 (0.18)	0.98 (0.26)
Gene-Gene					
Carriers (DrugBank)	1	**0.88** (0.78)	0.43 (0.99)	0.32 (0.57)	0.78 (0.75)
Transporters (DrugBank)	2	**1.00** (1.00)	0.96 (1.00)	0.54 (0.49)	**1.00** (0.35)
PharmGKB	3	**0.99** (0.78)	0.89 (0.99)	0.31 (0.55)	0.86 (0.49)
Complex (Reactome)	1	**1.00** (0.57)	**1.00** (0.70)	0.90 (0.60)	0.97 (0.50)
Reaction (Reactome)	1	**1.00** (0.58)	0.95 (0.80)	0.93 (0.59)	0.96 (0.46)

Table 7.6: Results for literature-based discovery task: Normalized macro-median ranks, full analogy and (B:D), full set of abstracts, 25 cues per target

Dataset	Avg. Targets	SGNS	ESP		
		+PubTator	+GNBR	+SemRep	+GNBR +SemRep
Chemical-Gene					
Enzymes (DrugBank)	2	**1.00** (1.00)	**1.00** (0.30)	0.77 (0.46)	**1.00** (0.10)
Targets (DrugBank)	4	**0.96** (0.85)	0.54 (0.45)	0.70 (0.66)	**0.96** (0.25)
PharmGKB	6	0.81 (0.83)	0.57 (0.41)	0.66 (0.57)	**0.89** (0.29)
Drug-Target (TTD)	1	0.81 (0.70)	0.57 (0.60)	**0.86** (0.70)	**0.86** (0.46)
Drug-Inhbitor (TTD)	1	0.82 (0.65)	0.54 (0.63)	**0.90** (0.82)	0.88 (0.51)
Chemical-Disease					
Side Effect (SIDER)	3	0.98 (1.00)	0.88 (0.36)	**0.99** (0.41)	0.98 (0.59)
Indication (SIDER)	2	0.95 (0.96)	0.90 (0.36)	**0.98** (0.42)	0.97 (0.54)
Biomarker-Disease (TTD)	1	0.95 (1.00)	0.68 (0.28)	**0.99** (0.29)	0.95 (0.60)
Indication (TTD)	1	0.95 (0.96)	0.98 (0.30)	**0.99** (0.24)	**0.99** (0.49)
Target-Disease (TTD)	2	0.86 (0.97)	0.55 (0.32)	**0.98** (0.44)	**0.98** (0.45)
Gene-Disease					
OMIM	1	0.85 (0.93)	0.52 (0.32)	**0.96** (0.33)	**0.96** (0.45)
PharmGKB	1	0.90 (0.97)	0.64 (0.42)	**0.98** (0.38)	**0.98** (0.57)
Gene-Gene					
Carriers (DrugBank)	1	0.69 (0.68)	0.64 (0.42)	0.34 (0.49)	**0.77** (0.58)
Transporters (DrugBank)	2	**1.00** (1.00)	0.92 (0.43)	0.96 (0.60)	**1.00** (0.33)
PharmGKB	5	**0.91** (0.91)	0.60 (0.49)	0.58 (0.55)	**0.91** (0.17)
Complex (Reactome)	1	0.66 (0.62)	0.66 (0.53)	**0.82** (0.62)	0.71 (0.42)
Reaction (Reactome)	1	0.78 (0.67)	0.56 (0.36)	**0.85** (0.74)	0.74 (0.47)

Table 7.7: Results for literature-based discovery task: Normalized macro-median ranks, full analogy and (B:D), intersection of abstracts, 25 cues per target

Dataset	Avg. Targets	SGNS	ESP		
		+PubTator	+GNBR	+SemRep	+GNBR +SemRep
Chemical-Gene					
Enzymes (DrugBank)	2	**1.00** (1.00)	0.89 (0.72)	0.86 (0.52)	**1.00** (0.44)
Targets (DrugBank)	4	**0.94** (0.83)	0.53 (0.52)	0.70 (0.50)	0.91 (0.70)
PharmGKB	6	0.79 (0.80)	0.54 (0.53)	0.66 (0.50)	**0.89** (0.56)
Drug-Target (TTD)	1	0.74 (0.64)	0.49 (0.52)	0.81 (0.34)	**0.82** (0.71)
Drug-Inhibitor (TTD)	1	0.75 (0.61)	0.56 (0.44)	**0.88** (0.28)	0.85 (0.70)
Chemical-Disease					
Side Effect (SIDER)	3	0.97 (0.99)	0.64 (0.47)	**0.99** (0.73)	0.98 (0.61)
Indication (SIDER)	2	0.92 (0.94)	0.66 (0.47)	**0.98** (0.76)	**0.98** (0.52)
Biomarker-Disease (TTD)	1	0.92 (0.99)	0.35 (0.24)	**0.99** (0.87)	**0.99** (0.54)
Indication (TTD)	1	0.94 (0.93)	0.66 (0.35)	**0.99** (0.69)	**0.99** (0.66)
Target-Disease (TTD)	2	0.85 (0.94)	0.54 (0.50)	**0.98** (0.72)	**0.98** (0.67)
Gene-Disease					
OMIM	1	0.77 (0.89)	0.49 (0.46)	**0.97** (0.60)	**0.97** (0.60)
PharmGKB	2	0.85 (0.94)	0.53 (0.55)	**0.98** (0.83)	**0.98** (0.59)
Gene-Gene					
Carriers (DrugBank)	1	0.64 (0.68)	0.51 (0.26)	0.33 (0.68)	**0.75** (0.41)
Transporters (DrugBank)	2	**1.00** (1.00)	0.73 (0.53)	0.98 (0.33)	**1.00** (0.48)
PharmGKB	5	0.89 (0.89)	0.57 (0.49)	0.60 (0.59)	**0.91** (0.54)
Complex (Reactome)	1	0.69 (0.54)	0.62 (0.64)	**0.73** (0.62)	0.69 (0.53)
Reaction (Reactome)	1	0.81 (0.63)	0.52 (0.65)	**0.86** (0.43)	0.78 (0.36)

for making official research recommendations at this stage.

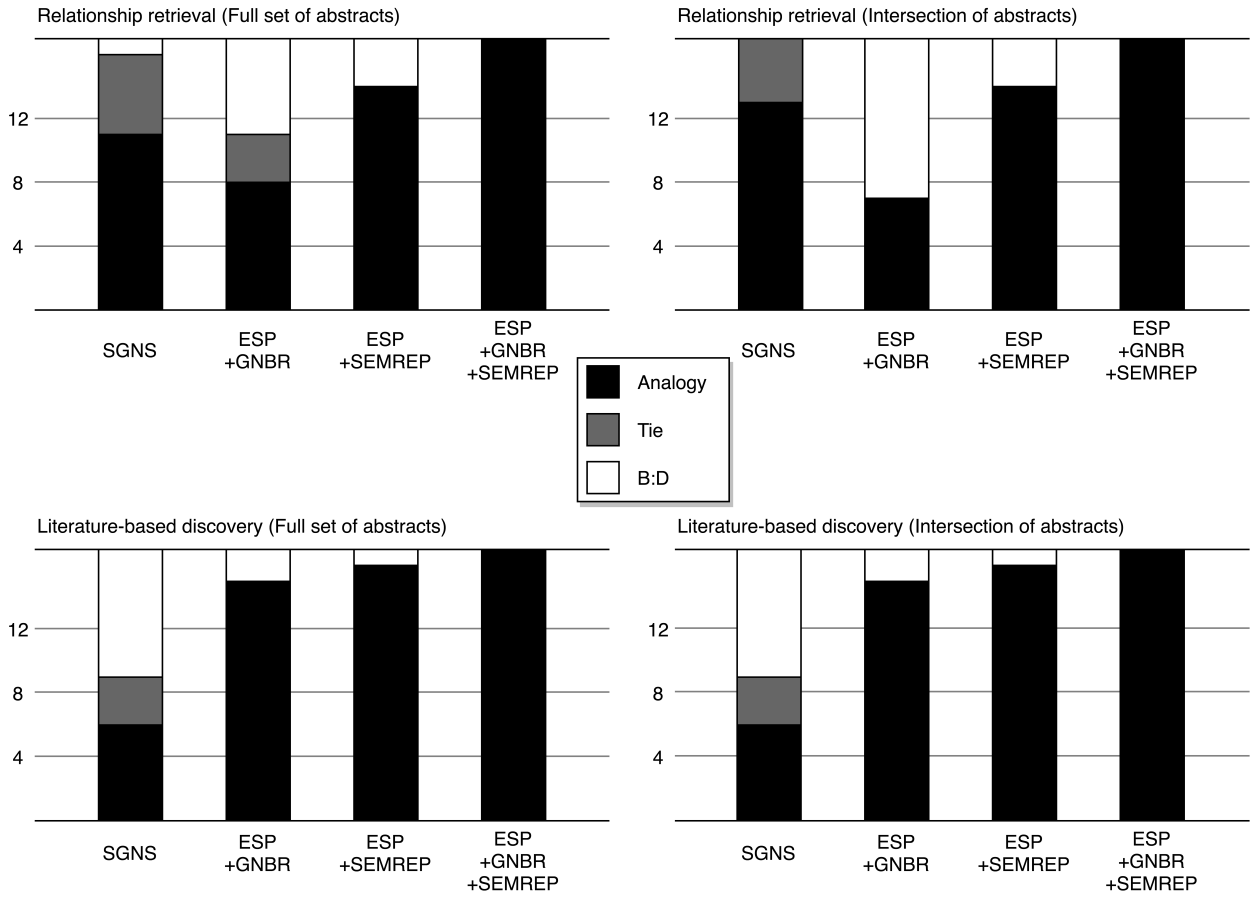


Figure 7.6: Proportion of model results in which analogical retrieval score is better than direct similarity (B:D) score. Columns sum to 17, the number of evaluation sets in our data.

7.9.2 Analogies vs. direct similarity

We compare models based on whether the performance for a given evaluation set was better when prompted with a full analogy or with just direct term similarity. These results are shown in Figure 7.6. Here, we find that the ESP+GNBR+SEMREP model, in all cases,

has better *analogical* retrieval performance (i.e., higher median ranks on average) than direct similarity-based retrieval for all sets. Although the SGNS model often outperformed the ESP based models in terms of ranking correct retrievals more highly in the RR task (Section 7.9.1), it appears to have relied more on direct term similarity than relational information for this performance.

7.9.3 Qualitative results

In Table 7.8, we show the results of a qualitative analysis to find potential treatment targets for *citalopram*, a selective serotonin reuptake inhibitor (SSRI) typically used to treat depression. Using 25 (drug, indication) pairs drawn from SIDER Indications as cues and *citalopram* as the prompt, we compose a query vector and perform a K-nearest neighbor search over terms that *do not* co-occur with *citalopram* in any of our three corpora. We focus on the skip-gram based model and ESP+GNBR+SemRep, the best-performing ESP-based model. As the table shows, many of the top ranked retrieved terms for SGNS are actually other drugs (indeed, the top result is 'citalopram' itself), which are categorically the wrong type to be potential treatment targets, while all of the top retrieved terms for the ESP-based model are the correct semantic type. Indeed, the DrugBank entry for *citalopram* notes that it has been found to alleviate symptoms of depression, obsessive-compulsive disorder (OCD), anxiety, eating disorders, and other mood disorders. Additional research, e.g. Anderberg et al. (2000), suggests that citalopram can alleviate some of the depressive symptoms comorbid with fibromyalgia.

7.9.4 Interpretation of results

Why might relationship retrieval performance not be predictive of literature-based discovery performance? The results in Section 7.9.2 illustrate how, for the literature-based discovery task, the SGNS model relies more on direct term similarity for performance, while the ESP based models appear to be using relational information to solve propor-

rank	SGNS	ESP+GNBR+SemRep
1	citalopram ×	**melancholia**
2	paroxetine ×	**major_depressive_disorder**
3	fluoxetine ×	**panic_disorder**
4	venlafaxine ×	**seasonal_affective_disorder**
5	**anxiety_disorders**	**agoraphobia**
6	sertraline ×	manic
7	reboxetine ×	**agitation**
8	mirtazapine ×	**premenstrual_dysphoric_disorder**
9	**mood_disorders**	**binge_eating_disorder**
10	**panic_disorder**	**bulimia**
11	fluvoxamine ×	catatonia
12	clomipramine×	**premenstrual_syndrome**
13	cipramil ×	sleeplessness
14	milnacipran ×	**fibromyalgia**
15	vilazodone ×	delirium

Table 7.8: Top 15 results for a search for potential treatment targets for the drug *citalopram*. **Bolded** terms indicate a condition that citalopram treats or might treat, based on a search to DrugBank or a cursory literature search; An × indicates terms that refer to a chemical, i.e., something that could not be a treatment target for a drug.

tional analogies. Table 7.8 showed how SGNS, even when prompted with an analogical retrieval cue, tends to rank similar terms, of the same semantic type as the prompt, more highly than legitimate completions to the analogy. Considering that the relationship retrieval task involves target term pairs that *co-occur*, SGNS may earn its advantage from the fact that its training involves maximizing the cooccurrence probability of words within a

linear context window of each other. In other words, the relationship retrieval task, as we have framed it, may really just be a test of a model's ability to model relatively narrow word contexts. Meanwhile, Table 7.8 shows how the ESP based model has highly ranked plausible completions to the analogy that are of the correct semantic type, suggesting that explicitly training on relational information leads to a better encoding of semantic information, rather than mere contextual similarity. This higher-order information is necessary for the LBD task, which involves retrieval of targets that were not expressed in the literature. In other words, a model based on linear cooccurrence is at a disadvantage compared to a model that has encoded relational contexts from which to extrapolate.

One limitation of this interpretation is that it is unclear the extent to which performance is impacted by which vocabularies and normalization strategies are shared between the knowledge bases and the processing pipelines for each model, partly because this information is multi-faceted and in some cases hard to find. Although we base our evaluations on the set of terms in common across each model, this does not account for what models may have encountered during training, and there is a chance that the models that perform best for a particular evaluation set simply have more vocabulary in common with that set, and thus have better representations (more training examples) for those terms.

7.10 Summary

We have described our approach to evaluating a selection of text representation models for their ability to encode biomedical relationships of interest. In the following chapter, we reflect on some of the limitations of this work in light of the data available to us.

Chapter 8

DISCUSSION

In this chapter, I discuss takeaways and limitations across the case studies, providing suggestions for future work.

8.1 The role for machine learning in biomedical research

In a recent conference about the use of machine learning in clinical research, with stakeholders including biomedical and machine learning researchers and representatives from organizations such as pharmaceutical companies, the US Food and Drug Administration, and patient advocacy groups, 'no consensus about best practices was reached' (Weissler et al., 2021, p. 2). Clearly, this is still an active area of research, and we should strive to shape it in a positive direction. Two angles to consider are the connection between machine learning developers and system users, and the connection between medical research subjects and wider systems of oppression.

Forsythe's (1993; 1996) extensive anthropological study of the people behind medical AI systems reveals how system builders are often detached from the key stakeholders, and how this leads to incorrect assumptions about what users want and need from computational interfaces to medical system. The programmers built abstract systems that users later rejected, because they had gone about the development process *assuming* what a user might want, without asking them first (Forsythe, 1993, 1996). In the work described in Chapter 7, my initial view of the knowledge base data I was using was predominantly as a 'proving ground' for validating different NLP pipelines for biomedical relationship extraction. However, the more time I spent analyzing the models' performance and trying to make sense of the patterns therein, and the more time I spent documenting the de-

tails of each knowledge base, it occurred to me that I have a poor understanding of what knowledge base curators, often human experts who perform painstaking reviews of the literature, actually want from a tool that could assist them in this practice. While my initial view of the project was as an abstract set of experiments for evaluating computational models of text, which could ostensibly inform later user-facing work, it seems to me that this way of working is backwards. Thus, in future work, I aim to incorporate more stakeholder feedback much earlier in the process.

Numerous studies in recent years have brought attention to the various systemic inequities that exist in biomedical research, ranging from which diseases are given the most attention and funding (Marshall et al., 2021), sex disparities in clinical trial enrollment (Feldman et al., 2019), and racial discrimination in medical care (Washington, 2006; Doll, 2018). Biomedical data on a particular question might not be comprehensive, as there are disparities in data collection at almost every step with respect to gender, patient condition, race, and even the patient's insurance status (or lack thereof) (Chen et al., 2019, 2020a). Thomas (2021) notes that diagnosis of Crohn's disease takes 12 months for men, and 24 months for women, while diagnosis of Ehlers-Danlos Syndrome takes 4 years for men, 16 for women, stressing that in light of these long wait times, people are likely to give up seeking a diagnosis, and that the research record is missing important data as a result. Studies also show how endometrial cancer is under-diagnosed in Black women because diagnostic methods underperform for them, or are not applied soon enough (Eichelberger et al., 2016; Doll et al., 2018; Doll, 2018).

The various data sources used in the experiments in Chapter 7, thus, are all subject to the biases incurred when research is not representative or is outright discriminatory, and so our findings must be taken with this context in mind. Our system operates on coarse-grained understandings of biomedical relationships — for example, we treat 'Drug X treats Condition Y' as a static fact in our evaluation, but we do not have finer-grained details for 'Under what conditions, and for whom?' Future work can provide a mechanism for a system user to explore this finer-grained information.

While there are gaps in our knowledge based on who is studied and who is represented, the people who are contributing to research are not necessarily those who are benefiting the most from it. For example, the Pima Indians Diabetes Dataset (PIDD) in the UCI Machine Learning Repository has been used thousands of times as a 'toy' classification task (Radin, 2017). The data were collected by the National Institutes of Health from the indigenous Akimel O'odham community, which had been extensively studied for their high prevalence of diabetes — which in and of itself stemmed from a history of displacement and settler-colonialism. Radin notes that PIDD "was used to refine algorithms that had nothing to do with diabetes or even to do with bodies, is exemplary of the history of Big Data writ large." (2017, p. 45). However, their participation in research had not yielded any significant decreases in obesity or diabetes amongst community members.

Another example involves the HeLa cells extracted from Henrietta Lacks, a Black American woman who died in 1951. A sample of cervical cells taken from her, without her knowledge or consent, have since become an 'immortal cell line' that has been the basis for innovations such as the polio vaccine, various cancer treatments, and more recently, COVID-19 vaccines. Her descendants were unaware of Henrietta's influential role in medical research, and did not receive any compensation; in fact, they had issues accessing medical care, themselves (Skloot, 2017). Thus, mere representation in medical data does not guarantee that subjects will benefit from the ensuing research, and we should look beyond strategies for diversifying data and toward eliminating exploitative research practices.

8.2 Implications for NLP system design

In my own practice, I have often felt I was missing guidance on how to structure and justify system design choices in ways that are rooted in consideration of broader impact. This is partially due to my own ignorance of relevant literature until too late. A variety of design frameworks have been proposed for considering ethics and values in the development of

technological systems (Donia et al., 2021). For example, Schnoebelen (2017) argues for a goal-oriented approach to NLP system design which includes consideration of the goals of those impacted by a system's deployment. With respect to designer agency, scholars vary in their assessment of how much responsibility designers can meaningfully take on in reflecting on their values and assumptions. There are often societal and structural issues that they must contend with, which produce different incentives or struggles even when designers are able to reflect meaningfully on the ethics and values of their process. Additionally, different frameworks include varying degrees of clarity around the ethical principles they suggest. For example, data statements are a somewhat neutral paradigm that promotes documentation and reflection on the production of a dataset, which is useful, but the framework does not require practitioners to explicitly comment on political or ethical commitments (Scheuerman et al., 2021). Documentation is just one step in the process towards normatively evaluating a dataset, and provides a basis upon which to ask critical questions (Bender and Friedman, 2018). In general, frameworks that promote reflection and transparency are essential as foundations for further evaluating appropriate uses for a system.

Examining the design process is important, but Shilton (2018) advocates asking whether designers even have much agency over the trajectory of their artifacts. That is, even if we carefully and thoughtfully design a system, we still ultimately do not have control over how it will be used in the world, and our imaginations are fundamentally limited.

We have shown how technology has the potential to reinforce or exacerbate power imbalances when it enters into an existing social and political context. To complete the picture, an interdisciplinary lens is required, as well as more user-studies of NLP systems 'in the wild' (e.g. Liebling et al. (2020)). Interestingly, NAACL 2022's announced theme is 'Human-centered NLP' which reflects burgeoning efforts to address the 'human' in NLP.

Model architectures may come and go, but the question of what problems to work on, who is included in this process, and how to evaluate performance and release systems into the world requires consistent reflection and accountability. However, the choice of model

architecture is not trivial, either, because this too shapes the impact of system development. Data-intensive and compute-intensive approaches have real impact on the physical world (Bender et al., 2021), and this infrastructure is often invisibilized, imagined as 'a cloud.' In the age of digital reproduction, it seems trivial to download and reuse data, and to fine-tune a model under various conditions, but each computation expends energy and relies on a vast physical infrastructure (Strubell et al., 2019).

8.3 Reflections and proposals for future work

As sketched out in the introduction, NLP technology is not created in, nor will it enter, a social vacuum – every step of the process reflects human realities. I opened with a quote from Franklin (1999) in which she said that technologies may 'enforce or destroy' a social structure, often 'in ways that are neither foreseen nor foreseeable.' I agree that human ingenuity is boundless, and have witnessed and participated firsthand in the use of technology far beyond the designers' original intentions. However, I argue that we can look to history and sociology to understand the material contexts that shape what might happen with the technologies we build — to foresee some of the likely outcomes. The environment that NLP technology enters into is one shaped by patterns of discrimination and injustice, of the 'matrix of domination' (Costanza-Chock, 2020). We must be aware of this, and, even better, strategize ways to correct these injustices. As a thought exercise, we can think of the worst possible use for something we build, and decide whether we are OK with that possibility.

One of Wagstaff (2012)'s 'impact' challenges posed for machine learning was 'a conflict between nations averted through high-quality translation provided by an ML system.' It is hard to imagine validating such a scenario without a counterfactual. However, we have seen some worst case scenarios for machine translation already. Researchers often point to the case of a Palestinian man in Israel wrongfully arrested based on a mis-translated

Facebook post[1]. He had posted a picture of a bulldozer and the caption 'Good morning' in Arabic, which was mis-translated via an automatic API into Hebrew as 'Attack them.' Concerned citizens alerted the police, who cited a previous incident in which a Palestinian man killed a pedestrian with a construction vehicle as cause for investigation. Would 'quality translation' have averted this violation? I argue that the answer is no. This situation was already incredibly loaded with social prejudices and anxieties, and any sort of mild misunderstanding was likely to be taken out of context. Even if the translation was flawless, there are idioms, sarcasm, and other forms of expression that can be willfully misread, even if 'correctly' translated. As an illustrative example, an English tourist was denied entry into the United States after DHS agents viewed his social media and cited a Twitter post saying he would 'destroy America' as grounds for rejecting his entry into the country (Patel et al., 2019). This was not a mistranslation, but a misreading of an idiom written in English, presumably a language the US DHS agents are proficient in. Thus, while improving machine translation is a noble goal, we cannot pretend that better translation will lead to utopian ideals of mutual understanding.

As the machine translation case study illustrated, NLP tools have been recruited into advanced techniques for surveillance. For those who have observed who has power in the world and what incentives they are driven by, abuse of this tool comes as no surprise. NLP experts have a key role to play in working with human rights advocates to identify cases of misuse, and more broadly, to inform policies for auditing and regulating language technology.

While there is increased sensitivity to issues of diversity in NLP systems, it is my fear that the benefits of development of NLP for so-called 'low resource' languages will ultimately accrue to the companies collecting the data and providing the infrastructure for this technology, and to those wealthiest in society who can afford to use the gadgets prerequisite to this technology. History has shown how technology developed in one context

[1]https://www.jpost.com/Arab-Israeli-Conflict/Palestinian-arrested-after-police-rely-on-mistranslation-of-Facebook-post-508107

does not travel well to other contexts (Irani et al., 2010). It should be noted that I am working within and focusing on technological development in the sociopolitical context of the United States, which is shaped by a history involving the displacement of indigenous peoples and enslavement of Africans. The technological imagination of those working in this context is thus shaped by these conditions, and thus my analysis and the set of issues that I am sensitive to is necessarily limited by my position.

8.4 Research questions revisited

We conclude by revisiting the research questions guiding the case studies.

- What has been the social impact of data-driven natural language processing technologies?
  - NLP technologies have facilitated access to information, in applications that range from assisting researchers in curation processes (Chapters 6 and 7) to powering translation between languages (Chapter 5). However, access to information is a double-edged sword, and when NLP enters into social contexts shaped by power imbalances, there is potential for harm. Additionally, the data collection processes that power modern NLP have ethical and legal implications for the rights of human data subjects and data workers (Chapter 3).
- What role does the design and development of these systems play in shaping that impact?
  - The choices made at every stage in the development process, ranging from the formulation of questions that inform data selection, to the incentives guiding model development practices, all impact the world outside the lab. For example, the research decision to rely on a web-scraping paradigm for data has led, in some cases, to public backlash when personal data is used in unexpected machine learning applications (Chapter 3).

  - Failing to include relevant stakeholders in the earliest stages of system development can lead to a rejection of the ensuing application (Chapter 5).

- What considerations should be made during the development process to anticipate and address this impact?

  - Researchers should strive for an awareness of the historical, social, and political contexts in which their work is developed and applied, which can help them make sense of data discrepancies and anticipate deployment outcomes.
  - Researchers should also gain literacy in the variety of frameworks that can help promote structured thinking about potential impacts of design choices throughout the development cycle (Sim et al., 2021).

Chapter 9
CONCLUSION

This work has explored some of the practical and ethical issues in developing and releasing natural language processing systems. I have structured this exploration around the stages of system development, in particular the conception, design, evaluation, and deployment of these systems, while applying a sociotechnical lens to these analyses. I have illustrated, through a handful of case studies, some of the considerations for the direct and indirect stakeholders of actions taken at various points in the NLP system development pipeline. We explored how current data practices have often flattened or obscured the social relations underlying their development, as well as the practical pitfalls that stem from the collection processes of data sets.

In light of observations of trends in the field and through my own work, I suggest that NLP researchers engage more meaningfully with critical questions about what applications we wish to power with our work, how these problems are situated in their broader social and political contexts, and who truly stands to benefit (or incur harms) from practices of data collection and system deployment. Promoting interdisciplinary work that emphasizes contributions from a variety of voices will continue to be vital in the future if we wish to have a positive impact. We should also strive to gain literacy in design frameworks that can inform our work.

I summarize my contributions as follows. The first case study (Chapter 4), on computational models of text, illustrated the impact of data selection choices for training such models. Researchers should justify training data curation choices by tying them to particular use contexts, distinguishing between prescriptive and descriptive uses of these models.

The second case study (Chapter 5), on machine translation, examined the social impact of machine translation due to its historical conception as a tool for surveillance. While

translation errors can cause harm, we should also be attentive to the power imbalances that shape scenarios that are inappropriate for the use of machine translation. In developing systems for 'low-resource' languages, we should consider issues of trust and consent in light of the historical inequities that lead to language endangerment.

In the third case study (Chapters 6 and 7) I reported results on evaluating methods for representing biomedical relationships learned from literature. In reflecting on the limitations of this study, I find that the needs of curators should inform the design of the system. In future work, I will more closely examine the commonalities across curation practices for different data resources, including frequently used resources and workflows and surfacing recurring pain points. I note that the body of literature in biomedicine is shaped by historical injustices and disparities; any system trained as such should come with caveats, and future work should focus on correcting these disparities.

Across each case study, a few patterns are apparent. Lack of informed consent, for using and re-using image data, medical data, linguistic data, erodes public trust and shows a lack of respect for human subjects. Failing to consider stakeholders, such as language speakers, curators of data bases, can lead to a rejection of system deployment (a failure to integrate) and a failure to anticipate the amplification of disparities in data. We should reconsider the goals & incentives that structure our research practices, and broaden where we look for guidance on this process by incorporating interdisciplinary perspectives and direct consultation with stakeholders.

The majority of my training thus far has been as a practitioner and researcher in NLP, and this dissertation reflects my active, ongoing engagement with critical, reflexive practices from scholarship in science and technology studies and critical data studies. I still have far to go in synthesizing these traditions to inform my future work.

BIBLIOGRAPHY

Philip E. Agre. Toward a critical technical practice: Lessons learned in trying to reform AI. In Geof Bowker, Les Gasser, Leigh Star, and Bill Turner, editors, *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*. Erlbaum, 1997.

Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy's new clothes. *medium. com*, 2017. URL `https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a`.

Blaise Agüera y Arcas, Alexander Todorov, and Margaret Mitchell. Do algorithms reveal sexual orientation or just expose our stereotypes? *medium. com*, 2018. URL `https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477`.

Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang. Assisted curation: does text mining really help? In *Biocomputing 2008*, pages 556–567. World Scientific, 2008.

Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798, 2015.

Ulla Maria Anderberg, Ina Marteinsdottir, and Lars von Knorring. Citalopram in patients with fibromyalgia—a randomized, double-blind, placebo-controlled study. *European journal of pain*, 4(1):27–35, 2000.

Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564. URL https://aaai.org/ojs/index.php/aimagazine/article/view/2564.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014.

Andrea Bear Nicholas. Linguicide: Submersion education and the killing of languages in canada. *Briarpatch Magazine*, 40(4):8, 2011.

Tanja Bekhuis. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical digital libraries*, 3(1):1–7, 2006.

Anja Belz and Adam Kilgarriff. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 133–135, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W06-1421.

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://www.aclweb.org/anthology/Q18-1041.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://www.aclweb.org/anthology/2020.acl-main.463.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of FAccT 2021*, 2021.

Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. Towards Standardization of Data Licenses: The Montreal Data License. *arXiv:1903.12262 [cs, stat]*, March 2019.

Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, 2019.

Janine Berg. Income security in the on-demand economy : findings and policy lessons from a survey of crowdworkers. Ilo working papers, International Labour Organization, 2016.

T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D.A. Forsyth. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages 848–854, Washington, DC, USA, 2004. IEEE. ISBN 978-0-7695-2158-9. doi: 10.1109/CVPR.2004.1315253.

Adam Berger. *Statistical machine learning for information retrieval*. Carnegie Mellon University, 2001.

Abeba Birhane. Algorithmic colonization of africa. *SCRIPTed*, 17:389, 2020.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1547, 2021.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

Geoffrey C Bowker. *Memory practices in the sciences*, volume 205. Mit Press Cambridge, MA, 2005.

Lynne Bowker and Jairo Buitrago Ciro. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Group Publishing, 2019.

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. Overcoming failures of imagination in ai infused system development and deployment. *arXiv preprint arXiv:2011.13416*, 2020.

danah boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5): 662–679, 2012.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

Sara Brown. The hidden work created by artificial intelligence programs. Retrieved from `https://mitsloan.mit.edu/ideas-made-to-matter/hidden-work-created-artificial-intelligence-programs` (2021/11/23), 2021.

Peter Bruza and Marc Weeber. *Literature-based discovery*. Springer Science & Business Media, 2008.

Michael Buckland. *Information and society*. MIT Press, 2017.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of Machine Learning Research*, volume 81, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Chris Callison-Burch. Crowd-workers: Aggregating information across turkers to help them find higher paying work. In *HCOMP*, 2014.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*, 2021.

Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4, 2020a.

Qingyu Chen, Kyubum Lee, Shankai Yan, Sun Kim, Chih-Hsuan Wei, and Zhiyong Lu. Bioconceptvec: creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS computational biology*, 16(4):e1007617, 2020b.

Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. The dataset nutrition label (2nd gen): Leveraging con-

text to mitigate harms in artificial intelligence. NeurIPS Workshop on Dataset Curation and Security, 2020. URL http://securedata.lol/.

Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company, 2020.

Kevin Bretonnel Cohen and Dina Demner-Fushman. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company, 2014.

Trevor Cohen and Dominic Widdows. Embedding of semantic predications. *Journal of biomedical informatics*, 68:150–166, 2017.

Trevor Cohen, G Kerr Whitfield, Roger W Schvaneveldt, Kavitha Mukund, and Thomas Rindflesch. Epiphanet: an interactive tool to support biomedical discoveries. *Journal of biomedical discovery and collaboration*, 5:21, 2010.

Trevor Cohen, Dominic Widdows, Roger Schvaneveldt, and Thomas C Rindflesch. Finding schizophrenia's prozac emergent relational similarity in predication space. In *International Symposium on Quantum Interaction*, pages 48–59. Springer, 2011.

Trevor Cohen, Dominic Widdows, Lance De Vine, Roger Schvaneveldt, and Thomas C Rindflesch. Many paths lead to discovery: analogical retrieval of cancer therapies. In *International Symposium on Quantum Interaction*, pages 90–101. Springer, 2012.

Samantha Cole. This horrifying app undresses a photo of any woman with a single click. *Vice*, (27):06, 2019.

B Jack Copeland. *The essential turing*. Clarendon Press, 2004.

Elif Kiesow Cortez. Data protection around the world. *Privacy laws in action*.

Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

Nick Couldry and Ulises A Mejias. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4):336–349, 2019.

Kate Crawford. *"The Trouble with Bias"*, 2017. URL https://www.youtube.com/watch?v=fMym_BKWQzk. NeurIPS keynote.

Kate Crawford and Trevor Paglen. *Excavating AI: The Politics of Images in Machine Learning Training Sets*, 2019. URL https://www.excavating.ai/.

Kate Crawford, Mary Gray, and Kate Miltner. Big data— critiquing big data: Politics, ethics, epistemology — special section introduction. *International Journal of Communication*, 8 (0), 2014. ISSN 1932-8036. URL https://ijoc.org/index.php/ijoc/article/view/2167.

Marie-Catherine De Marneffe and Christopher D Manning. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics, 2008.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

Emily L. Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. In *Proceedings of the Participatory Approaches to Machine Learning Workshop*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186,

Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE, 2019. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.html.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://www.aclweb.org/anthology/D19-1224.

Kemi M Doll. Investigating black-white disparities in gynecologic oncology: theories, conceptual models, and applications. *Gynecologic oncology*, 149(1):78–83, 2018.

Kemi M Doll, Sara Khor, Katherine Odem-Davis, Hao He, Erika M Wolff, David R Flum, Scott D Ramsey, and Barbara A Goff. Role of bleeding recognition and evaluation in black-white disparities in endometrial cancer. *American journal of obstetrics and gynecology*, 219(6):593–e1, 2018.

Mimi Ọnụọha. The Point of Collection. *Points*, 2016. URL https://points.datasociety.net/the-point-of-collection-8ee44ad7c2fa.

Joseph Donia, James Shaw, et al. Ethics and values in design: A structured review and theoretical critique. *Science and Engineering Ethics*, 27(5):1–32, 2021.

Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 294, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3373157. URL https://doi.org/10.1145/3351095.3373157.

Paul N. Edwards. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Infrastructures Series. The MIT Press, Cambridge, Massachusetts London, England, first paperback edition edition, 2013. ISBN 978-0-262-51863-5 978-0-262-01392-5.

Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.

Kacey Y Eichelberger, Kemi Doll, Geraldine E Ekpo, and Matthew L Zerden. Black lives matter: claiming a space for evidence-based outrage in obstetrics and gynecology, 2016.

Joseph Errington. Colonial linguistics. *Annual Review of Anthropology*, 30:19–39, 2001.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of nlp leaderboards. In *arXiv:2009.13888*, 2020.

Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-009-0275-4.

Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487, 2016.

A Fazzari, Maria Graziella Catalano, A Comba, M Becchis, M Raineri, Roberto Frairia, and N Fortunati. The control of progesterone receptor expression in mcf-7 breast cancer cells: effects of estradiol and sex hormone-binding globulin (shbg). *Molecular and cellular endocrinology*, 172(1-2):31–36, 2001.

Eric S Felber. Botulinum toxin in primary care medicine. *Journal of Osteopathic Medicine*, 106(10):609–614, 2006.

Sergey Feldman, Waleed Ammar, Kyle Lo, Elly Trepman, Madeleine van Zuylen, and Oren Etzioni. Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA network open*, 2(7):e196700–e196700, 2019.

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Tracy L Fischer, John A Pieper, Donald W Graff, Jo E Rodgers, Jeffrey D Fischer, Kimberly J Parnell, Joyce A Goldstein, Robert Greenwood, and J Herbert Patterson. Evaluation of potential losartan-phenytoin drug interactions in healthy volunteers. *Clinical Pharmacology & Therapeutics*, 72(3):238–246, 2002.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dan-

gana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.195.

Diana E Forsythe. The construction of work in artificial intelligence. *Science, technology, & human values*, 18(4):460–479, 1993.

Diana E Forsythe. New bottles, old wine: hidden cultural assumptions in a computerized explanation system for migraine sufferers. *Medical anthropology quarterly*, 10(4):551–574, 1996.

Marion Fourcade and Kieran Healy. Seeing like a market. *Socio-economic review*, 15(1): 9–29, 2017.

Ursula Franklin. *The real world of technology*. House of Anansi, 1999.

Susan Gal. Politics of translation. *Annual Review of Anthropology*, 44:225–240, 2015.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. On making reading comprehension more comprehensive. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 105–112, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5815. URL https://www.aclweb.org/anthology/D19-5815.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F.

Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.117.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Robert W Gehl, Lucas Moyer-Horner, and Sara K Yeo. Training computers to see internet pornography: Gender and sexual discrimination in computer vision science. *Television & New Media*, 18(6):529–547, 2017.

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372862. URL https://doi.org/10.1145/3351095.3372862.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Andrew Gelman, Greggor Mattson, and Daniel Simpson. Gaydar and the fallacy of decon-

textualized measurement. *Sociological Science*, 5(12):270–280, 2018. ISSN 2330-6696. doi: 10.15195/v5.a12. URL http://dx.doi.org/10.15195/v5.a12.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL https://www.aclweb.org/anthology/D19-1107.

Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, and Klaus Mueller. Measuring social biases of crowd workers using counterfactual queries. *arXiv preprint arXiv:2004.02028*, 2020.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL-HLT SRW*, pages 47–54, San Diego, California, June 12-17, 2016, 2016. ACL. doi: 10.18653/v1/N16-2002. URL https://www.aclweb.org/anthology/N/N16/N16-2002.pdf.

Kevin Gold. Norvig vs. chomsky and the fight for the future of ai, 2011. URL https://www.tor.com/2011/06/21/norvig-vs-chomsky-and-the-fight-for-the-future-of-ai/.

Mary L. Gray and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517, 2005.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, 2020.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Adam Harvey and Jules LaPlace. MegaPixels: Origins and endpoints of biometric datasets "In the Wild". https://megapixels.cc, 2019.

He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6115. URL https://www.aclweb.org/anthology/D19-6115.

Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *ACM Future of Computing Blog*, 2018. URL https://acm-fca.org/2018/03/29/negativeimpacts/.

Benjamin Heinzerling. Nlp's clever hans moment has arrived. *The Gradient*, 2019.

Miguel Helft. Google's Computing Power Refines Translation Tool. *The New York Times*, March 8 2010. URL https://www.nytimes.com/2010/03/09/technology/09translate.html.

Sam Henry and Bridget T McInnes. Literature based discovery: models, methods, and trends. *Journal of biomedical informatics*, 74:20–32, 2017.

Hendrik Heuer and Daniel Buschek. Methods for the design and evaluation of hci+nlp systems. *arXiv preprint arXiv:2102.13461*, 2021.

Anna Lauren Hoffmann. Terms of inclusion: Data, discourse, violence. *new media & society*, page 1461444820958725, 2020.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.345.

Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016.

Dimitar Hristovski, Carol Friedman, Thomas C Rindflesch, and Borut Peterlin. Exploiting semantic relations for literature-based discovery. In *AMIA annual symposium proceedings*, volume 2006, page 349. American Medical Informatics Association, 2006.

Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300637. URL https://doi.org/10.1145/3290605.3300637.

John Hutchins. From first conception to first demonstration: the nascent years of machine translation, 1947–1954. a chronology. *Machine Translation*, 12(3):195–252, 1997.

W John Hutchins. *Early years in machine translation: memoirs and biographies of pioneers*, volume 97. John Benjamins Publishing, 2000.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Craig Denuyl. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of ACL 2020*, 2020.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2021.

Lilly Irani. The cultural work of microwork. *New Media & Society*, 17(5):720–739, 2015a.

Lilly Irani. Difference and dependence among digital workers: The case of amazon mechanical turk. *South Atlantic Quarterly*, 114(1):225–234, 2015b.

Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1311–1320, 2010.

Lilly C. Irani and M. Six Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 611–620. Association for Computing Machinery, 2013. ISBN 9781450318990.

Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*, 2019.

Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372829. URL https://doi.org/10.1145/3351095.3372829.

Pentti Kanerva. Binary spatter-coding of ordered k-tuples. In *International Conference on Artificial Neural Networks*, pages 869–873. Springer, 1996.

Rehab A Karam, Haidy E Zidan, and Mohamed Hamed Khater. Polymorphisms in the tnf-α and il-10 gene promoters and risk of psoriasis and correlation with disease severity. *Cytokine*, 66(2):101–105, 2014.

Yarden Katz. Manufacturing an artificial intelligence revolution. *Available at SSRN 3078224*, 2017.

Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1546. URL https://www.aclweb.org/anthology/D18-1546.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2020.

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2021.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. Broad-coverage biomedical relation extraction with semrep. *BMC bioinformatics*, 21:1–28, 2020.

Abdelfattah Kilito. *Thou Shalt Not Speak My Language*. Syracuse University Press, 2008.

Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

John Kuchtey, Ta Chen Chang, Lampros Panagis, and Rachel W Kuchtey. Marfan syndrome caused by a novel fbn1 mutation with associated pigmentary glaucoma. *American Journal of Medical Genetics Part A*, 161(4):880–883, 2013.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

Claire Larsonneur. The disruptions of neural machine translation. *spheres: Journal for Digital Cultures*, 5:1–10, 2019.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. *International Conference on Machine Learning (ICML)*, 2020.

Claire Leibowicz, Emily Saltz, and Lia Coleman. Creating ai art responsibly: A field guide for artists. *Diseña*, (19):5–5, 2021.

James Lennon. If you're de-biasing the model, it's too late, 2020. URL https://scale.com/blog/if-youre-de-biasing-the-model-its-too-late.

Amanda Levendowski. How copyright law can fix artificial intelligence's implicit bias problem. *Wash. L. Rev.*, 93:579, 2018.

Hector J Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180, 2014.

Gideon Lewis-Kraus. The Great A.I. Awakening. *New York Times Magazine*, December 14 2016. URL https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

Xiaochang Li. *Divination Engines: A Media History of Text Prediction*. PhD thesis, New York University, 2017.

Daniel J Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. Unmet needs and opportunities for mobile translation ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Tal Linzen. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2503. URL https://www.aclweb.org/anthology/W16-2503.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 1 (Long and Short Papers), pages 2171–2179, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1225. URL https://www.aclweb.org/anthology/N19-1225.

William Nash Locke and Andrew Donald Booth. Insights from Tom Pedtke: text of MT Summit keynote address. *Language Today*, 6:6–15, March 1998.

Hoyt Long. *The Values in Numbers: Reading Japanese Literature in a Global Information Age*. Columbia University Press, 2021.

Sean Macavaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. Community-level research on suicidality prediction in a secure environment: Overview of the clpsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, 2021.

Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

Iain James Marshall, Veline L'Esperance, Rachel Marshall, James Thomas, Anna Noel-Storr, Frank Soboczenski, Benjamin Nye, Ani Nenkova, and Byron C Wallace. State of the evidence: a survey of global disparities in clinical trials. *BMJ Global Health*, 6(1):e004145, 2021.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 233–243, 2019.

Anna J Matheson and Caroline M Spencer. Ropinirole. *Drugs*, 60(1):115–137, 2000.

Ryan Merkley. Use and Fair Use: Statement on shared images in facial recognition AI, March 2019.

Jacob Metcalf and Kate Crawford. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211, 2016.

Milagros Miceli, Martin Schuessler, and Tianling Yang. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), October 2020. doi: 10.1145/3415186. URL https://doi.org/10.1145/3415186.

Julian Michael. To dissect an octopus: Making sense of the form/meaning debate. Blog post, 2020. URL https://blog.julianmichael.org/2020/07/23/to-dissect-an-octopus.html.

Shotaro Michinaga, Akinori Hisatsune, Yoichiro Isohama, and Hiroshi Katsuki. An anti-parkinson drug ropinirole depletes orexin from rat hypothalamic slice culture. *Neuroscience research*, 68(4):315–321, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2013a. URL https://arxiv.org/pdf/1301.3781.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1090.

Ishan Misra, C. Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards

for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, pages 1–26, 2020.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

Marissa Moorman. Can an algorithm be racist? Africa Is A Country, blog, 2014. URL https://africasacountry.com/2014/09/can-an-algorithm-be-racist/.

Madhumita Murgia. Who's using your face? The ugly truth about facial recognition. *Financial Times*, September 2019.

Rahul Nadkarni, David Wadden, Iz Beltagy, Noah A Smith, Hannaneh Hajishirzi, and Tom Hope. Scientific language models for biomedical knowledge base completion: An empirical study. *arXiv preprint arXiv:2106.09700*, 2021.

Timothy Niven and Hung-Yu Kao. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL https://www.aclweb.org/anthology/P19-1459.

Safiya Umoja Noble. *Algorithms of oppression*. New York University Press, 2018.

Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732, 2021.

Jason D Oliver, H Llewelyn Roderick, David H Llewellyn, and Stephen High. Erp57 functions as a subunit of specific complexes formed with the er lectins calreticulin and calnexin. *Molecular biology of the cell*, 10(8):2573–2582, 1999.

Margaret Pugh O'Mara. *The Code: Silicon Valley and the Remaking of America*. Penguin Press, 2019.

Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

Liz O'Sullivan. Don't steal data. NeurIPS Workshop on Dataset Curation and Security, 2020. URL `http://securedata.lol/`.

Ben Packer, M. Mitchell, Mario Guajardo-Céspedes, and Yoni Halpern. Text embeddings contain bias. Here's why that matters. Technical report, Google, 2018.

Soon-Yong Pak and Keumjoong Hwang. Assimilation and segregation of imperial subjects:"educating" the colonised during the 1910–1945 japanese colonial rule of korea. *Paedagogica Historica*, 47(3):377–397, 2011.

Despoina Panou. Equivalence in translation theories: A critical evaluation. *Theory and Practice in Language Studies*, 3(1):1, 2013.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, 2018.

Irene V. Pasquetto, Bernadette M. Randles, and Christine L. Borgman. On the Reuse of Scientific Data. *Data Science Journal*, 16:8, March 2017. ISSN 1683-1470. doi: 10.5334/dsj-2017-008.

Faiza Patel, Rachel Levinson-Waldman, Raya Koreh, and Sophia DenUyl. Social media monitoring. *Brennan Center for Justice*, 2019. URL `https://www.brennancenter.org/our-work/research-reports/social-media-monitoring`.

Amandalynne Paullada. Machine translation shifts power. *The Gradient*, 2021.

Amandalynne Paullada, Bethany Percha, and Trevor Cohen. Improving Biomedical Analogical Retrieval with Embedding of Structural Dependencies. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 38–48. Association for Computational Linguistics, 2020. URL `https://www.aclweb.org/anthology/2020.bionlp-1.4`.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily L. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2, 2021.

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

Kenneth L Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Alastair Pennycook. The modern mission: The language effects of christianity. *Journal of Language, Identity, and Education*, 4(2):137–155, 2005.

Bethany Percha. Modern clinical text mining: A guide and review. *Annual Review of Biomedical Data Science*, 4, 2021.

Bethany Percha and Russ B Altman. Learning the structure of biomedical relationships from unstructured text. *PLoS Comput Biol*, 11(7):e1004216, 2015.

Bethany Percha and Russ B Altman. A global network of biomedical relationships derived from text. *Bioinformatics*, 34(15):2614–2624, 2018.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1179. URL https://www.aclweb.org/anthology/D18-1179.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Mohammad Taher Pilehvar and Jose Camacho-Collados. Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4):1–175, 2020.

Stelios Piperidis. Machine translation and its philosophical accounts. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 1–1, 2009.

Everest Pipkin. On lacework: watching an entire machine-learning dataset. *unthinking.photography*, July 2020.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://www.aclweb.org/anthology/S18-2023.

Eric A Posner and E Glen Weyl. *Radical Markets*. Princeton University Press, 2019.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32 (10):6363–6381, 2020.

Precarity Lab. *Technoprecarious*. Goldsmiths Press, 2020.

Judita Preiss. Seeking informativeness in literature based discovery. In *Proceedings of BioNLP 2014*, pages 112–117, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3417. URL https://aclanthology.org/W14-3417.

Willard Van Orman Quine. *Word and object*. MIT press, 2013.

Joanna Radin. "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. *Osiris*, 32(1):43–64, September 2017. ISSN 0369-7827, 1933-8287. doi: 10.1086/693853.

Deborah Raji, Anna Lauren Hoffmann, Nyalleng Moorosi, Vinay Prabhu, Jake Metcalf, and Sherry Stanley. Panel discussion: Harms from ai research. NeurIPS Workshop on Navigating the Broader Impacts of AI Research, 2020. URL https://nbiair.com/.

Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. *arXiv preprint arXiv:2102.00813*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of Machine Learning Research*, volume 97, pages 5389–5400, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Yim Register and Amy J Ko. Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics. In *Proceedings of the 2020 ACM Conference on International Computing Education Research*, pages 67–78, 2020.

Daniel R Rhodes, Bushra Ateeq, Qi Cao, Scott A Tomlins, Rohit Mehra, Bharathi Laxman, Shanker Kalyana-Sundaram, Robert J Lonigro, Beth E Helgeson, Mahaveer S Bhojani, et al. Agtr1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an agtr1 antagonist. *Proceedings of the National Academy of Sciences*, 106(25): 10284–10289, 2009.

Neil M Richards and Jonathan H King. Big data ethics. *Wake Forest L. Rev.*, 49:393, 2014.

Emma Rodman. A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, 28(1):87–111, 2020.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8: 842–866, 2020.

Alice Rulcova, Lucie Krausova, Tomas Smutny, Radim Vrzal, Zdenek Dvorak, Ramiro Jover, and Petr Pavek. Glucocorticoid receptor regulates organic cation transporter 1 (oct1, slc22a1) expression via hnf4α upregulation in primary human hepatocytes. *Pharmacological Reports*, 65(5):1322–1335, 2013.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of*

Computer Vision, 115(3):211–252, December 2015. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-015-0816-y.

Jathan Sadowski. *Too Smart: How Digital Capitalism is Extracting Data, Controlling Our Lives, and Taking Over the World*. MIT Press, 2020.

Matthew Sag. The new legal landscape for text mining and machine learning. *Journal of the Copyright Society of the USA*, 66, 2019.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, 2020.

Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1621–1630, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://www.aclweb.org/anthology/P19-1163.

David Sarno. Need a translation? Google awaits your call. *The Los Angeles Times*, March 8 2010a. URL https://www.latimes.com/archives/la-xpm-2010-mar-08-la-fi-google-translate9-2010mar09-story.html.

David Sarno. Franz Josef Och, Google's translation uber-scientist, talks about Google Translate. *The Los Angeles Times*, March 11 2010b. URL https://latimesblogs.latimes.com/technology/2010/03/the-web-site-translategooglecom-was-done-in-2001-we-were-just--licensing-3rd-party-machine-translation-technologies-tha.html.

Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020. doi: 10.1145/3392866. URL https://doi.org/10.1145/3392866.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. 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 670–674, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.85. URL https://aclanthology.org/2021.acl-short.85.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. Beyond leaderboards: A survey of methods for revealing weaknesses in natural language inference data and models. *arXiv preprint arXiv:2005.14709*, 2020.

Tyler Schnoebelen. Goal-oriented design for ethical machine learning and nlp. In *Proceed-*

ings of the First ACL Workshop on Ethics in Natural Language Processing, pages 88–93, 2017.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. 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 3410–3416, 2019.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019. URL `http://arxiv.org/abs/1907.10597`.

James C Scott. *Seeing like a state*. Yale University Press, 2008.

D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and A. Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *ICLR*, 2018.

Andrew D Selbst, danah boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

Alana Semuels. The internet is enabling a new kind of poorly paid hell. *The Atlantic*, 2018. URL `https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/`.

Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. Turkers, scholars, "Arafat" and "peace": Cultural communities and algorithmic gold standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 826–838, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450329224.

Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pages 49–58, 2005.

Naomi Shapiro, Amandalynne Paullada, and Shane Steinert-Threlkeld. A multilabel approach to morphosyntactic probing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4486–4524, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.382.

Katie Shilton. Values and ethics in human-computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 12(2), 2018.

M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. Responsible research with crowds: Pay crowdworkers at least minimum wage. *Commun. ACM*, 61 (3):39–41, February 2018. ISSN 0001-0782. doi: 10.1145/3180492. URL https://doi.org/10.1145/3180492.

Kate Sim, Andrew Brown, and Amelia Hassoun. Thinking through and writing about research ethics beyond" broader impact". *arXiv preprint arXiv:2104.08205*, 2021.

Rebecca Skloot. *The immortal life of Henrietta Lacks*. Broadway Paperbacks, 2017.

Neil R Smalheiser. Rediscovering Don Swanson: The past, present and future of literature-based discovery. *Journal of data and information science (Warsaw, Poland)*, 2(4):43, 2017.

Olivia Solon. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. In *NBC News*, 2019.

Margaret Speas. Language ownership and language ideologies. In Laetitia La Follette, editor, *Negotiating Culture: Heritage, Ownership, and Intellectual Property*, pages 101–121. University of Massachusetts Press, Amherst, 2013.

Robyn Speer. How to make a racist ai without really trying. Retrieved from http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/, July 2017.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. *International Conference on Machine Learning (ICML)*, 2020.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL `https://www.aclweb.org/anthology/P19-1164`.

Susan Leigh Star and Anselm Strauss. Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW)*, 8(1):9–30, 1999.

Victoria Stodden. The data science life cycle: A disciplined approach to advancing data science as a science. *Communications of the ACM*, 63(7):58–66, June 2020. ISSN 0001-0782, 1557-7317. doi: 10.1145/3360646.

Victoria Stodden and Sheila Miguez. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software*, 2(1):e21, July 2014. ISSN 2049-9647. doi: 10.5334/jors.ay.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

Bruno J. Strasser and Paul N. Edwards. Big Data Is the Answer . . . But What Is the Ques-

tion? *Osiris*, 32(1):328–345, September 2017. ISSN 0369-7827, 1933-8287. doi: 10.1086/694223.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

Lucy Suchman. Making work visible. *Commun. ACM*, 38(9):56–64, September 1995. ISSN 0001-0782. doi: 10.1145/223248.223263. URL https://doi.org/10.1145/223248.223263.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9. 2021.

Don R Swanson. Searching natural language text by computer. *Science*, 132(3434):1099–1104, 1960.

Don R Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986a.

Don R Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986b.

Don R Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557, 1988.

Don R Swanson. Intervening in the life cycles of scientific knowledge. 1993.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP*, 2020. URL https://arxiv.org/abs/2009.10795.

Rachael Tatman and Conner Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech*, pages 934–938, 2017.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*, 2020a.

Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020b.

Carol Tenopir, Donald W King, Sheri Edwards, and Lei Wu. Electronic journals and changes in scholarly article seeking and reading patterns. In *Aslib proceedings*. Emerald Group Publishing Limited, 2009.

Menasha Thilakaratne, Katrina Falkner, and Thushari Atapattu. A systematic review on literature-based discovery: General overview, methodology, & statistical analysis. *ACM Computing Surveys (CSUR)*, 52(6):1–34, 2019.

Rachel Thomas. Medicine's machine learning problem. In Daren Acemoglu, editor, *Redesigning AI*, volume 18 of *Boston Review Forum*, pages 127–138. Boston Review, 2021.

Jer Thorp. *Living In Data*. MCD, 2021.

Yeganeh Torbani. Google Says Google Translate Can't Replace Human Translators. Immigration Officials Have Used It to Vet Refugees. *ProPublica*, 2019. URL https://www.propublica.org/article/google-says-google-translate-cant-replace-human-translators-immigration-officials-have-used-it-to-vet-refugees.

John C Torpey. *The invention of the passport: Surveillance, citizenship and the state*. Cambridge University Press, 2018.

Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, November 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.128. URL https://doi.org/10.1109/TPAMI.2008.128.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. *International Conference on Machine Learning (ICML)*, 2020.

Peter D Turney and Michael L Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1):251–278, 2005.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

Ramya Vajapey, David Rini, Jeremy Walston, and Peter Abadir. The impact of age-related dysregulation of the angiotensin system on mitochondrial redox balance. *Frontiers in physiology*, 5:439, 2014.

Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset. In *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*, pages 1–4, 2016.

Joaquin Vanschoren and Serena Yeung. Announcing the neurips 2021 datasets and benchmarks track, 2021. URL https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

Salomé Viljoen. Democratic data: A relational theory for data governance. *Available at SSRN 3727562*, 2020.

Nicholas Vincent, Brent Hecht, and Shilad Sen. "data strikes": Evaluating the effectiveness of a new form of collective action against technology companies. In *The World Wide Web Conference*, pages 1931–1943, 2019.

Kiri Wagstaff. Machine learning that matters. *arXiv preprint arXiv:1206.4656*, 2012.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. *European Conference on Computer Vision (ECCV)*, 2020a.

Lucy Wang, Oyvind Tafjord, Arman Cohan, Sarthak Jain, Sam Skjonsberg, Carissa Schoenick, Nick Botner, and Waleed Ammar. Supp. ai: finding evidence for supplement-drug interactions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 362–371, 2020b.

Lucy Lu Wang and Kyle Lo. Text mining approaches for dealing with the rapidly expanding literature on covid-19. *Briefings in Bioinformatics*, 22(2):781–799, 2021.

Yunxia Wang, Song Zhang, Fengcheng Li, Ying Zhou, Ying Zhang, Zhengwen Wang, Runyuan Zhang, Jiang Zhu, Yuxiang Ren, Ying Tan, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic acids research*, 48(D1):D1031–D1041, 2020c.

Harriet A Washington. *Medical apartheid: The dark history of medical experimentation on Black Americans from colonial times to the present*. Doubleday Books, 2006.

Warren Weaver. Translation. *Machine translation of languages*, 14(15-23):10, 1955.

Marc Weeber, Henny Klein, Lolkje TW de Jong-van den Berg, and Rein Vos. Using concepts in literature-based discovery: Simulating swanson's raynaud–fish oil and migraine–magnesium discoveries. *Journal of the american society for information science and technology*, 52(7):548–557, 2001.

E Hope Weissler, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F Freitag, James Benoit, Michael C Hughes, Faisal Khan, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1):1–15, 2021.

Michelle Whirl-Carrillo, Ellen M McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, Russ B Altman, and Teri E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012.

Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. Fair work: Crowd work minimum wage with one line of code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):197–206, Oct. 2019.

Dominic Widdows and Trevor Cohen. Real, complex, and binary semantic vectors. In *International Symposium on Quantum Interaction*, pages 24–35. Springer, 2012.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

Langdon Winner. Do artifacts have politics? *Daedalus*, pages 121–136, 1980.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

T Elizabeth Workman, Marcelo Fiszman, Thomas C Rindflesch, and Diane Nahl. Framing serendipitous information-seeking behavior for facilitating literature-based discovery: A proposed model. *Journal of the Association for Information Science and Technology*, 65 (3):501–512, 2014.

T Elizabeth Workman, Marcelo Fiszman, Michael J Cairelli, Diane Nahl, and Thomas C Rindflesch. Spark, an application based on serendipitous knowledge discovery. *Journal of biomedical informatics*, 60:23–37, 2016.

Jin Yang and Elke D Lange. SYSTRAN on AltaVista: a user study on real-time machine translation on the Internet. In *Conference of the Association for Machine Translation in the Americas*, pages 275–285. Springer, 1998.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 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*, FAT* '20, page 547–558, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367.

Meliha Yetisgen-Yildiz and Wanda Pratt. A new evaluation methodology for literature-based discovery systems. *Journal of biomedical informatics*, 42(4):633–643, 2009.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In

proceedings of the ACM Conference on Health, Inference, and Learning, pages 110–120, 2020.

Rui Zhang, Dimitar Hristovski, Dalton Schutte, Andrej Kastrin, Marcelo Fiszman, and Halil Kilicoglu. Drug repurposing for covid-19 via knowledge graph completion. *Journal of biomedical informatics*, 115:103696, 2021.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.

Yongjun Zhu, Olivier Elemento, Jyotishman Pathak, and Fei Wang. Drug knowledge bases and their applications in biomedical informatics research. *Briefings in bioinformatics*, 20 (4):1308–1321, 2019.

Shoshana Zuboff. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1):75–89, 2015.

Appendix A

FULL RESULTS TABLES

Table A.1: LBD: 25 cues, full set of abstracts

Entity Pair Type	Relation	ESP-GNBR				ESP-GNBR-SEMREP				ESP-SEMREP				SGNS				Targets per term	
		CD		DC		CD		DC		CD		DC		CD		DC			
		AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	CD	DC
Chem-Gene	Enzymes (DrugBank)	1.00	0.30	0.91	0.45	1.00	0.10	0.97	0.62	0.77	0.46	0.92	0.48	1.00	1.00	0.89	0.91	2.0	8.0
	Targets (DrugBank)	0.54	0.45	0.60	0.48	0.96	0.25	0.95	0.70	0.70	0.66	0.90	0.45	0.96	0.85	0.82	0.80	4.0	4.0
	PharmGKB	0.57	0.41	0.59	0.44	0.89	0.29	0.87	0.47	0.66	0.57	0.66	0.50	0.81	0.83	0.85	0.93	6.0	5.0
	Drug-Target (TTD)	0.57	0.60	0.61	0.38	0.86	0.46	0.85	0.64	0.86	0.70	0.75	0.53	0.81	0.70	0.82	0.80	1.0	2.0
	Drug-Inhibitor (TTD)	0.54	0.63	0.62	0.45	0.88	0.51	0.85	0.59	0.90	0.82	0.82	0.49	0.82	0.65	0.84	0.79	1.0	2.0
Chem-Disease	Side Effects (SIDER)	0.88	0.36	0.74	0.59	0.98	0.59	0.88	0.47	0.99	0.41	0.83	0.42	0.98	1.00	0.93	0.95	3.0	5.0
	Indications (SIDER)	0.90	0.36	0.77	0.48	0.97	0.54	0.88	0.55	0.98	0.42	0.83	0.56	0.95	0.96	0.97	0.95	2.0	3.0
	Biomarker-Disease (TTD)	0.68	0.28	0.54	0.74	0.95	0.60	0.34	0.16	0.99	0.29	0.08	0.12	0.95	1.00	0.69	0.56	1.0	1.0
	Indications (TTD)	0.98	0.30	0.81	0.52	0.99	0.49	0.87	0.56	0.99	0.24	0.85	0.53	0.95	0.96	0.94	0.92	1.0	3.0
	Target-Disease (TTD)	0.55	0.32	0.47	0.54	0.98	0.45	0.78	0.35	0.98	0.44	0.74	0.60	0.86	0.97	0.67	0.66	2.0	1.0
Gene-Disease	OMIM	0.52	0.32	0.39	0.58	0.96	0.45	0.80	0.41	0.96	0.33	0.71	0.57	0.85	0.93	0.77	0.73	1.0	2.0
	Gene-Disease (PGKB)	0.64	0.42	0.44	0.39	0.98	0.57	0.83	0.23	0.98	0.38	0.77	0.61	0.90	0.97	0.78	0.77	1.0	3.0
Gene-Gene	Carriers (DrugBank)	0.64	0.42	0.60	0.46	0.77	0.58	0.87	0.73	0.34	0.49	0.76	0.47	0.69	0.68	0.67	0.75	1.0	8.0
	Transporters (DrugBank)	0.92	0.43	0.89	0.48	1.00	0.33	0.95	0.69	0.96	0.60	0.94	0.48	1.00	1.00	0.83	0.86	2.0	9.0
	Gene-Gene (PGKB)	0.60	0.49	0.60	0.52	0.91	0.17	0.91	0.16	0.58	0.55	0.58	0.55	0.91	0.91	0.90	0.91	5.0	5.0
	Complex (Reactome)	0.66	0.53	0.64	0.38	0.71	0.42	0.82	0.40	0.82	0.62	0.91	0.63	0.66	0.62	0.82	0.67	1.0	1.0
	Reaction (Reactome)	0.56	0.36	0.71	0.79	0.74	0.47	0.92	0.46	0.85	0.74	0.88	0.46	0.78	0.67	0.86	0.69	1.0	1.0
AB:CD vs BD		AB:CD (15/17)		AB:CD (13/17)		AB:CD (17/17)		AB:CD (17/17)		AB:CD (16/17)		AB:CD (16/17		BD (8/17)		AB:CD (11/17)			
CD vs DC	AB:CD	CD (11/17)				CD (13/17)				CD (10/17)				CD (11/17)					
	BD	DC (13/17)				DC (9/17)				Tied				CD (10/17)					

Table A.2: RR: 25 Cues, Full set of abstracts.

Entity Pair Type	Relation	ESP-GNBR				ESP-GNBR-SEMREP				ESP-SEMREP				SGNS				Targets per term	
		CD		DC		CD		DC		CD		DC		CD		DC			
		AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	CD	DC
Chem-Gene	Enzymes (DrugBank)	1.00	1.00	0.97	0.98	1.00	0.83	0.97	0.33	0.64	0.40	0.96	0.64	1.00	1.00	0.96	0.92	2.0	7.0
	Targets (DrugBank)	0.80	0.76	0.84	0.95	0.83	0.52	0.94	0.45	0.83	0.47	0.95	0.51	0.99	0.66	0.99	0.75	2.0	2.0
	PharmGKB	0.88	0.97	0.92	0.99	0.91	0.62	0.98	0.43	0.52	0.49	0.98	0.51	0.98	0.82	0.99	0.96	2.0	2.0
	Drug-Target (TTD)	0.99	0.71	1.00	0.89	0.99	0.48	0.98	0.51	0.99	0.39	0.95	0.47	1.00	0.64	1.00	0.75	1.0	3.0
	Drug-Inhibitor (TTD)	0.99	0.66	1.00	0.88	0.99	0.53	0.98	0.66	0.99	0.36	0.95	0.45	1.00	0.70	1.00	0.74	1.0	3.0
Chem-Disease	Side Effects (SIDER)	1.00	1.00	0.98	0.98	0.99	0.57	0.98	0.51	0.99	0.22	0.98	0.58	1.00	1.00	0.97	0.95	3.0	6.0
	Indications (SIDER)	1.00	0.99	1.00	0.97	0.99	0.41	0.99	0.35	1.00	0.28	0.99	0.67	1.00	0.97	1.00	0.94	2.0	3.0
	Biomarker-Disease (TTD)	0.95	0.99	0.85	0.85	0.99	0.36	0.87	0.54	0.99	0.09	0.94	0.63	0.99	1.00	0.82	0.67	1.0	3.0
	Indications (TTD)	1.00	0.99	1.00	0.95	1.00	0.45	0.99	0.32	1.00	0.39	0.99	0.68	1.00	0.96	1.00	0.89	1.0	4.0
	Target-Disease (TTD)	0.86	0.99	0.71	0.80	0.98	0.62	0.88	0.48	0.99	0.20	0.93	0.70	0.99	0.99	0.90	0.59	1.0	3.0
Gene-Disease	Omim	0.98	0.96	0.92	0.73	0.99	0.44	0.86	0.55	0.99	0.21	0.90	0.52	1.00	0.95	0.99	0.72	1.0	2.0
	PharmGKB	0.78	0.99	0.71	0.96	0.99	0.57	0.76	0.59	0.99	0.37	0.65	0.55	1.00	1.00	0.94	0.73	1.0	3.0
Gene-Gene	Carriers (DrugBank)	0.84	0.88	0.74	0.96	0.77	0.34	0.93	0.43	0.33	0.64	0.98	0.62	0.86	0.77	0.97	0.77	2.0	2.0
	Transporters (DrugBank)	1.00	1.00	0.97	0.98	1.00	0.48	0.94	0.32	0.50	0.74	0.93	0.46	1.00	1.00	0.97	0.90	2.0	5.0
	PharmGKB	0.97	0.99	0.97	0.99	0.87	0.57	0.87	0.63	0.33	0.71	0.32	0.72	0.98	0.80	0.99	0.80	3.0	3.0
	Complex (Reactome)	1.00	0.74	1.00	0.71	0.99	0.48	1.00	0.43	0.91	0.45	0.94	0.48	1.00	0.62	1.00	0.61	1.0	1.0
	Reaction (Reactome)	0.98	0.78	0.99	0.68	0.99	0.53	0.99	0.50	0.95	0.54	0.97	0.48	1.00	0.65	1.00	0.61	1.0	1.0
AB:CD vs BD		AB:CD (8/17)		BD (8/17)		AB:CD (17/17)		AB:CD (17/17)		AB:CD (14/17)		AB:CD (16/17)		AB:CD (11/17)		AB:CD (17/17)			
CD vs DC	AB:CD	CD (8/17)				CD (10/17)				CD (10/17)				CD (7/17) + Tied (7/17)					
	BD	CD (11/17)				CD (10/17)				DC (14/17)				CD (7/17) + Tied (7/17)					

Table A.3: LBD: 25 cues, intersection of abstracts

Entity Pair Type	Relation	ESP-GNBR				ESP-GNBR-SEMREP				ESP-SEMREP				SGNS				Targets per term	
		CD		DC		CD		DC		CD		DC		CD		DC			
		AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	CD	DC
Chem-Gene	Enzymes (DrugBank)	0.89	0.72	0.84	0.43	1.00	0.44	0.96	0.22	0.86	0.52	0.94	0.42	1.00	1.00	0.86	0.88	2.0	8.0
	Targets (DrugBank)	0.53	0.52	0.53	0.45	0.91	0.70	0.94	0.40	0.70	0.50	0.91	0.34	0.94	0.83	0.80	0.76	4.0	4.0
	PharmGKB	0.54	0.53	0.53	0.46	0.89	0.56	0.91	0.30	0.66	0.50	0.81	0.34	0.79	0.80	0.83	0.92	6.0	4.0
	Drug-Target (TTD)	0.49	0.52	0.60	0.49	0.82	0.71	0.83	0.38	0.81	0.34	0.77	0.48	0.74	0.64	0.80	0.76	1.0	2.0
	Drug-Inhibitor (TTD)	0.56	0.44	0.54	0.49	0.85	0.70	0.87	0.45	0.88	0.28	0.85	0.53	0.75	0.61	0.84	0.74	1.0	2.0
Chem-Disease	Side Effects (SIDER)	0.64	0.47	0.56	0.46	0.98	0.61	0.88	0.27	0.99	0.73	0.85	0.39	0.97	0.99	0.91	0.94	3.0	5.0
	Indications (SIDER)	0.66	0.47	0.64	0.55	0.98	0.52	0.88	0.16	0.98	0.76	0.86	0.42	0.92	0.94	0.97	0.95	2.0	2.0
	Biomarker-Disease (TTD)	0.35	0.24	0.48	0.30	0.99	0.54	0.39	0.62	0.99	0.87	0.11	0.51	0.92	0.99	0.51	0.52	1.0	1.0
	Indications (TTD)	0.66	0.35	0.74	0.52	0.99	0.66	0.87	0.22	0.99	0.69	0.86	0.40	0.94	0.93	0.96	0.90	1.0	3.0
	Target-Disease (TTD)	0.54	0.50	0.41	0.47	0.98	0.67	0.82	0.57	0.98	0.72	0.69	0.38	0.85	0.94	0.59	0.58	2.0	1.0
Gene-Disease	OMIM	0.49	0.46	0.52	0.40	0.97	0.60	0.80	0.65	0.97	0.60	0.68	0.50	0.77	0.89	0.81	0.74	1.0	2.0
	PharmGKB	0.53	0.55	0.57	0.53	0.98	0.59	0.82	0.60	0.98	0.83	0.74	0.51	0.85	0.94	0.74	0.75	2.0	3.0
Gene-Gene	Carriers (DrugBank)	0.51	0.26	0.57	0.58	0.75	0.41	0.88	0.40	0.33	0.68	0.80	0.25	0.64	0.68	0.71	0.73	1.0	8.0
	Transporters (DrugBank)	0.73	0.53	0.70	0.45	1.00	0.48	0.95	0.32	0.98	0.33	0.94	0.41	1.00	1.00	0.76	0.81	2.0	9.0
	PharmGKB	0.57	0.49	0.56	0.50	0.91	0.54	0.91	0.54	0.60	0.59	0.60	0.60	0.89	0.89	0.89	0.89	5.0	5.0
	Complex (Reactome)	0.62	0.64	0.43	0.45	0.69	0.53	0.84	0.60	0.73	0.62	0.88	0.62	0.69	0.54	0.75	0.58	1.0	1.0
	Reaction (Reactome)	0.52	0.65	0.78	0.54	0.78	0.36	0.90	0.58	0.86	0.43	0.89	0.50	0.81	0.63	0.86	0.59	1.0	1.0
AB:CD vs BD		AB:CD (13/17)		AB:CD (14/17)		AB:CD (17/17)		AB:CD (16/17)		AB:CD (16/17)		AB:CD (15/17)		BD (8/17)		AB:CD (9/17)			
CD vs DC	AB:CD	CD (9/17)				CD (9/17)				CD (10/17)				DC (9/17)					
	BD	CD (11/17)				CD (11/17)				CD (11/17)				CD (10/17)					

Table A.4: RR: 25 cues, intersection of abstracts

Entity Pair Type	Relation	ESP-GNBR				ESP-GNBR-SEMREP				ESP-SEMREP				SGNS				Targets per term	
		CD		DC		CD		DC		CD		DC		CD		DC			
		AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	AB:CD	BD	CD	DC
Chem-Gene	Enzymes (DrugBank)	0.99	1.00	0.95	0.98	1.00	0.86	0.96	0.51	0.69	0.46	0.96	0.41	1.00	1.00	0.95	0.87	2.0	7.0
	Targets (DrugBank)	0.67	0.72	0.67	0.94	0.81	0.55	0.91	0.46	0.80	0.48	0.95	0.50	0.99	0.62	0.99	0.70	2.0	2.0
	PharmGKB	0.75	0.97	0.72	0.99	0.88	0.54	0.97	0.42	0.45	0.52	0.97	0.36	0.98	0.78	0.99	0.93	2.0	2.0
	Drug-Target (TTD)	1.00	0.63	1.00	0.84	0.98	0.39	0.97	0.48	0.99	0.47	0.93	0.40	1.00	0.60	1.00	0.66	1.0	3.0
	Drug-Inhibitor (TTD)	1.00	0.63	1.00	0.86	0.98	0.36	0.96	0.36	0.99	0.59	0.94	0.51	1.00	0.72	1.00	0.69	1.0	3.0
Chem-Disease	Side Effects (SIDER)	0.96	1.00	0.94	0.98	0.99	0.07	0.97	0.58	0.99	0.41	0.98	0.45	1.00	1.00	0.97	0.93	3.0	6.0
	Indications (SIDER)	1.00	0.99	1.00	0.96	1.00	0.12	0.99	0.50	1.00	0.22	1.00	0.46	1.00	0.95	1.00	0.92	2.0	3.0
	Biomarker-Disease (TTD)	0.60	0.99	0.70	0.87	0.98	0.55	0.90	0.53	0.99	0.03	0.94	0.57	0.99	0.99	0.85	0.61	1.0	3.0
	Indications (TTD)	1.00	0.99	1.00	0.94	1.00	0.20	0.99	0.38	1.00	0.16	0.99	0.55	1.00	0.94	1.00	0.85	1.0	4.0
	Target-Disease (TTD)	0.58	0.99	0.66	0.77	0.98	0.37	0.89	0.53	0.99	0.07	0.93	0.57	0.99	0.96	0.88	0.56	1.0	3.0
Gene-Disease	OMIM	0.93	0.88	0.91	0.69	0.99	0.32	0.87	0.60	0.99	0.32	0.92	0.49	1.00	0.91	0.99	0.71	1.0	2.0
	PharmGKB	0.58	1.00	0.58	0.95	0.98	0.26	0.79	0.42	0.99	0.18	0.62	0.44	1.00	0.99	0.92	0.74	1.0	2.0
Gene-Gene	Carriers (DrugBank)	0.43	0.99	0.48	0.96	0.78	0.75	0.92	0.47	0.32	0.57	0.98	0.54	0.88	0.78	0.96	0.74	1.0	2.0
	Transporters (DrugBank)	0.96	1.00	0.87	0.98	1.00	0.35	0.92	0.56	0.54	0.49	0.94	0.41	1.00	1.00	0.95	0.84	2.0	4.0
	PharmGKB	0.89	0.99	0.92	0.99	0.86	0.49	0.86	0.44	0.31	0.55	0.32	0.49	0.99	0.78	0.99	0.78	3.0	3.0
	Complex (Reactome)	1.00	0.70	1.00	0.65	0.97	0.50	0.99	0.50	0.90	0.60	0.92	0.61	1.00	0.57	1.00	0.55	1.0	1.0
	Reaction (Reactome)	0.95	0.80	0.92	0.72	0.96	0.46	0.97	0.53	0.93	0.59	0.95	0.53	1.00	0.58	1.00	0.58	1.0	1.0
AB:CD vs BD		BD (10/17)		BD (10/17)		AB:CD (17/17)		AB:CD (17/17)		AB:CD (14/17)		AB:CD (16/17)		AB:CD (13/17)		AB:CD (17/17)			
CD vs DC	AB:CD	CD (6/17); Tied (7/17)				CD (11/17)				Tied				CD (7/17); Tied (8/17)					
	BD	CD (12/17)				DC (9/17)				DC (9/17)				CD (12/17)					