VICTOR PETRÉN BACH HANSEN

TOWARDS FAIRNESS IN CONVERSATIONAL NATURAL LANGUAGE PROCESSING

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen.

TOWARDS FAIRNESS IN CONVERSATIONAL NATURAL LANGUAGE PROCESSING

VICTOR PETRÉN BACH HANSEN

UNIVERSITY OF
COPENHAGEN

Doctor of Philosophy (Ph.D.)
Department of Computer Science
Faculty of Science
University of Copenhagen

August 2021

SUPERVISORS:
Anders Søgaard, University of Copenhagen
Stig Geer Pedersen, Topdanmark

AFFILIATIONS:
University of Copenhagen,
Faculty of Science,
Department of Computer Science

Topdanmark A/S

ABSTRACT

With the emergence of deep learning-based Natural Language Processing (NLP), the field of conversational Artificial Intelligence (AI) has gone from a mere pipe-dream to full-blown commercial integration into our society in the span of less than a decade. The recent advances of conversational systems are primarily driven by data-hungry models that require vast quantities of data, which poses several challenges in terms of the models' performance and their societal impacts.

This thesis presents work that contributes to the field of NLP and conversational AI in multiple ways. The first part explores methods for building more intelligent dialogue systems. Here we examine how user feedback can be incorporated to transfer knowledge from one domain to another more efficiently, how to resolve elliptical structures in a conversational context and how bias in dialogue-based data collection guidelines can manifest itself in the resulting corpora.

The second part looks at how NLP models adhere to socio-demographic fairness principles under different constraints, namely compression and privacy. While compressing neural models for the sake of a reduced memory footprint and inference cost is an attractive trait, we find that pruning methods in text classification systems lead to an increase in disparity of performance among different groups. Similarly, model privacy is also shown to be at odds with fairness principles, but we find that combined with distributionally robust optimization, it can lead to both private and fair models.

RESUMÉ

Med fremgangen af sprogteknologi (NLP) baseret på dyb læring har konversationel kunstig intelligens (AI) bevæget sig fra at være ønsketænkning til en fuldt ud kommerciel integrering i vores samfund på mindre end et årti. De nylige fremskridt inden for sprogteknologiske konversationelle løsninger har primært været drevet af modeller som kræver enorme mængder af data hvilket giver giver flere udfordringer, både med hensyn til deres ydeevne men også hvordan de påvirker vores samfund.

Denne afhandling presenterer ny forskning der bidrager til NLP og konversationel AI på flere måder. Den første del udforkser nye metoder til at konstruere mere intelligente dialogsystemer. Her undersøges det hvordan man mere effektivt kan anvende feedback fra brugere til at overføre viden fra et domæne til et andet, hvordan man kan løse elliptiske konstruktioner i en konversationel kontekst og hvordan bias i retningslinjer til dialog-baseret dataindsamling kan manifistere sig i det resluterende korpus.

I den anden del tager kigger vi nærmere på hvordan NLP modeller overholder socio-demografiske retfærdighedsprincipper når de bliver udsat for forskellige restriktioner, mere specifikt komprimering og modelprivathed. At komprimere neurale modeller for at reducere deres hukommelsesafktryk og inferensomkostninger er en attraktiv egenskab, men vi viser at denne process i tekstklassificeringsssystemer kan medføre en ubalance i modellers ydeevne på tværs af forskellige grupper. Tilsvarende ser vi også at modelprivathed som udgangspunkt er i modstrid med disse retfærdighedsprincipper, men når det kombineres med robuste optimeringsmetoder kan det medføre både private og retfærdige modeller.

PUBLICATIONS INCLUDED IN THIS THESIS

This is an article-based dissertation consisting of papers that are either peer-reviewed or currently under review. The articles appear in their original published form, with the exception of formatting changes, added references as well as the correction of typos. The articles are as follows:

Bingel, Joachim, **Victor Petrén Bach Hansen**, Ana Valeria González-Garduño, Pawel Budzianowski, Isabelle Augenstein, and Anders Søgaard (2019). "Domain Transfer in Dialogue Systems without Turn-Level Supervision." In: *The 3rd NeurIPS workshop on Conversational AI: Today's Practice and Tomorrow's Potential*.

Hansen, **Victor Petrén Bach**, Atula Tejaswi Neerkaje, Ramit Sawhney, Lucie Flekova, and Anders Søgaard (2021). "The Impact of Differential Privacy on Group Disparity Mitigation." *Currently under review*.

Hansen, **Victor Petrén Bach** and Anders Søgaard (Aug. 2021a). "Guideline Bias in Wizard-of-Oz Dialogues." In: *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Online: Association for Computational Linguistics, pp. 8–14.

Hansen, **Victor Petrén Bach** and Anders Søgaard (Aug. 2021b). "Is the Lottery Fair? Evaluating Winning Tickets Across Demographics." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 3214–3224.

Hansen, **Victor Petrén Bach** and Anders Søgaard (2020). "What Do You Mean 'Why?': Resolving Sluices in Conversations." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 7887–7894.

PUBLICATIONS NOT INCLUDED IN THIS THESIS

I have also contributed to the following publications that were not included in this dissertation:

González, Ana Valeria, **Victor Petrén Bach Hansen**, Joachim Bingel, and Anders Søgaard (June 2019). "CoAStaL at SemEval-2019 Task 3: Affect Classification in Dialogue using Attentive BiLSTMs." In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 169–174.

ACKNOWLEDGMENTS

The past three years throughout this PhD have been very special to me for several different reasons, and now that I'm writing these concluding remarks, it leaves me feeling a bit bittersweet. Happy and relieved because I managed to finish the goal I set for myself initially, but also sad because I realize that it has come to an end as it has been a wonderful experience that I wouldn't be without. I jumped into this journey with a very narrow idea of what it meant to do research and it feels somewhat ironic that it's ending just as it's starting to click. Doing a PhD can in many ways be a hard and unforgiving endeavour, and the fact that I managed to stumble my way to the finish line can only be accredited to the many great people I've had the privilege of surrounding myself with.

First of all, I would like to thank my supervisor, Anders. For your guidance, never-ending optimism, and for fostering the most welcoming research environment one could ever hope for. I am truly honored for having been a part of it, and it is a big reason I'm where I am today. I've met a ton of new people here that I feel lucky to call my friends, thanks to all of you: Mostafa, Ana, Desmond, Yova, Mareike, Heather, Emanuele, Vinit, Simon, Lasse, Sheng, Joachim, Miryam, Rahul, Daniel to only name a *few*. In light of the ongoing pandemic this past year and a half, I want to emphasize my admiration of each one of you for finding the motivation to endure the very lonely process of doing a PhD in isolation. I genuinely mean it when I say that it has been what kept me going as well. And to all the new additions to the group that I hardly got chance to acquaint myself with, enjoy it as much as you can!

I also want to give a heartfelt *thank you* to all my colleagues in the machine learning team at Topdanmark: Stig, Søren, Lise, Mathias, Christian (x2), Kåre, Asger, Tue, Nicolaj and Vladimir for your trust and support throughout the process. I've truly enjoyed working with all of you over the past years and really appreciate you listening to my numerous laments when things weren't going the way I wanted.

Thanks to my family: mom, dad, Lisa and Max for providing me with an invaluable oasis of non-academic distractions and refuge in times of turbulence.

Last but not least, I also want to thank Pernille. For always standing by me and supporting me during the rough times and for being able to share the good times with you. You mean everything to me!

CONTENTS

I INTRODUCTION 1

1 INTRODUCTION 3
- 1.1 Conversational artificial intelligence 4
  - 1.1.1 Open-ended dialogue systems 5
  - 1.1.2 Task-oriented dialogue systems 5
  - 1.1.3 Challenges in conversational AI systems 6
- 1.2 Fairness in Natural Language Processing 7
  - 1.2.1 The Notion of fairness 8
  - 1.2.2 Common sources of bias 9
  - 1.2.3 Mitigating bias 10
- 1.3 Research questions 10
- 1.4 Thesis Overview and Contributions 12

II IMPROVING GENERALIZATION OF DIALOGUE SYSTEMS 16

2 DOMAIN TRANSFER IN DIALOGUE SYSTEMS WITHOUT TURN-LEVEL SUPERVISION 18
- 2.1 Introduction 18
- 2.2 Baseline Architecture 20
- 2.3 Domain Transfer Using Reinforcement Learning 21
- 2.4 Experiments 23
  - 2.4.1 Data 23
  - 2.4.2 Implementation Details 24
  - 2.4.3 Experimental Protocol 24
- 2.5 Results 25
- 2.6 Analysis 25
  - 2.6.1 Error Analysis 27
  - 2.6.2 Comparisons to Weak Supervision 27
- 2.7 Related Work 29
- 2.8 Conclusion 31

3 WHAT DO YOU MEAN 'WHY?': RESOLVING SLUICES IN CONVERSATIONS 32
- 3.1 Introduction 32
- 3.2 Background 33
- 3.3 A Conversational Sluicing Dataset 35
- 3.4 Experiments 39
  - 3.4.1 Baseline models 39
  - 3.4.2 Results 42
- 3.5 Analysis 43
- 3.6 Conclusion 45

4 GUIDELINE BIAS IN WIZARD-OF-OZ DIALOGUES 47
- 4.1 Introduction 47

4.2 Bias in CCPE-M 49
4.3 Bias in Taskmaster-1 53
4.4 Related Work 54
4.5 Discussion & Conclusion 55

III EXAMINING FAIRNESS IN NATURAL LANGUAGE PROCESSING 57

5 IS THE LOTTERY FAIR? EVALUATING WINNING TICKETS ACROSS DEMOGRAPHICS 59
5.1 Introduction 59
5.2 Related Work 61
5.3 Pruning methodology 62
5.4 Experiments 62
5.4.1 Data 62
5.4.2 Models 64
5.4.3 Measuring group disparity 65
5.5 Results 65
5.6 Conclusion 67

6 THE IMPACT OF DIFFERENTIAL PRIVACY ON GROUP DISPARITY MITIGATION 68
6.1 Introduction 68
6.2 Fairness and Privacy 69
6.3 Experiments 71
6.3.1 Algorithms 71
6.3.2 Tasks and architectures 73
6.3.3 Results 76
6.3.4 Discussion 78
6.4 Related Work 79
6.5 Ethics Statement 80
6.6 Conclusions 80

IV CONCLUSION 82

7 DISCUSSION AND CONCLUSION 84

V APPENDIX 87

A SUPPLEMENTARY MATERIAL FOR INDIVIDUAL STUDIES 89
A.1 Chapter 2 89
A.2 Chapter 4 90
A.3 Chapter 5 90
A.4 Chapter 6 95
A.4.1 Additional Figures 95
A.4.2 Experimental Details 95

BIBLIOGRAPHY 99

LIST OF FIGURES

Figure 1 Illustration of our proposed domain transfer dialogue state tracker, using a model M^P trained with turn-level supervision on d^P as a starting point for the fine-tuning policy $\pi_\theta(s|a)$ on domain d^F. 19

Figure 3 Example of conversational sluicing. Q_1 and A_1 provides a context for the second question Q_2 which has multiple correct resolutions, denoted in brackets, such as R_1 and R_2. 33

Figure 4 Illustration of the attention weights from all the 8 attention heads in the final decoder layer of the Transformer network. The x-axis corresponds to the position in the input sequence, whereas the y-axis corresponds to the output sequence. 42

Figure 5 Illustration of the attention weights from a single attention head in the 3-layer Transformer network, during decoding. The x-axis corresponds to the position in the input sequence, whereas the y-axis corresponds to the output sequence. 42

Figure 6 Conversational sluice resolution by the fine-tuned GPT-2 model that is judged better than the gold standard by our annotators. 44

Figure 7 A case where resolving the sluice in the an instance of the CoQA dataset improves the performance of QA system. $A_{no-sluice}$ is the answer generated when information contained in the bracket is included. 45

Figure 8 The percentage of sentences with the word *like* in the CCPE-M annotation guidelines (Guidelines), the suggested questions to ask users, in the guidelines (Suggestions), the *actual* first turns by the assistants (1st turn), and the actual replies by the users (2nd turn). In all cases, more than half of the sentences contain the word *like*. 48

Figure 9 Example of test sentence permutations. 50

Figure 10 Probability that a verb that describes a preference towards a movie is mentioned, given a priming word by the annotator is mentioned. 53
Figure 11 Probability that a guideline goal x_1 is mentioned before another one x_2 in an actual dialogue, given that x_1 comes before x_2 in the agent's guideline. 56
Figure 12 Fairness Sensitivity to Pruning (FSP): the gradient of the linear fit of (the logarithm of) the pruning ratio to min-max group-level disparity. We use this to quantify the sensitivity of Rawlsian min-max fairness to weight pruning across architectures, pruning strategies and datasets. 60
Figure 13 Macro-averaged performance of our feed-forward networks as a function of pruning ratio. Fairness Sensitivity to Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs. Results are for CivilComments. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. See the Appendix for similar plots for the Trustpilot Corpus. 61
Figure 14 FSP for Distributional Robust Optimization 66
Figure 15 Examples of the different subgroups that appear in a subset of the datasets we train on. CelebA (left) contains images of celebrities, using hair-color as our target variable and gender as our protected attribute. Blog Authorship Corpus (right) contains text-based blogposts on two topics {Technology, Arts} our targets, using $\mathcal{G} : \{\text{Man}, \text{Woman}\} \times \{\text{Young}, \text{Old}\}$ as our protected subgroups. 71
Figure 16 **Face Attribute Detection:** Performance of individual groups of increasing levels of ε. Comparing baseline ERM to Group DRO, we find that Group DRO performance on the minority group (blond males) perform much better under privacy constraints; we return to this in § 6.3.4. 76
Figure 17 **Topic Classification:** Performance of individual groups of increasing levels of ε. Group DRO, compared to baseline ERM, results in a more balanced performance across all groups, even on a low privacy budget. 76

Figure 18 **Volatility Forecasting:** A comparison of group-disparity between subgroups for increasing temporal volatility windows (τ) and privacy budgets (ε), over 5 independent runs. 78
Figure 19 CCPE-M Guidelines to Assistants 90
Figure 20 Macro-averaged performance of our feed-forward networks as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs. 91
Figure 21 Macro-averaged performance of our LSTMs as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs. 92
Figure 22 Macro-averaged performance of our layer-wise and globally pruned feed-forward networks trained with DRO as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs. 93
Figure 23 Macro-averaged performance of our layer-wise and globally pruned LSTM networks trained with DRO as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs. 94
Figure 24 Performance of individual groups of increasing levels of ε for the Trustpilot-US corpus. Error bars show standard deviation over 3 individual seeds. 95

Figure 25 Performance of individual groups of increasing levels of ε for the Trustpilot-UK corpus. Error bars show standard deviation over 3 individual seeds. 95

LIST OF TABLES

Table 1 Statistics of the MultiWOZ dataset. The reported numbers are from our processed dataset. 22

Table 2 Accuracy scores for our pre-trained baseline (BL) and the policy gradient fine-tuning (PG). The colored results along the left-to-right downward diagonal are in-domain results, dark red being the supervised results and light green the policy gradient fine-tuned results, and each pair of columns compare baseline and system results for each target domain. The AVERAGES row presents the average out-of-domain transfer scores for each domain. Note that while the PG method has access to more data, this does not invalidate the comparison, seeing that the additional data is relatively easy to obtain in an applied setting. 26

Table 3 Statistics of the *wh*-word distribution across the different splits for our conversational sluicing dataset. 37

Table 4 Results on our conversational sluicing dataset for a series of baseline architectures. We measure the performance using BLEU, GLEU and character n-gram F-score, precision and recall on the test split. In the last row, ANN AGREE denotes the inter-annotator agreement as the average between two randomly sampled gold annotations from each data point of the test set. 39

Table 5 The results of the human judgement experiment. To obtain human judgments, we asked three annotators to rank the output of three systems and the crowd-sourced gold annotations. MRR is the mean reciprocal ranking, and r_1 refers to the fraction of presented examples where the model was ranked as number 1. Our results show that the fine-tuned GPT-2 model produces favorable resolutions, both in terms of automatic as well as human evaluation and $1/5$ instances *better* than gold annotations. 40

Table 6 Generated output from our series of baselines, given a question-answer context, (Q_1, A_1) and follow-up one-word question. Examples are taken from the test split. 43

Table 7 Comparison of in-sample F_1 performance, performance on the same data with *like* replaced with phrases with similar meaning, and performance on Reddit data. Results are reported for training models on biased CCPE-M as well as a debiased CCPE-$M_{thesaurus}$ which improves model performance in almost all cases. 50

Table 8 CCPE-M and Reddit sentence-level statistics 51

Table 9 Detailed dataset statistics. N refers to the number of discrete demographics in the dataset and S is the size of each demographic test set. 63

Table 10 FFNN and LSTM hyperparameters. E_{dim} is embedding layer size, h_{dim} is hidden layer size, B is batch size and N is number of epochs. Both the layer-wise and global pruning structures use the same set of hyperparameters. 64

Table 11 FSP values across architectures, layer-wise (lw) and global (gl) pruning, and the four datasets. Our main observation is that FSP values are almost consistently positive, and slightly higher for global pruning. DRO does not consistently reduce FSP; we highlight cases where it does. 65

Table 12 Performance (top) and Δ-Fairness (bottom) of ERM and Group DRO across different degrees of differential privacy (ε). ε_1, ε_2 and ε_3 corresponds to ε-values of roughly 10, 5 and 1 respectively (see table for exact values). We report F1 scores for sentiment and topic classification, accuracy for face recognition and MSE for volatility forecasting. Group disparity (GD) is measured by the absolute difference between the best and worst performing sub-group (Δ-Fairness; see Definition 2.1). The performance and corresponding uncertainties are based on several individual runs of each configuration, see § A.4 in the Appendix for further details. Differential privacy consistently hurts fairness for ERM. For Group DRO, we **bold-face** numbers where strict differential privacy (ε_3) *increases* fairness; this happens in 4/5 datasets. We see large increases for face recognition and small increases for topic classification and sentiment analysis. 74

Table 13 Comparison of example turn predictions from the MultiWOZ dataset between the baseline model trained on the HOTEL domains, and the policy gradient fine-tuned model. Green indicates a correct prediction whereas red indicates a wrong prediction. 89

Table 14 Group distribution in the training set of CelebA 96

Table 15 Group distribution in the training set of Blog Authorship corpus 96

Table 16 Group distribution in the training set of Earnings Conference Calls 96

Table 17 Group distribution in the training set of Trustpilot-US 97

Table 18 Group distribution in the training set of Trustpilot-UK 97

Part I

INTRODUCTION

1 INTRODUCTION

Conversations are one of the most central parts of human-to-human communication, and while it, for you and me, may seem like a matter of course, it remains an intricate process between multiple subjects that in the field of Artificial Intelligence (AI) has been studied for decades (Austin, 1962). Natural Language Processing (NLP) is the interdisciplinary field between linguistics, computer science and AI that studies natural language from a computational perspective to understand the interaction between humans and machines. The creation of intelligent conversational systems, using NLP, that mimics the intricacies of human dialogue, in terms of expression but also comprehension, has been a long outstanding goal of AI (Green et al., 1961; Weizenbaum, 1966). As an academic field, NLP, and thus also conversational AI, has seen massive a growth with the emergence of deep learning (Gao, Galley, and Li, 2018; Ni et al., 2021) and particularly within the last few years with the rise of large pre-trained language models (LM) (Bommasani et al., 2021).

Due to this growth, conversational systems using NLP have seen a surge of commercial applications and success over the recent years. Personal virtual assistants (such as the Google Assistant, Amazon's Alexa and Apple's Siri) that enable users to interact with their smart devices through speech to get directions, book appointments and control appliances are being deployed at an unprecedented rate. Meanwhile, organizations are at an increasing rate utilizing conversational agents, e.g. for customer support related purposes. In 2019, the global market for conversational systems was forecast to grow by almost 30% annually over a five year period.[1]

At the same time, critical decisions in our society, be it in areas such as healthcare (Obermeyer et al., 2019), criminal justice (Rigano, 2018), or finance (Phaneuf, 2020), are increasingly being made with the assistance of intelligent systems based on Machine Learning (ML) and NLP. Machine-aided decision-making has many upsides, such as a significant increase in efficiency and reduced monetary cost; however, it can also have unintended consequences if e.g. the models at the foundation of the decisions favour one socio-demographic group over another. This bias is harmful from a decision-making standpoint and can unknowingly exacerbate itself in the long term. Without informing the models we train with an implicit understanding of social bias and negative stereotyping, they will inevitably reflect the bias

1 https://markets.businessinsider.com/news/stocks/global-chatbot-market-anticipated-to-reach-9-4-billion-by-2024-robust-opportunities-to-arise-in-retail-ecommerce-1028759508

contained in the data itself. Studying the ethical impacts of the algorithms we deploy was for a time overshadowed by the many breakthroughs that NLP has experienced, but in recent years an increasing number of concerning instances of biases has started to surface (see examples in Section 1.2). As such, Ethical AI as a field on its own is rapidly establishing itself[2], with some of the biggest academic ML conferences now requiring authors to discuss possible harmful impacts of their research.[3]

In this dissertation, we explore two main directions of research. The first is in the field of conversational AI, where we investigate how to improve data efficiency and quality for training better dialogue and question-answering systems and how we need to be mindful of potential biases when collecting new resources. The second direction studies the impact of the NLP models we deploy through the lens of fairness, more specifically, how neural models satisfy fairness principles under different circumstances.

1.1 CONVERSATIONAL ARTIFICIAL INTELLIGENCE

Dialogue systems[4] that communicate with users through text, speech or a combination once only existed in the world of Sci-Fi books and movies; however, with the technological advances we have been experiencing in the past decades, it seems like we have achieved what was previously only possible in fiction. The field of Conversational AI has come a long way from its infancy of rule-based systems in the 1960's (Weizenbaum, 1966) to the new era of data-driven neural models that deep learning has unlocked (Gao, Galley, and Li, 2018; Ni et al., 2021; Serban et al., 2018). This section briefly summarises the current landscape of dialogue systems, their role in society, and the challenges we still face and, to some extent, address in this project. This section only covers a fraction of the current research of dialogue systems. For a more comprehensive review, refer to (Chen et al., 2017; Jurafsky and Martin, 2020; Ni et al., 2021; Santhanam and Shaikh, 2019). When referring to conversational systems, we generally distinguish between two main categorizations, namely *open-ended dialogue systems* (or *chatbots*) and *task-oriented dialogue systems*. The differences are briefly outlined below.

2 For example the AIES conference: (*AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* 2018)

3 NeurIPS ethics statement: https://nips.cc/FAQ/EthicsFairnessInclusivityandCodeofConduct

4 Generally, the literature uses "dialogue systems" and "conversational systems/AI" interchangeable; this chapter does so as well.

1.1.1 *Open-ended dialogue systems*

Chatbots seek to simulate human-human dialogues by engaging users in natural or informal conversations that do not necessarily have a specific objective, and their application mainly lean toward entertainment. Its use-cases have its roots in clinically psychology, with systems such as ELIZA (Weizenbaum, 1966) and PERRY (Colby, Weber, and Hilf, 1971), but it has since evolved into conversational agents that can pretty much talk about anything, like Cleverbot[5] and Facebooks BlenderBot (Roller et al., 2021). The new school of open-ended dialogue systems, using large-scale pre-trained LMs as their backbone, now offers features like incorporating personalities that provides the agents with more human-like mannerisms such as empathy, more engaging talking points, and a notion of knowledge (Adiwardana et al., 2020; Dinan et al., 2019; Roller et al., 2021; Wolf et al., 2019; Zhang et al., 2020). Nevertheless, human evaluation studies of current open-ended dialogue systems indicate that these systems still are far from perfect. They sometimes suffer from memory issues during extended conversations, repetitions and contradictions, as well as hallucinogenic behaviour (Roller et al., 2021).

1.1.2 *Task-oriented dialogue systems*

Where open-ended dialogue systems often do not have a specific purpose in their communicative efforts, task-oriented dialogue systems on the other hand are concerned with providing a service that achieves a goal. As mentioned previously, these can be virtual personal assistants, such as Siri or Alexa, that can search, play music, or set alarms via voice-assisted commands. Compared to open-ended dialogue systems, task-oriented dialogue systems often have a more limited scope of interaction and are often designed using a modular approach. The traditional task-oriented dialogue pipeline is usually divided into four major components, namely (i) Natural Language Understanding (NLU) for interpreting user intents, (ii) the Dialogue State Tracking (DST) module to keep track of the history and current internal state of the conversation, (iii) a dialogue policy for determining the next course of action and (iv) a Natural Language Generation (NLG) module for generating a response to the user based on the chosen action and internal state (Chen et al., 2017; Jurafsky and Martin, 2020). Although the modular design allows for highly optimized individual components, it might not improve the overall performance when employed in conjunction. As an alternative, complete end-to-end neural dialogue systems have been suggested as a more straightforward solution (Budzianowski and Vulić, 2019; Ham et al., 2020;

5 https://www.cleverbot.com/

Hosseini-Asl et al., 2020; Le et al., 2020; Lei et al., 2018; Li et al., 2017; Peng et al., 2020).

Although open-ended, as well as task-oriented, dialogue research overlaps in some aspects, such as generating responses in a natural language, the research in Part ii of this thesis mainly focuses on the task-oriented setting.

1.1.3 *Challenges in conversational AI systems*

Even though conversational AI as a field has come a long way, many issues persist. Among the most prevalent ones are issues related to data and ethics which are unfolded below.

DATA Data-driven dialogue systems are heavily dependent on high quality data. Task-oriented dialogue systems are usually trained on fully annotated large-scale corpora of human-human dialogue such as MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020; Zang et al., 2020). Large-scale corpora can, due to their detailed level of annotation, be costly to collect, and can thus be a severely limiting aspect of improving the user experience. New unsupervised transfer learning methods that leverage large quantities of raw text, and conversations, have, for this reason, been proposed. Wolf et al. (2019) show that using transformer-based (Vaswani et al., 2017) generative pre-training (GPT) (Radford et al., 2018) can result in more personable conversational agents. In the task-oriented setting Budzianowski and Vulić (2019) simplifies the pipeline by providing a unified end-to-end framework, based on GPT-2 (Radford et al., 2019). Zhang et al. (2020) trains a tunable response generation model, DialoGPT, on a large corpora of Reddit comments, leading to improved conversational representations for downstream dialogue tasks. In Part ii we explore another approach of transferring knowledge when expanding the domains supported by task-oriented dialogue systems that relies on weaker supervision signals. Additionally, we also show how data curators should be mindful of ways annotation guidelines negatively can impact the quality of the data.

ETHICAL CONCERNS Since many of the state-of-the-art conversational systems are trained using large-scale dialogue datasets, this makes them vulnerable to encoding implicit biases in the data (Henderson et al., 2018). Cercas Curry and Rieser (2018) examine how conversational AI systems respond to sexual and offensive requests and find that responses range from non-engaging behaviour in commercial systems to flirtatious behaviour in data-driven systems. Dinan et al. (2020) analyze and identify gender bias in a number of different dialogue datasets and find that the dialogue systems induced from said data reflect and even amplify this bias. They propose to mitigate

gender bias through data augmentation (gender swapping) and bias controlled training. Dinan et al. (2021) surveys the landscape of bias and safety issues in end-to-end conversational models that are based on large-scale LMs and provides a framework for how and when to, or when not to, release them. In a similar line of research, (Bender et al., 2021) takes a step back and analyzes, among others, the ethical risks that these foundational models are associated with.

In the next section we delve further into the topic of bias and fairness in NLP and steps we can take to mitigate bias.

1.2 FAIRNESS IN NATURAL LANGUAGE PROCESSING

With the rapid integration of NLP and ML in everyday software solutions, it has become imperative to monitor the societal impact of the algorithms we deploy (Hovy and Spruit, 2016). It is vital that these system does not exhibit discriminatory behaviour against protected groups and it is imperative that they ensure equal opportunity for individuals across the entire population.[6]

A series of recent instances of systematic and algorithmic bias are particularly concerning. Amazon discovered, that an attempt to use AI to automate their recruitment process led to the model favouring male candidates over their female counterparts due to lack of representation in the data (Dastin, 2018). For the computer vision field, Buolamwini and Gebru (2018) show that bias are also very present in commercial gender classifications systems, such as facial recognition systems, exhibiting a significantly higher error rate among darker-skinned females compared to light-skinned males, again due to under-representation in the data. Obermeyer et al. (2019) reveals how a racial bias in prediction algorithms used in the U.S. health care system, due to the reliance on biased predictor variables, underestimates the medical needs of black patients compared to white patients, affecting millions of people. They use health costs as a proxy to reflect their health needs, and since less money was spent on black patients, the model falsely estimated that black patients had the same needs as healthier white patients. In the context of NLP, unsupervised representations of words, such as word embeddings (Mikolov et al., 2013; Pennington, Socher, and Manning, 2014) or contextual representations derived from pre-trained LMs (Devlin et al., 2019; Peters et al., 2018; Radford et al., 2019) are synonymous with state-of-the-art systems. They are often extracted from vast quantities of text corpora that have been curated from the Internet. Since bias at the fundamental level is inherently integrated into our society, models

6 Here we refer to groups that are legally protected due to attributes such as age, gender and race. See the following URL for a complete list: https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en

can unintentionally reflect the same systematic biases that we project. Bolukbasi et al. (2016) show how word embeddings end up reinforcing negative gender stereotypes by examining word analogies such as "man is to computer programmer as woman is to X", where it turns out "homemaker" is the most likely candidate. Similarly, we also see that text generated by LMs, such as OpenAI's GPT-3 (Brown et al., 2020), have been shown to generate racist and toxic language when prompted with specific snippets.[7] Both Brown et al. (2020) and Abid, Farooqi, and Zou (2021) also discuss the societal biases that GPT-3 encodes, such as occupation discrimination based on genders, how the religion of Islam is associated with negatively loaded words like "terrorist", and how different sentiments are attributed to different ethnic races.

Adopting these representations blindly are therefore destined to propagate these biases further downstream to the task at hand.

1.2.1 *The Notion of fairness*

As previously mentioned, one of the focus areas of this thesis is that of *fairness*, which can be defined as the absence of prejudice and discrimination for a group based on their descriptive attributes. We consider an algorithm, or model, to be *fair* if the output is independent of a set of given variables that should not influence the outcome of the prediction. The work mainly focuses on personal traits such as gender, disability, race, sexual orientation, etc.

In machine learning systems we usually quantify fairness by observing potential divergences in prediction rates on sub-populations with labelled attributes. There exists many ways of measuring model fairness. Gajane (2017) and Verma and Rubin (2018) formally defines several agreed upon measures of fairness for analysing bias in machine learning. In Part iii of this thesis, we mainly concern ourselves with *group fairness*, which are derived from the notion of collectivist egalitarianism for distributive justice (Gajane, 2017; Rawls, 1971). We outline a three different variations of group fairness and the conditions a model is considered fair under: (i) **demographic parity:** when protected groups should have equal rates of positive outcomes. (ii) **equality of opportunity:** when protected groups should have equal rates of true positives.and (iii) **equalized odds:** when protected groups should have equal rates of true positives *and* false positives.

7 https://www.technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/

1.2.2 *Common sources of bias*

Properly identifying the origin of potential bias is an important tool in reducing harmful societal effects. We generally distinguish between three sources of bias, namely the *data* we train our models on, the *models* themselves and the *modelers* behind them (Bommasani et al., 2021).

DATA In NLP, text corpora lie at the foundation of our predictive models. It is important that the data is representative of the real world in the context we deploy them in. However, human subjectivity, as well as underlying social biases, will manifest itself in the estimator if not taken into account for (Caliskan, Bryson, and Narayanan, 2017; Garg et al., 2018; Henderson et al., 2018; Paullada et al., 2020; Voigt et al., 2018). When some demographics or other members of protected groups represented in the dataset do not reflect the true population, it can lead to poor generalization. To reiterate a previous example, Buolamwini and Gebru (2018) show that gender classification systems fail to correctly classify dark-skinned females more so than light-skinned males due to being vastly underrepresented in the dataset. In Chapter 4 we show how a dataset containing a lexical bias can lead to worse model generalization on benchmarks where it is removed.

MODELS While data is a common perpetrator when it comes to bias, Hooker (2021) argues against the notion that a biased model is only a reflection of the underlying data it has been trained on and that model design is an essential aspect of mitigating bias. The decisions we make during the development of a model, be it choosing an appropriate objective function, optimizer or the architecture itself, can help tackle the amplification of bias. Jiang et al., 2020 show that the tail-end, or underrepresented features in general, are learned later in the training process, demonstrating that choosing a learning rate can also affect the fairness of a model (Hooker, 2021). As an example, in Chapter 5 we show how model compression can exacerbate the bias of NLP models.[8] In Chapter 6 we show how differentially private models also suffer the same fate if not explicitly accounted for in the objective function.

MODELERS Just like raw textual data in many ways reflect the humans (along with their systematic biases) it is curated from, the models are a reflection of the developers designing it. If representation and diversity among the people building the system are poor, the design decisions might mirror that. Bommasani et al. (2021) argues that for e.g. multilingual model, flawed data handling of underrepresented languages in multilingual datasets (Caswell et al., 2021), result-

8 Hooker et al. (2019) also shows this phenomenon in the realm of facial recognition.

ing in biased models, could be prevented if a better representation of developers could have identified the issue earlier. Furthermore, they also note that as the end-users of the models most likely are more diverse than those developing it and integrating user feedback in the model design to reduce bias is one way to move forward.

1.2.3 *Mitigating bias*

With the recent focus on uncovering unintended biases in ML and NLP, efforts have similarly been made in avoiding and rectifying them, some which have been hinted at in the previous section. There is a long list of literature that attempts to address the issue of bias, especially gender bias, in many aspects of NLP ranging from word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018b), coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a), sentiment analysis (Kiritchenko and Mohammad, 2018), machine translation (Vanmassenhove, Hardmeier, and Way, 2018) and language modeling (Bordia and Bowman, 2019). However, while many of these post-hoc debiasing methods have shown to reduce gender bias significantly, they make no guarantees for gender neutrality (Ethayarajh, Duvenaud, and Hirst, 2019; Gonen and Goldberg, 2019).

Proactive measures have also been suggested to help practitioners document potential pitfalls when it comes to identifying bias in the development stage. For example, Gebru et al. (2018) and Bender and Friedman (2018) suggest to create datasheets and data statements for your datasets, which urges the creater to better characterize aspects of the data, including potential biases. Mitchell et al. (2019) proposes an analogous concept for the models themselves, e.g. the setting the trained model is best suited to be deployed in, along with extensive evaluation on different demographic groups. Similarly, in their survey of bias in NLP, Garrido-Muñoz et al. (2021) lists a number of steps that helps software engineers deal with stereotyping bias for applications of large-scale LMs.

Several authors approaches bias mitigation by optimizing directly for out-of-distribution mixtures of sub-populations using distributionally robust optimization techniques (Hashimoto et al., 2018; Hu et al., 2018; Levy et al., 2020; Oren et al., 2019; Sagawa et al., 2020a), which we also adopt in Chapters 5 and 6.

1.3 RESEARCH QUESTIONS

This thesis is the product of an Industrial PhD project executed in collaboration with the Danish insurance company Topdanmark and the Department of Computer Science at the University of Copenhagen. The project is funded partially by Topdanmark and partially by the Innovation Fund Denmark. The role of this project in the context of

Topdanmark has mainly been exploring possible research ideas related to expanding its current use of conversational agents used for automating part of the more than 1 million annual customer support related queries regarding, e.g. their insurance policies. A specific challenge that the research in this thesis tackles relates, in two ways, to the *data* we base our dialogue systems on: Firstly, how user feedback can be beneficial when expanding task-oriented dialogue systems to new domains, where high-quality annotations are unavailable (Chapter 2) —secondly, the importance of guideline formulations regarding bias when setting out to collect and annotate new dialogue data (Chapter 4). We also explore how to resolve sluices, a frequent and challenging elliptical structure in informal conversations, which can lead to downstream improvements for conversational Question-Answering (QA) systems (Chapter 3).

As we saw in the previous section, systematic biases and fairness of NLP models, with respect to socio-demographic groups, is a major concern in commercial applications as models can seemingly appear to perform well on the surface while still failing to accommodate minority groups (Sagawa et al., 2020b). Based on some of the challenges we outlined in Section 1.2, we can conclude that the bias problem in NLP is of great relevance for industrial adaptation, for both legal as well as ethical reasons. It is crucial that when we deploy NLP solutions, be it text classification or conversational agents, based on pretrained representations that have been shown to propagate negative stereotypes, practitioners need to verify that the fairness principles of the model in its intended environment still is satisfied. In this thesis, we investigate how well they do this under different constraints, such as model compression (Chapter 5) and privacy (Chapter 6).

As previously outlined, dialogue systems are an increasingly adopted paradigm that has already seen commercial success, including at Topdanmark. Due to the dependence on high-quality datasets, we still face many challenges that require further attention, both in terms of model performance and the ethical issues that follow. As a response to the problems highlighted in the previous sections, we motivate the research in this thesis with five research questions.

The first three questions concern how we can improve the generalization performance of our dialogue systems, both in terms of the datasets we rely on as well as how we adapt to new domains, but also how we can improve the quality of our existing data by resolving ambiguous language in conversation:

- How do we leverage user feedback to more efficiently improve the generalization capabilities of our dialogue systems?
- How can we resolve implicit content from a conversational context to improve the quality of our dialogue systems?

- To what extent does the formulation of conversational data collection guidelines influence the resulting corpora?

In the next part we shift our focus from improving generalization of dialogue systems to the ethical challenges that we face in NLP, namely model fairness. More specifically, we are interested in examining socio-demographic fairness of NLP models when subjected to certain limitations. The last set of research questions is as follows:

- How well does our NLP models satisfy fairness principles when subject to compression techniques?
- How is fairness affected by group robust optimization objectives when under the influence of privacy preserving methods?

1.4 THESIS OVERVIEW AND CONTRIBUTIONS

This thesis explores multiple aspects of NLP and is divided into four parts, spanning two main research directions. This first part (Part i) introduces the PhD project and the context in which it was conducted. In Part ii, our first research direction, we investigate methods that improve dialogue systems in multiple ways. Firstly, by exploring how to, more efficiently, transfer Dialogue State Tracking systems to new domains using user feedback at the dialogue-level in combination with Reinforcement Learning (Chapter 2). We then study how to resolve conversational sluices, a complex elliptical structure, in order to ultimately improve downstream performance of conversational QA models that struggle doing so implicitly (Chapter 3). Lastly, we investigate how a bias in annotation guidelines for dialogue-based data collection frameworks can insert itself in the resulting dataset, ultimately leading to poor generalization on data where the bias is not present (Chapter 4). In Part iii, our second direction, we explore concepts of model fairness and bias in ML and NLP, main how well they do so under different constraints, such as model compression (Chapter 5) and privacy (Chapter 6). Lastly, in Part iv we summarize and discuss our findings from Chapters 2 to 6 and suggest directions for furthering the research presented here.

We summarize the contributions of the chapters that constitutes this dissertation (Parts ii and iii) as follows:

- Task-oriented dialogue systems are typically trained using manually annotated data at the turn-level, which are cumbersome and expensive to obtain. Chapter 2 explores how we more efficiently can, using reinforcement learning, transfer DST models to new domains, by instead leveraging reward signals at the dialogue-level, that are more easily accessible in a realistic scenario. Our experiments demonstrates how our policy gradient based method quickly adapts to new domains as well as

improves in-domain performance of already converged model trained with regular turn-level supervision.

- Stand-alone *wh*-word questions, such as *When?*, are generally trivial for people to understand in a conversation, but can pose a real challenge for dialogue systems to interpret correctly when the context has to be retrieved from past turns. In Chapter 3 we introduce the task of conversational sluice resolution, a pervasive and challenging ellipsis phenomenon. We crowd-source a new dataset consisting of roughly 4000 annotated conversational sluices, collected from pre-existing Question-Answering datasets and present a series of baselines based on both sequence-to-sequence models as well as large pre-trained language models. Our human evaluation of automatically resolved sluices show that they can at times rival the performance of human annotators.

- In Chapter 4 we introduce and analyse the concept of guideline bias, the unintended bias that arises from how guidelines are formulated, in datasets collected using the Wizard-of-Oz framework. We show two things: (i) how a simple bias toward the verb *like* easily leads us to overestimate performance in the wild by showing performance drops on semantically innocent perturbations of the test data and how we through data augmentation can, to some extent, mitigate it and (ii) how the order of the instructions influence the structure of the resulting conversation.

- Compressing models while minimizing loss of performance is crucial for storage and inference coast in the age of portable devices but can have unintended consequences. In Chapter 5, we study the impact of weight pruning on fairness in NLP. We evaluate demographic group disparity across two architectures, two pruning strategies and two datasets, including multilingual sentiment classification and English toxicity classification. We introduce a new metric, *fairness sensitivity to pruning* that measures how Rawlsian min-max fairness across demographic groups decreases with weight pruning. Our results suggest that pruning increases group-level performance disparities, mostly at high pruning rates and with some variance across architectures and pruning strategies. Group-level disparities seem to be in part a result of the instability of weight pruning. Our results also indicate that weight pruning in combination with distributionally robust optimization objectives can *sometimes* be used to induce fairer, sparse classifiers.

- Chapter 6 investigates how Differential Privacy (DP) impacts model fairness in two different settings, namely under (i) a baseline empirical risk minimization and (ii) a group distribu-

tionally robust optimization. In line with previous work, our results show how DP disproportionately impacts minority subpopulations negatively during training in the baseline setting; more interestingly, however, we show that DP not only mitigates the decrease but also can improve fairness compared to our non-private experiments in the distributionally robust setting.

Part II

IMPROVING GENERALIZATION OF DIALOGUE SYSTEMS

2

DOMAIN TRANSFER IN DIALOGUE SYSTEMS WITHOUT TURN-LEVEL SUPERVISION

ABSTRACT

Task oriented dialogue systems rely heavily on specialized dialogue state tracking (DST) modules for dynamically predicting user intent throughout the conversation. State-of-the-art DST models are typically trained in a supervised manner from manual annotations at the turn level. However, these annotations are costly to obtain, which makes it difficult to create accurate dialogue systems for new domains. To address these limitations, we propose a method based on reinforcement learning for transferring DST models to new domains without turn-level supervision. Across several domains, our experiments show that this method quickly adapts off-the-shelf models to new domains and performs on par with models trained with turn-level supervision. We also show our method can improve models trained using turn-level supervision by subsequent fine-tuning optimization toward dialog-level rewards.

2.1 INTRODUCTION

Intelligent personal assistants, such as Amazon Alexa, Apple Siri and Google Assistant, are becoming everyday technologies. These assistants can already be used for tasks such as booking a table at your favorite restaurant or routing you across town. Such dialogue systems potentially allow for smooth interactions with a myriad of online services, but rolling them out to new tasks and domains requires expensive data annotation. In developing goal-oriented dialogue systems, dialogue state tracking (DST) refers to the subtask of incrementally inferring a user's intent as expressed over a sequence of turns. The detected user intent is then used by the dialogue policy in order to decide what action the system should take (Henderson, 2015). For example, in a chatbot-based train reservation system, DST amounts to understanding key information provided by the user as *slot-value pairs*, such as the desired departure and arrival stations, the day and time of travel, among others. With the introduction of the Dialogue State Tracking Challenges (Williams et al., 2013), this line of research has received considerable interest.

State-of-the-art models for dialogue state tracking are typically learned in a fully supervised setting from datasets where slots and values are annotated manually at the turn level (Mrkšić et al., 2017a; Nouri

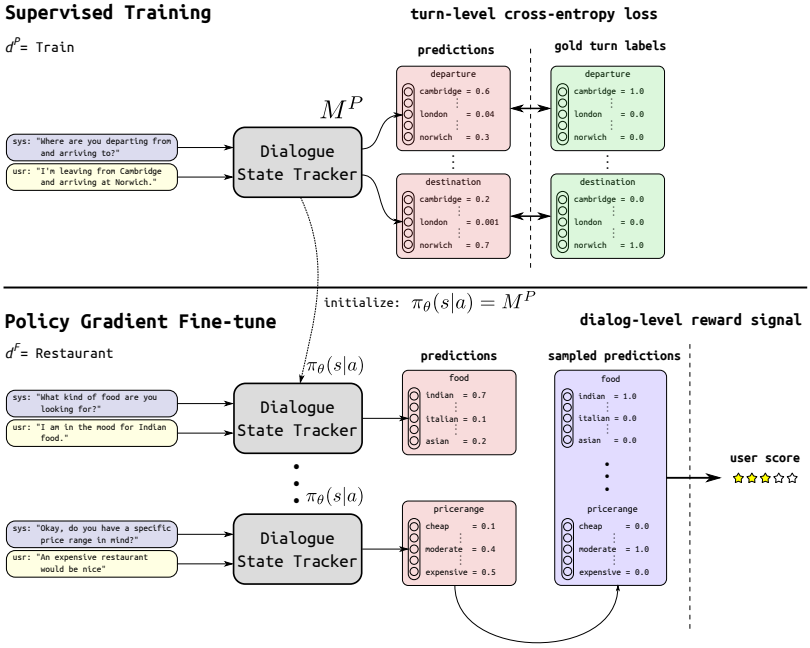


Figure 1: Illustration of our proposed domain transfer dialogue state tracker, using a model M^P trained with turn-level supervision on d^P as a starting point for the fine-tuning policy $\pi_\theta(s|a)$ on domain d^F.

and Hosseini-Asl, 2018; Ren et al., 2018; Zhong, Xiong, and Socher, 2018). This allows for high-accuracy models in a select number of domains, where turn-level annotations are available. However, such annotations are cumbersome and costly to obtain, and, in practice, a bottleneck for producing dialogue systems for new domains.

In this paper, we present an approach to DST that pre-trains a model on a source domain for which turn-level annotations exist, then fine-tunes to other target domains for which no turn-level annotation is directly available. In particular, we use standard maximum likelihood training to induce a supervised model for the source domain, and resort to Reinforcement Learning (RL) from dialog-level signals (e.g., user feedback) for transferring to the target domain, improving target domain performance and potentially saving massive annotation efforts. In addition to this, we also report consistent gains using dialogue-level feedback to further improve supervised models in-domain.

CONTRIBUTIONS To summarize, our contributions are: Relying on *only dialogue-level signals* for target domain fine-tuning, we show that it is possible to transfer between domains in dialogue state tracking using reinforcement learning, gaining a significant increase in performance over baselines trained using source-domain, turn-level annotations. Second, we show that policy gradient methods can also be used to boost the in-domain accuracy of already converged models trained in the usual supervised manner.

2.2 BASELINE ARCHITECTURE

Our proposed model is based on StateNet (Ren et al., 2018), which uses separate encoders for the two basic inputs that define a turn: the user utterance and the system acts in the previous turn. These inputs are represented as fixed-size vectors that are computed from n-gram based word vector averages, then passed through a number of hidden layers and non-linearities. We concatenate these representations, and, for every candidate slot, we compare the result to slot representations, again derived from word vectors and intermediate layers. We update the hidden state of a GRU encoding the dialogue history.and compare this representation to all candidate values for a given slot. From this, we compute the probability of slot-value pairs. For efficiency reasons, we modify the original StateNet model to only update the GRU that tracks the inner dialogue state after every turn and once all slots are processed within that turn, rather than after every computation of slot values.

Embedding slots and values, and treating them as an input to the model rather than as predefined classes, are important features of StateNet: These features enable zero-shot learning and make the architecture a natural choice for domain transfer experiments, even if it is not the first to enable zero-shot learning in dialogue state tracking in such a way (Ramadan, Budzianowski, and Gasic, 2018; Zhong, Xiong, and Socher, 2018). In addition to being well suited for domain transfer, StateNet also produces state-of-the-art results on the DSTC2 and WOZ 2.0 datasets (Henderson, Thomson, and Williams, 2014; Mrkšić et al., 2017b).

Training our model is split into two distinct phases. From a pre-training domain d^P for which manual turn-level annotations are available, we learn a model M^P, using the available dialogues to train our system until convergence on a held-out development set. Then, for a further domain $d^F \notin D - d^P$, where D is the set of available domains, we use a policy gradient training to fine-tune M^P to the new domain, based on simulated user feedback, corresponding to how many goals we met at the end of the conversation. Figure 1 presents an overview of this training process.

PRE-TRAINING In the pre-training phase, we use our implementation of the StateNet model. Just as Ren et al. (2018), we focus on predicting the user state and use the information about the system acts contained in the data. During pre-training, we rely on turn level supervision, training models on a single domain and evaluating on a held out set from that same domain.

2.3 DOMAIN TRANSFER USING REINFORCEMENT LEARNING

DIALOGUE STATE TRACKING WITH RL Given a pre-trained model M^P trained on a domain d^P, we fine-tune it on a new domain d^F. Since we do not have turn-level annotations for the target domain, we cannot use maximum likelihood training to adapt to d^F. This also means that standard domain adaptation methods (Blitzer, McDonald, and Pereira, 2006; Daume III and Marcu, 2006; Jiang and Zhai, 2007) are *not* applicable. Instead, we frame our transfer learning task as a reinforcement learning problem and use policy gradient training. This allows us to use dialogue-level signals as a reward function. Policy gradient training has advantages over value-based RL algorithms, including better convergence properties, ability to learn optimal stochastic policies and effectiveness in high-dimensional action spaces (Sutton and Barto, 1998). Within this paradigm, the dialogue state tracker can be seen as an *agent* that interact in the *environment* of a dialogue. Throughout the conversation, the DST model tracks the presence of slots in the conversation and assigns a probability distribution over the values, if present. At the end of a dialogue, represented by a state s, our model goes through the slots and performs an action, a, by sampling a value from the present slot-value probability distribution. It then receives a reward based on how well it predicted slot-value pairs. We illustrate this training regime using dialog-level feedback in the lower half of Figure 1.

DIALOG-LEVEL REWARD SIGNAL In a real-world setting, dynamically obtaining turn-level rewards, for instance from user feedback, is not only costly, but undesirable for the user experience. In contrast, acquiring user feedback at the end of a dialogue, for instance in the form of a 5-star scale, is more feasible and common practice in commercial dialogue systems.

For practical reasons, we simulate this feedback in our experiments by the success our model achieves in correctly predicting slot-value pairs, assuming that model performance is correlated with user satisfaction. Concretely, we use the Jaccard index between the predicted (S_P) and ground-truth (S_G) final belief state:

$$R_{goal} = \frac{|S_G \cap S_P|}{|S_G \cup S_P|} \tag{1}$$

POLICY GRADIENT METHODS We define the policy network π_θ as the StateNet network, which is initialized with a pre-trained model M^P. The weights of the StateNet network are then fine-tuned using stochastic gradient ascent, i.e., in the direction of the gradient of the

Domain	Dialogues	Dialogues with only one domain	Turns/ Dialogue	Slots	Values (processed)	Split sizes (train-dev-test)
TAXI	2057	435	7.66	4	610	326-57-52
TRAIN	4096	345	10.26	6	81	282-30-33
HOTEL	4197	634	10.95	9	187	513-56-67
RESTAURANT	4692	1310	8.78	6	330	1199-50-61
ATTRACTION	3515	150	7.69	2	186	127-11-12

Table 1: Statistics of the MultiWOZ dataset. The reported numbers are from our processed dataset.

objective function $\nabla J(\theta)$. The update in the vanilla policy gradient algorithm is:

$$\nabla J(\theta) = \nabla_\theta \log \pi_\theta(a|s) R_{goal} \tag{2}$$

We update the policy of the network after each iteration, following Sutton and Barto (1998).

VARIANCE REDUCTION METHODS Policy gradient methods suffer form certain shortcomings. For instance, they frequently converge to local, instead of global, optima. Furthermore, the evaluation of a policy is inefficient and suffers from high variance (Sutton and Barto, 1998). A common way to circumvent the above-mentioned issues is to introduce a baseline model (Weaver and Tao, 2001). It is typically initialized as a frozen copy of the pre-trained model M^P. The baseline models the reward B_{goal} at the end of the dialog. We can then define an *advantage* of an updated model over the initial one as $A_{goal} = R_{goal} - B_{goal}$. In addition to subtracting the baseline, we also add the entropy $\mathcal{H}(\pi_\theta(a|s))$ of the policy to the gradient to encourage more exploration (Williams and Peng, 1991), in order to counteract the local optima convergence shortcoming. With these modifications to the policy update in Eq. (2), we can rewrite the final gradient as:

$$\nabla J(\theta) = \nabla_\theta \log \pi_\theta(s|a) A_{goal} + \alpha \mathcal{H}(\pi_\theta(s|a)), \tag{3}$$

where α is a term that control influence of the entropy.

HILL CLIMBING WITH ROLLBACKS Since the policy gradient methods are prone to suffer from performance degradation over time (Kakade, 2002), we employ a rollback method when the policy starts to deviate from the objective. The performance of the model is monitored every few iterations on the development set. If the new model achieves greater rewards than the previously best model, the new model is saved. Contrarily, we roll back to the previous model that performed

best and continue from there following other exploration routes if the reward failed to improve for a while. When the policy degrades beyond recovery, the rollback in combination with the slot-value distribution sampling can give a way to a path that leads to greater rewards. We note our hill climbing with rollbacks strategy is an instance of a generalized version of the win-or-learn-fast policy hill climbing framework (Bowling and Veloso, 2001).

2.4 EXPERIMENTS

2.4.1 *Data*

We use the MultiWOZ dataset (Budzianowski et al., 2018) which consists of 10,438 dialogues spanning 7 domains: ATTRACTION, HOSPITAL, POLICE, HOTEL, RESTAURANT, TAXI and TRAIN. The dataset contains few dialogues in the POLICE and HOSPITAL domains, so we do not include these as the single domain dialogues in these domains did not contain belief state labels. The MultiWOZ dataset consists of natural conversations between a tourist and a clerk from an information center in a touristic city. There are two main types of dialogues. Single-domain dialogues include one domain with a possible booking sub-task. Multi-domain dialogues, on the other hand, include at least two main domains. MultiWOZ is much larger and more complex than other structured dialogue datasets such as WOZ2.0 (Mrkšić et al., 2017b), DSTC2 (Henderson, Thomson, and Williams, 2014) and FRAMES (El Asri et al., 2017). In addition, unlike the previous datasets, users can change their intent throughout the conversation, making state tracking much more difficult. Table 1 presents statistics of domains used in experiments with the distinction between the case when the dialogue consists of only one or more domains.

PREPROCESSING MULTIWOZ The user utterances and system utterances used to trained our models contain tokens that were randomly created during the creation of the data to simulate reference numbers, train IDs, phone numbers, arrival and departure times and post codes. We delexicalize all utterances by replacing these randomly generated values with a special generic token. In addition, we replace the turn label values with this special token and add that to the ontology. As mentioned by Mrkšić et al. (2017a), delexicalizing all values is not scalable to large domains as that requires to always have a dictionary holding all possible values. Therefore, we do not delexicalize any other values. Since MultiWOZ only contains the current belief state at each turn, we create the labels by registering the changes in the belief state from one turn to the next. The annotators were given instructions on specific goals to follow, however at times they did not

follow this goal. This lead to errors in the belief state such as wrong labels or missing information. These instances also propagate further down to our assigned gold turn labels. Furthermore, while preprocessing the data, we found that there are more values present than reported iin the ontology, therefore the number of values presented here is higher than what is reported in Budzianowski et al. (2018). We release our preprocessed data and preprocessing scripts.[1]

2.4.2 *Implementation Details*

Our pre-trained StateNet model is implemented without parameter sharing and is not initialized with single-slot pre-training as in Ren et al. (2018). We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-3}. We use an n-gram utterance representation size of 3 and 3 multi-scale receptors per n-gram. The supervised models are trained using a batch size of 16. The size of the GRUs hidden state is 200 and the size of the word embeddings is 400. In line with recent methods for dialogue state tracking, we use fixed pre-trained embeddings and do not update them during the training (Mrkšić et al., 2017a; Ren et al., 2018; Zhong, Xiong, and Socher, 2018). We use the established data splits for train, development and testing and apply early stopping if the joint goal accuracy has not improved over 20 epochs.

When fine-tuning with policy gradient, we evaluate on the development set every 5 batches, saving the model if the reward has increased since last. We use an independent hill climbing patience factor of 15, reverting back to the previous best model if no improvements were made in that period. We use a batch size of 16 in our fine-tuning experiments. When applying policy gradient methods in practice, larger batch sizes have shown to lead to more accurate policy updates (Papini, Pirotta, and Restelli, 2017), but due to the relatively small training sets we found a batch size of 16 gave us the best sample efficiency trade-off. Our implementation uses PyTorch (Paszke et al., 2017) and is publicly available.[1]

2.4.3 *Experimental Protocol*

SETUPS In our experiments, we report a number of different results: 1) Training a DST model M^P with the usual turn-level supervision on the different domains. We only use dialogues which strictly contains the labels of that single domain. We hypothesize that this serves as an upper bound to the performance of the policy gradient fine-tuning. 2) Evaluating the pre-trained models as a cross-domain zero-shot baseline. We take a model pre-trained on d^P and measure its performance on d^F for all domains in $D - d^P$. This serves as the

1 https://github.com/coastalcph/dialog-rl

lower bound for the performance of the policy gradient fine-tuned models. We use this baseline and not a model fine-tuned on d^F with cross entropy training with dialogue level supervision on the final belief state, as we simulate not having gold labels for each slot-value pair, but rather only a scalar rating as the sole signal. 3) Fine-tuning the pre-trained model M^P to all other domains with policy gradient as described in Section 2.3. We experiment with domain transfer from d^P to all domains in $D - d^P$ using only the user simulated dialog-level reward using policy gradient. 4) Lastly, we report the results of fine-tuning a model using policy gradient on the same domain it was pre-trained on, d^P, after convergence in order to see if the dialog-level reward signal can further improve its performance. We here use the same training and development data as the supervised model was trained on.

METRIC We measure the performance of our models with what we refer to as the *turn level accuracy* metric, which measures the ratio of how many of the gold turn labels are predicted by the DST model at each turn. The reported accuracy is the mean of all turns in the evaluation set.

2.5 RESULTS

In Table 2 we present the results from our baseline StateNet model and from policy gradient training for the in- and out-of domain scenarios. We also report the average out-of-domain accuracies for each domain, to illustrate how policy gradient training in general performs compared to the baseline. The table show the performance of transferring from each domain to all other domains. From the results we observe that in almost all domain transfer settings, with the exception of RESTAURANT to ATTRACTION, we get a consistent increase in performance when applying policy gradient fine-tuning, compared to the zero-shot transfer baselines. In some instances we also see an increase in performance from further fine-tuning a model after turn-level supervision convergence using only the dialogue-level reward feedback. In the case of ATTRACTION, we are even able to increase the accuracy by a large margin using in-domain policy gradient fine-tuning. On average, we see relative improvements of the accuracy, ranging from 0.03 to 0.2, when applying our proposed method of fine-tuning for DST domain transfer.

2.6 ANALYSIS

In order to illustrate the effectiveness of doing PG fine-tuning compared to doing zero-shot domain transfer, we plot in Figure 2a the results of training a model on the source domain HOTEL while eval-

Pre-train \ Finetune	TAXI		TRAIN		HOTEL		RESTAURANT		ATTRACTION	
	BL	PG	BL	PG	BL	PG	BL	PG	BL	PG
TAXI	0.35	0.35	0.17	**0.27**	0.04	**0.10**	0.12	**0.29**	0.00	**0.11**
TRAIN	0.13	0.13	0.43	0.43	0.07	**0.08**	0.08	**0.22**	0.00	0.00
HOTEL	0.004	**0.26**	0.02	**0.19**	0.30	**0.33**	0.10	**0.19**	0.06	**0.11**
RESTAURANT	0.04	**0.25**	0.13	**0.27**	0.11	**0.13**	0.33	**0.34**	**0.11**	0.05
ATTRACTION	0.00	**0.27**	0.00	**0.39**	0.00	**0.08**	0.05	**0.10**	0.11	**0.17**
AVERAGES	0.04	**0.23**	0.08	**0.28**	0.06	**0.10**	0.09	**0.2**	0.04	**0.07**

Table 2: Accuracy scores for our pre-trained baseline (BL) and the policy gradient fine-tuning (PG). The colored results along the left-to-right downward diagonal are in-domain results, dark red being the supervised results and light green the policy gradient fine-tuned results, and each pair of columns compare baseline and system results for each target domain. The AVERAGES row presents the average out-of-domain transfer scores for each domain. Note that while the PG method has access to more data, this does not invalidate the comparison, seeing that the additional data is relatively easy to obtain in an applied setting.

uating, on the development set, its zero-shot accuracy on the target domain TAXI, until convergence on the source domain. After convergence we show how the PG fine-tuning uses the pre-trained model as a starting point to further improve the accuracy on the target domain using only the dialog-level feedback. Figure 2a also illustrates the importance of the hill climbing technique we employ. When the performance starts to deteriorate, it manages to revert back to a reasonable baseline and improve performance from there instead. From the blue baseline curve, we also observe that even though the accuracy continuously improves on the source domain, this is not necessarily an indication of the performance on the target domain. On the contrary, performance suddenly starts to deteriorate for the latter when the model overfits to the source domain.

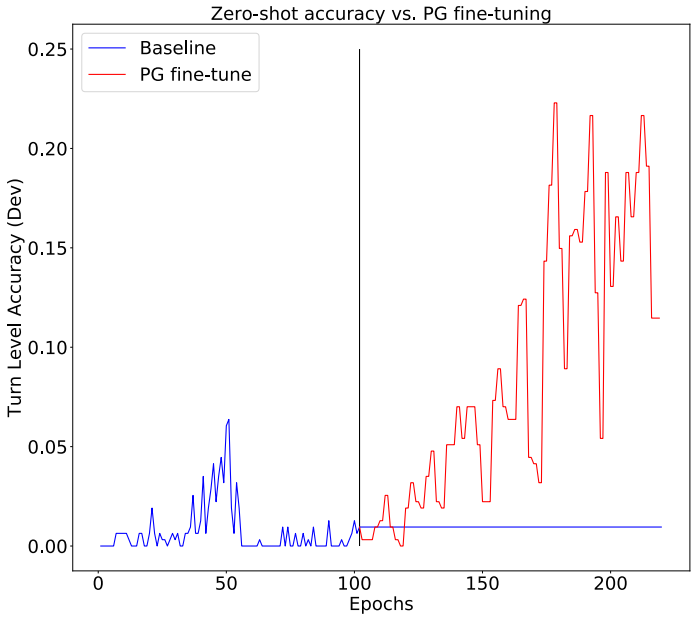
2.6.1 *Error Analysis*

In general we observe lower scores for both the baseline models and in-domain fine-tuning on the ATTRACTION domain. We believe this can be attributed to the fact that it only contains 150 dialogues, leaving very little data for the development and test splits. Coupled with the fact that it has 2 slots and 180 values, the risk of encountering unseen slot-value pairs increases significantly.

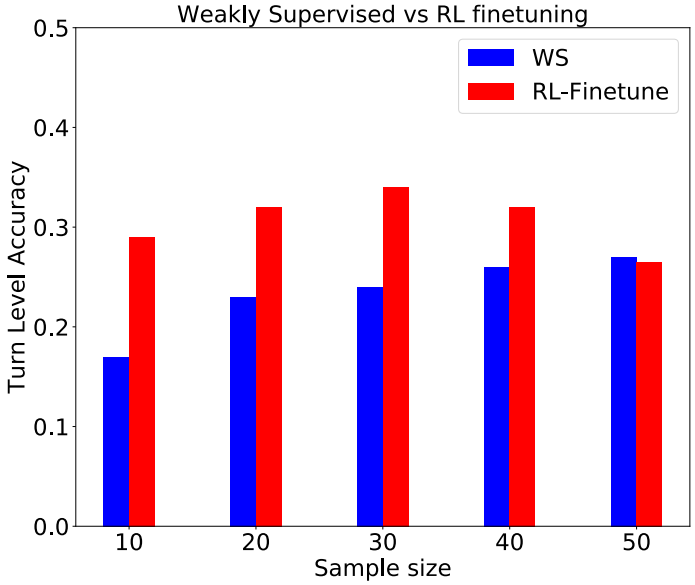
In Table 13 (see Appendix) we present a couple of example turns from the test set of the RESTAURANT domain, with the system utterance, user utterance and the predicted slot-value pairs for both the baseline model, which has been trained on the HOTEL domain, and the PG fine-tuned model. The slot-value pairs in green show correct predictions, whereas pairs in red show incorrect predictions. From the predicted slot-value pairs, we can for example see how the fine-tuned model to a better extent is able to utilize the user and system utterances to correctly predict what price range the user is looking for, even though the baseline correctly predicts the slot presence.

2.6.2 *Comparisons to Weak Supervision*

We also pose the question of how many annotated dialogues in the target domain are needed before policy gradient fine-tuning with dialogue-level rewards is no longer beneficial, compared to fine-tuning a model trained with turn-level cross entropy. In order to further investigate this, we use our pre-trained model in the TAXI domain and further finetune with varying amounts of dialogues i.e. $s \in [10, 20, 30, 40, 50]$ using turn level supervision for the RESTAURANT domain. We then fine-tuned each of the models on the RESTAURANT domain using the dialogue-level reward only. The results for these experiments are shown in Figure 2b. Overall, we find that when we annotate just 10 complete dialogues and then fine-tune our model using reinforcement learning



(a) The performance of the supervised model trained on the HOTEL domain while evaluated on the development set of the TAXI domain after each epoch until convergence on HOTEL versus the improvements we get from the policy gradient fine-tuning using the supervised model as starting point.



(b) The turn level accuracy of our weakly supervised fine-tuning compared to fine-tuning using PG. Performance plateaus after about 50 samples for both methods.

we still see an increase in performance. We observe that as we increase the sample size s for our weakly supervised models, fine-tuning using policy gradient comes with diminishing returns. At around 50 samples, the performance of the weakly supervised baseline reaches the performance of our system, and improvements from reinforcement learning, if any, become significantly smaller.

2.7 RELATED WORK

DST ARCHITECTURES The goal of Dialogue State Tracking is to predict the user intent or *belief state* at each turn of the conversation. The range of user goals or, *slots* and *value* pairs, that can possibly be recognized by the system are contained in the domain ontology. DST has for long been a part of spoken dialogue systems, however, before the Dialogue State Tracking challenges (Henderson, Thomson, and Williams, 2014; Williams et al., 2013) many of the early architectures relied on hand crafted rules (Sun et al., 2014, 2016; Wang and Lemon, 2013). Later research has proposed RNN models that exploit delexicalized features (Henderson, Thomson, and Young, 2014; Mrkšić et al., 2015; Rastogi, Hakkani-Tür, and Heck, 2017) in order to allow the model to perform better and achieve generalization by reducing the amount of labels. Delexicalization requires that all possible mentions of a slot and value are contained in a lexicon which does not become scalable in larger domains. To address this, Mrkšić et al. (2017a) proposed a neural belief tracker which uses pre-trained word embeddings to represent user utterances, system acts and current candidate slot-value pairs and utilizes these as inputs into a neural network. Recent approaches have proposed sharing parameters across estimators for the slot-value pairs (Nouri and Hosseini-Asl, 2018; Ramadan, Budzianowski, and Gasic, 2018; Ren et al., 2018; Zhong, Xiong, and Socher, 2018). Although not extensively investigated, this would make the model more scalable as the amount of parameters would not increase while the ontology size grows. In our experiments, we adopt the model by Ren et al. (2018) as our supervised baseline.

DOMAIN TRANSFER A key issue that remains unexplored by many of the existing methods within DST is domain adaptation. Williams (2013) presented some of the earliest work dealing with multi-domain dialogue state tracking, investigating domain transfer in two dimensions: 1) sharing parameters across slots, 2) sharing parameters across single domain systems. Later research further expanded by using disparate data sources in order to train a general multi-domain belief tracker (Mrkšić et al., 2015). The tracker is then fine-tuned to a single domain to create a specialized system that has background knowledge across various domains. Furthermore, Rastogi, Hakkani-

Tür, and Heck (2017) proposed a multi-domain dialogue state tracker that uses a bidirectional GRU to encode utterances from user and system which are then passed in combination with candidate slots and values to a feed-forward network. Unlike our proposed method, they rely on delexicalization of all values. In addition, their GRU shares parameters across domains. Ramadan, Budzianowski, and Gasic (2018) introduced an approach which leverages the semantic similarities between the user utterances and the terms contained in the ontology. In their proposed model, domain tracking is learned jointly with the belief state following Mrkšić and Vulić (2018). We want to emphasize that all previous models assume the existence of dialogue data annotated at the turn level in the new domain. In our proposed method, we model a more realistic scenario in which we only have a score of how accurate the system was at the end of the dialogue given the final user goal.

REINFORCEMENT LEARNING IN DIALOGUE In task-oriented dialogues, the reinforcement learning framework has mostly been used to tackle dialogue policy learning (Li, Williams, and Balakrishnan, 2009; Liu et al., 2018a; Singh et al., 2002; Williams and Young, 2007). Gasic et al. (2013) proposed a method to expand a domain to include previously unseen slots using Gaussian process POMDP optimization. While they discuss the potential of their model in adapting to new domains, their study does not present results in multi-domain dialogue management. Recent work has attempted to build end-to-end systems that can learn both user states and dialogue policy using reinforcement learning. Zhao and Eskenazi (2016) propose an end-to-end dialogue model that uses RL to jointly learn state tracking and dialogue policy. This model augments the output action space with predefined API calls which modify a query hypothesis which can only hold one slot value pair at a time. Dhingra et al. (2017) instead show that providing the model with the posterior distribution of the user goal over a knowledge base, and integrating that with RL, leads to higher task success rate and reward. In contrast to our work, Gašić et al. (2017) have tackled the problem of domain adaptation using RL to learn generic policies and derive domain specific policies. In a similar study, Chen et al. (2018) approach the problem of domain adaptation by introducing slot-dependent and slot-independent agents. Our approach differs from the previously presented models in several ways: a) we track the user state using RL, however, we do not learn generic and specific policies ; b) we use RL to adapt models across many domains and a large number of *slot,value* pairs; and c) we assume that a reward is only known for target domain dialogues at the end of each dialogue.

2.8 CONCLUSION

This paper tackles the challenge of transferring dialogue state tracking models across domains without having target-domain supervision at the turn level; that is, without manual annotations, which are costly to obtain. Our setup is motivated by the fact that in a practical setting it is much more feasible to obtain dialogue level signals such as user satisfaction. We introduce a transfer learning method to address this, using supervised learning to learn a base model and then using reinforcement learning for fine-tuning using our dialogue level reward. Our results show consistent improvements over domain transfer baselines without fine-tuning, at times showing similar performance to in-domain models. This suggests that with our approach, dialog-level feedback is almost as useful as turn-level labels. In addition, we show that using the dialogue-level reward signal for fine-tuning can further improve supervised models in-domain.

3

WHAT DO YOU MEAN 'WHY?': RESOLVING SLUICES IN CONVERSATIONS

ABSTRACT

In conversation, we often ask one-word questions such as 'Why?' or 'Who?'. Such questions are typically easy for humans to answer, but can be hard for computers, because their resolution requires retrieving both the right semantic frames and the right arguments from context. This paper introduces the novel ellipsis resolution task of resolving such one-word questions, referred to as *sluices* in linguistics. We present a crowd-sourced dataset containing annotations of sluices from over 4,000 dialogues collected from conversational QA datasets, as well as a series of strong baseline architectures.

3.1 INTRODUCTION

Stand-alone *wh*-word questions, such as *When?* in Figure 3, are easy for us to understand, but in order to interpret them we need to retrieve implicit information from context. Learning to do so is an instance of *sluicing*, an ellipsis phenomenon, defined by Ross (1969) as 'the effect of deleting everything but the preposed constituent of an embedded question, under the condition that the remainder of the question is identical to some other part of the sentence, or a preceding sentence.' In the context of conversations, one-word *wh*-word questions are particularly frequent (Anand and Hardt, 2016; Rønning, Hardt, and Søgaard, 2018), and because they are often hard to resolve, they seem to be a frequent source of error in conversational question answering (Choi et al., 2018; Reddy, Chen, and Manning, 2018) and dialogue understanding (Vlachos and Clark, 2014). We refer to this type of sluicing as *conversational sluicing*.

Unlike previous work where sluice resolution is treated as predicting the span of the antecedent (Anand and Hardt, 2016; Rønning, Hardt, and Søgaard, 2018), we frame conversational sluice resolution as a Natural Language Generation (NLG) task, in which we seek to automatically generate the full question, given a question-answer context and a one-word question. To this end, we provide a novel corpus of conversational sluice annotations and explore a series of strong baselines and their performance on this dataset.

CONTRIBUTIONS In this paper we introduce the task of resolving conversational sluicing, a pervasive and challenging ellipsis phe-

Q_1:	Where was the bombing?
A_1:	San Diego's Edward J. Schwartz Federal Courthouse.
Q_2:	When?
R_1:	When [*was the bombing*?]
R_2:	When [*was the bombing of San Diego's Edward J. Schwartz Federal Courthouse*]?

Figure 3: Example of conversational sluicing. Q_1 and A_1 provides a context for the second question Q_2 which has multiple correct resolutions, denoted in brackets, such as R_1 and R_2.

nomenon. We crowd-source a new dataset containing over 4000 annotated sluices, gathered from existing conversational QA datasets. We conduct a series of baseline experiments on this task, using both encoder-decoder frameworks, as well as language modelling objectives, and show through human evaluation of the predicted resolutions that these baselines are quite strong and at times even rival the quality of human annotators.

3.2 BACKGROUND

SLUICING Ellipsis is the linguistic phenomenon that describes the omission of one or more words from a phrase that can be retrieved from a previous context. Sluicing is a case of ellipsis where content is elided from a question, leaving behind only the *wh*-remnant. Anand and Hardt (2016) and Rønning, Hardt, and Søgaard (2018) consider two types of sluices, namely *embedded* sluices and *root* sluices, also sometimes referred to as *bare* sluices.

(1) My neighbor said he would stop by, but I don't know when [*he would stop by*].

(2) a. My neighbor is stopping by.
b. When [*is the stopping by*]?

In Example (1), we see an instance of embedded sluicing where the question is a part of a larger structure, and (2) is an example of a root sluice where the *wh*-fronted ellipsis is an utterance in itself, i.e. in a root environment. Anand and Hardt (2016) note that sluicing in dialogue often differs from sluicing in single-authored text, with root sluices being more prevalent in dialogue. In dialogue, using sluices – and ellipsis in general – requires a level of mutual understanding. Colman, Eshghi, and Healey (2008) therefore use ellipsis in dialogue as a means of quantifying mutual understanding in conversations.

Fernández, Ginzburg, and Lappin (2007) focus on the task of classifying occurrences of single-word sluices in conversations and call

these *bare* sluices. They categorize such sluices into distinct categories; (i) *direct*, which is the case where the sluice queries for additional information that was quantified, either explicitly or implicitly, in the previous utterance; (ii) *reprise*, where the speaker is unable to understand an aspect of the previous utterance, which the initial speaker assumed as presupposed; (iii) *clarification*, where the speaker uses the sluice to ask for clarification of the entire preceding utterance; (iv) *Wh-anaphor*, where the antecedent is a wh-phrase; and (v) *unclear*, the case where it is difficult understand what the sluice conveys, usually because of a lack of proper context. Note that the direct, reprise and clarification sluices are relatively easier to resolve, since their answer can always be retrieved from the previous sentence. Our corpus therefore ignores the first three types of conversational sluices and focuses on (bare or stand-alone) *wh*-anaphors; in our annotation experiments below, we also allow annotators to skip unclear instances. Similarly, Baird, Hamza, and Hardt (2018) presented classification experiments learning to distinguish between different types of sluices in dialogue.

Conversational sluices usually depend on their question-answer context, and can span both the previous utterances, i.e. the answer, as well as the previous question, whereas direct/reprise/clarification sluices only requires retrieval of context from the previous utterance. Consider the multi-turn example:

A: Did Ned have family?
B: Yes.
A: Who [*was Ned's family*]?

Resolving this sluice, depends on both the question initially asked by speaker A in addition to the outcome of the answer from speaker B. Looking only at the previous utterance, in this case, would not provide sufficient context, as the *Yes/No* utterance of speaker B determines what information from speaker A is relevant for the resolution.

The first efforts to resolve (non-conversational, standard) sluices, by identifying the antecedent of the *wh*-remnant, is due to Anand and McCloskey (2015), who describe a linguistically-informed annotation scheme for resolving sluices. They present a dataset of 3.100 annotated examples of sluices extracted from the New York Times section of the English Gigaword corpus. Anand and Hardt (2016) presented the first sluice resolution system, achieving decent performance, but Rønning, Hardt, and Søgaard (2018) subsequently presented a neural multi-task architecture outperforming their original model by some margin.

A few researchers have explored ellipsis resolution in dialogue: Kazuhide and Eiichiro (1998) discussed the importance of being able to resolve sluices to understand dialogue. They showed that for certain types of conversational ellipsis, it is possible to achieve good results with simple classification algorithms. Their results are not com-

parable to other results in the literature, because they focus on a small subset of phenomena, rely on linguistic preprocessing, and consider ellipsis phenomena in Japanese. Rønning, Hardt, and Søgaard (2018) also evaluates on conversational data from English Open Subtitles. Their results suggest that resolving sluices in dialogue is harder than domains such as newswire, with F_1 resolution scores dropping from > 0.7 in newswire to around 0.5 for conversations. As stated, these previous approaches to sluice resolution differs from ours, as we seek to generate a reconstruction of the sluice, not predict the span of the antecedent. Due to the fact that in a conversational context, the antecedent is conditioned on the response to the initial question in our question-answer context, it often results in disjoint antecedent spans, which cannot be represented in the architecture proposed by Rønning, Hardt, and Søgaard (2018). The advantage of resolving the sluice using NLG approaches is that for most downstream purposes, a fluent paraphrase of the *wh*-word and the antecedents is preferred and not only an antecedent span, that as stated above, can be non-coherent.

QUESTION GENERATION Researchers have worked on question generation from text paragraphs (Zhao et al., 2018c), relative clauses (Khullar et al., 2018), SQL queries (Guo et al., 2018), knowledge bases (Serban et al., 2016), etc. Khullar et al. (2018), which is probably the problem set-up most similar to ours, albeit much simpler, consider relative clauses such as in *I am giving fur balls to John who likes cats*. Their simple observation is that relative clauses translate almost straightforwardly into questions, e.g., *Who likes cats?*. Using a small set of heuristic rules, they extract relative clauses and use them to generate training data for machine comprehension. Our task is considerably harder, since we deal with an ellipsis phenomenon that requires us to find antecedents in the previous dialogue turns. Our approach is also very different. While Khullar et al. (2018) can solve their problem with simple rules, we cannot, and we therefore present neural baseline architectures originally developed for language modeling and transduction tasks.

3.3 A CONVERSATIONAL SLUICING DATASET

In this work, we present a crowd-sourced annotated sluicing dataset. The dataset consists of sluice occurrences in conversational question answering contexts. The conversations are teacher-student dialogues, where the teacher asks questions about a background text passage, and the student has to answer the teacher's questions. Sluices, and ellipsis in general, are frequent in the data. Each datapoint consists of (i) an initial question, Q_1, (ii) an answer to Q_1, A_1, together forming the QA context (Q_1, A_1), (iii) a one word follow-up *wh*-question, Q_2, (iv) a gold annotated resolution, R, to the sluice in (iii), written

in free-text. The resolutions are what we crowd-source to construct the new conversational sluicing dataset. Given question-answer context pairs (Q_1, A_1) and one-word follow-up questions Q_2, we seek to resolve conversational sluices by generating the full questions R by explicitly generating the elided context, therefore framing it like a NLG task, rather than an antecedent selection task as done by Rønning, Hardt, and Søgaard (2018) and Anand and Hardt (2016). This also dramatically simplifies the annotation process as we only seek a resolved sluice in the form of R instead of the annotation scheme used by Anand and McCloskey (2015) and Rønning, Hardt, and Søgaard (2018), i.e. explicitly annotating the antecedent, sluiced expression, main predicate of the antecedent clause as well as potential correlates in addition to annotations for the auxiliary tasks.

This section describes the process of collecting and cleaning the annotations, and presents a quantitative and qualitative analysis of the dataset.

DATA COLLECTION METHODOLOGY In order to obtain our conversational sluicing dataset, we crawl existing conversational QA datasets, namely QuAC[1] and CoQA[2] (Choi et al., 2018; Reddy, Chen, and Manning, 2018), for question-answer contexts with one-word follow-up questions. Specifically, we identify all occurrences of five one-word questions: *Why?*, *What?*, *Where?*, *Who?* and *When?*. For each such question, we construct a tuple of the previous QA context and the follow-up question. This process results in roughly 4200 examples of conversational sluices.

We then proceeded to ask Amazon Mechanical Turkers (AMT) to fill out the remainder of the question as asked by the interrogator based on the the question-answer context pair. In order to not impose too many restrictions on the annotators, we left it up to the AMT workers to decide how much of the elided information they wanted to include in their answer, as a conversational sluice can often be solved in multiple ways. For example, in Figure 3 we consider both R_1 and R_2 as correct resolutions to the conversational sluice, even if R_1 did not specify the PPN *San Diego's Edward J. Schwartz Federal Courthouse* as the location of the bombing. In general, annotations often differed in whether modifiers and relative clauses were included, whether or not previous anaphora was resolved, etc. If the previous question and answer did not provide enough context to fill out the elided information, the workers were informed to simply skip it and move on to the next example. We collected a single annotation for each sluice in the training and test splits, and three annotations for each sluice in the test set. For the test set, we use each unique annotation as a separate datapoint. We allocated 1 minute per annotation

1 https://quac.ai/
2 https://stanfordnlp.github.io/coqa/

and paid the workers $0.13 for each accepted annotation. The average time spent per assignment was around 20 seconds, which resulted in an hourly rate of $23.4. The total cost of the crowd-sourcing process was $797.

For our final corpus, we filter out the examples skipped by the annotators, in addition to the conversational sluices whose Q_1 context is less than 3 words, as these showed empirically to not contain enough information, usually due to Q_1 being a sluice itself. Consider, for example:

Q_1: By who?
A_1: Unknown assailants.
Q_2: Where?

Without first resolving the sluice *By who?*, we are unable to properly identify the antecedent, as it is unclear whether or not Q_2 refers to the current location of the assailant or the location of the actual assault. These are also the sluices categorized as *Unclear* by Fernández, Ginzburg, and Lappin (2007). After cleaning, we reduced the initial size from 4980 to 4175 datapoints.

CORPUS STATISTICS In Table 3, we show the distribution of the different *wh*-questions across the various splits in our corpus. The dataset contains both instances of conversational sluices as well as reprise/direct/clarification sluices. We release the raw annotated version of the conversational sluicing corpus, as well as our cleaned version which we report our results on, including the splits used.[3]

Split	*Why*	*Where*	*Who*	*What*	*When*	Total
train	851	714	513	302	702	3082
val	84	71	54	39	52	300
test	229	183	97	83	201	793
Total	1164	968	664	424	955	4175

Table 3: Statistics of the *wh*-word distribution across the different splits for our conversational sluicing dataset.

Empirically, we did not observe many long distance dependencies between the sluice and corresponding antecedent, as it was found within a three-turn window a majority of the time (around 95%). Rønning, Hardt, and Søgaard (2018) similarly reports that long term dependencies (3 or more sentences between sluice and antecedent) are

3 https://github.com/vpetren/conv_sluice_resolution

very rare (around 1%). Solving these rare dependencies would also be an interesting task, but is however outside the scope of this work. This dataset provides a reasonable limitation for a stab at an already challenging phenomenon.

PERFORMANCE METRICS Natural language generation systems are often evaluated in terms of BLEU scores (Papineni et al., 2002) and on subsamples of standard corpora. Neither are likely to be optimal. Finding an appropriate performance metric that correlates with human judgments of resolution quality, is crucial to ensure progress on conversational sluicing resolution; and evaluating across different samples is equally important to avoid community-wide over-fitting to one particular sample. We hope to be able to contribute to improving both performance metrics and the data situation, but for now we also report the performance of our baseline systems in terms of BLEU scores on a random subsample. In order to combat the bias introduced by BLEU, we supplement the scores with alternative performance metrics, as well as with human judgments from professional annotators. BLEU originally was intended for corpus-level evaluation and has several limitations when applied at the sentence-level (Rapp, 2009). We therefore also include the GLEU metric, as proposed by (Wu et al., 2016), which according to their experiments, is better suited for sentence-level evaluation, while still correlating well with BLEU on the corpus-level.[4] In addition to BLEU and GLEU we also measure the the character n-gram F-score (CHRF) (Popović, 2015), as well as the precision (CHRP) and recall (CHRR). We use $\beta = 3$, i.e. assigning a higher weight to recall, as it has been shown to correlate better with human judgements than other popular automatic machine translation metrics, such as BLEU and ROGUE-L. For n we use 4-grams.

Given the shortcomings of automatic evaluation metrics, we also include a human evaluation study. We sample n contexts along with the gold sluice resolution and the resolutions generated by our baseline models from the test set and ask human evaluators to rank them according to relative quality. We obtained judgments of 100 document instances and report on these experiments in §3.5.

ANNOTATION QUALITY In the last row of Table 4, ANN AGREE, we report the inter-annotator agreement scores of the test set. For each of the 3 collected annotation per conversational sluice instance, we sample 2 of them and calculate BLEU, GLEU, chrF, chrP and chrR scores between them as a measurement of annotator agreement. As different annotations can be considered correct sluice resolutions, we use this measurement as a means to set an expectation for the performance

4 We use the sentence-level GLEU and BLEU implementations provided by NLTK with the smoothing function introduced by Lin and Och (2004)

Model	GLEU	BLEU	CHRF	CHRP	CHRR
C&E Q1	0.035	0.043	0.114	0.034	0.166
C&E A	0.010	0.016	0.034	0.011	0.048
LSTM-SEQ2SEQ	0.232	0.304	0.276	0.311	0.274
TRANSFORMER	0.337	**0.391**	0.443	0.461	0.442
GPT-2	0.067	0.117	0.138	0.109	0.167
GPT-2 (FT)	**0.348**	**0.391**	**0.467**	**0.499**	**0.470**
ANN AGREE	0.570	0.589	0.712	0.704	0.720

Table 4: Results on our conversational sluicing dataset for a series of baseline architectures. We measure the performance using BLEU, GLEU and character n-gram F-score, precision and recall on the test split. In the last row, ANN AGREE denotes the inter-annotator agreement as the average between two randomly sampled gold annotations from each data point of the test set.

ceiling of our models. In general we observe that there seems to be reasonably high annotator agreement scores compared to the best performing models, but still indicates that the sluices can be solved in multiple correct ways.

3.4 EXPERIMENTS

In our experiments, we use the splits outlined in Table 3 (also made publicly available). We preprocess our data by appending the QA context and one-word question together, converting the input sequence into the format `<s> Q1 <del> A1 <del> Q2 </s>` and the target sequence we seek to generate as `<s> R </s>`. Here `<del>` is a special delimiter token, and `<s>` and `</s>`, denote the beginning and end of the sequence. In addition to this, we only preprocess the data by performing lower-casing and tokenization.

3.4.1 *Baseline models*

In this section, we present a number of different baseline architectures and heuristics for the task of conversational sluice resolution.

COPY & EDIT HEURISTICS Seeing as the structure of the resolved sluice in some cases takes on the form of either Q_1, especially in the cases where a yes/no answer precedes it, or A, as seen in Figure 3, we propose two simple copy and edit heuristics. (i) Given the QA-context and our conversational sluice Q_2, we simply replace the wh-question word in Q_1 with Q_2 and use this augmented question as the

Model	MRR	r_1
LSTM-SEQ2SEQ	0.295	0.005
TRANSFORMER	0.381	0.030
GPT-2 (FT)	**0.529**	**0.190**
GOLD	0.879	0.775

Table 5: The results of the human judgement experiment. To obtain human judgments, we asked three annotators to rank the output of three systems and the crowd-sourced gold annotations. MRR is the mean reciprocal ranking, and r_1 refers to the fraction of presented examples where the model was ranked as number 1. Our results show that the fine-tuned GPT-2 model produces favorable resolutions, both in terms of automatic as well as human evaluation and 1/5 instances *better* than gold annotations.

resolution to our sluice. We refer to this as C&E Q1. (ii) Similarly, we can copy the answer from A and prepend the Q_2 sluice to it. We refer to this as C&E A.

LSTM-SEQ2SEQ Sequence-to-sequence models (Sutskever, Vinyals, and Le, 2014), or seq2seq, have previously been successfully applied to conversational modelling tasks (Vinyals and Le, 2015). They use the encoder-decoder framework, where an input context is encoded by an encoder-module, usually a variant of Recurrent Neural Networks (RNNs), and decoded by a decoder-module, into the target sequence. For both the encoder and decoder, we use a standard two-layer LSTM (Hochreiter and Schmidhuber, 1997), with a hidden state size of 512, and regularized using a dropout rate of 0.5. We initialize the embedding matrix with 300 dimensional GloVe (Pennington, Socher, and Manning, 2014), which remains fixed during training. We optimize the end-to-end network using Adam (Kingma and Ba, 2015), with the default learning rate of 0.001.[5]

TRANSFORMER The transformer architecture (Vaswani et al., 2017) is now the *de facto* standard architecture in machine translation and has paved the way for state-of-the-art pre-trained contextual language encoders such as BERT (Devlin et al., 2019) and the OpenAI GPT-2 (Radford et al., 2019). While still adopting the encoder-decoder framework, instead of processing the source and target sequences sequentially, it relies on a multi-headed self-attention mechanism, attending over the entire sequence at same time, allowing for greater parallelization and a positional encoding of the sequence, ensures that contextual information is maintained. As our conversational sluicing resolution corpus is small in comparison to the corpora used in the

5 Implementation is based on https://github.com/bentrevett/pytorch-seq2seq.

experiments by Vaswani et al. (2017), we limit ourselves to three encoder/decoder layers to 3 (compared to 6 in their work), after observing improvements on our validation data.[6] As with the LSTM-seq2seq model, we initialize the embedding matrix with 300 dimensional GloVe embeddings, but otherwise we use the defualt hyperparameters.

GPT-2 The generative pre-trained transformer (GPT-2) (Radford et al., 2019), trained to simply predict the next word in 40GB of Internet text, has since its introduction been used to generate state-of-the-art performance on multiple language modelling datasets. The GPT-2 architecture, as mentioned above, is based on the transformer architecture. In our experiments, we use the small pre-trained model released by OpenAI (117M parameters). We experiment both with the pre-trained GPT-2 model as is, as well as with fine-tuning it on our sluicing corpus. When fine-tuning the model, we simply concatenate the input and output sequences together and input them to the language model. Unlike the LSTM-seq2seq and Transformer, we do not fine-tune the GPT-2 model until convergence, but instead we ran it for 18 hours on an Nvidia TitanX GPU. We also report the performance of the GPT-2 model on our task when no fine-tuning has taken place.

OTHER BASELINES CONSIDERED Inspired by Hill, Cho, and Korhonen (2016) and Lample et al. (2018), we also experimented with pre-training the SEQ2SEQ-LSTM and TRANSFORMER architectures with sequential de-noising autoencoder objectives. We collected a dataset consisting of 350.000 questions from CoQA, QuAC and SQuAD 2.0, making sure not to include cases of sluices, hypothesizing that this would allow the encoder and decoder to learn the internal structure and representation of questions. After pre-training, we fine-tune the architectures on our conversational sluicing data. These experiments did, however, not lead to any improvements in the performance when using automatic metrics. A manual inspection of the generated resolutions did not reveal any noticeable improvements over their non pre-trained counterparts, so we do not report the results below.

Again, we stress that due to the reasons listed above, i.e. incompatible annotation schemes between our work and that of Rønning, Hardt, and Søgaard (2018) as well as the lack of flexibility that a span-prediction model provides, we do not use their work as a baseline. We hypothesize that our heuristics, C&E Q1 and C&E A, will serve as an indication as to what we can expect from these types of models.

6 Implementation is based on https://github.com/jadore801120/attention-is-all-you-need-pytorch/

3.4.2 *Results*

Table 4 summarizes the results from our baseline models on our conversational sluicing corpus, using standard automatic performance metrics. The results suggest that the fine-tuned GPT-2 architecture is superior to all other baselines across the board, achieving scores closest to the inter-annotator ceiling, with the TRANSFORMER model rivalling it on the BLEU score. Although the C&E Q1 and C&E A heuristics could seem like strong baselines, as some of the examples in Table 6 and Figure 3 might suggest, our results tells a different story. Again, this illustrates the flexibility that is required to resolve these conversational sluices, which a non-disjoint antecedent span fails to capture. We can observe that without the task-specific fine-tuning, the GPT-2 model falls short, as it ultimately just proceeds to generate what comes after the sluice, not resolving it. However, this extensive pre-training does shine through compared to the TRANSFORMER model, when fine-tuned on our dataset as we also can see from our human evaluation (illustrated in Table 5), which we discuss in the next section.

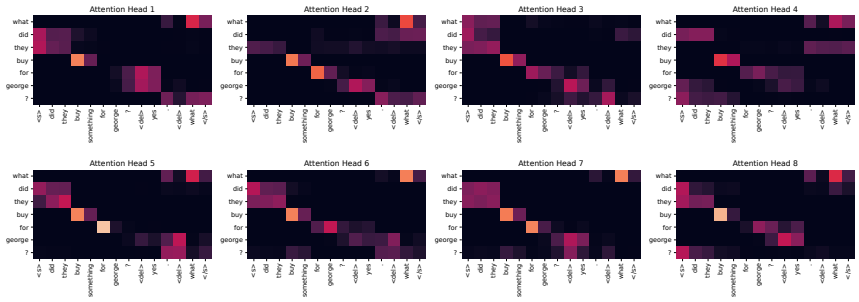


Figure 4: Illustration of the attention weights from all the 8 attention heads in the final decoder layer of the Transformer network. The x-axis corresponds to the position in the input sequence, whereas the y-axis corresponds to the output sequence.

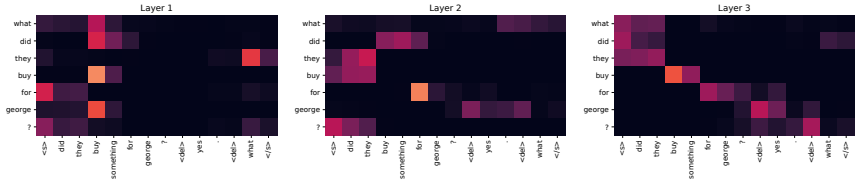


Figure 5: Illustration of the attention weights from a single attention head in the 3-layer Transformer network, during decoding. The x-axis corresponds to the position in the input sequence, whereas the y-axis corresponds to the output sequence.

CONTEXT		LSTM-S2S	TRANSFORMER	GPT-2	GOLD
Q_1:	What did Susie do?	When did they go??	When did Susie woke up?	When did Susie wake up?	When did Susie wake up?
A_1:	Woke up.				
Q_2:	When?				
Q_1:	Did the island ever change its form of government?	When did the objective of the?? Scotland?	When did the island change its form of government?	When did the island form?	When did the island change its form of government?
A_1:	Yes.				
Q_2:	When?				
Q_1:	Is there any mysterious character?	Who is the other??	Who are the character in?	Who was the famous person that was added to the story?	Who is the mysterious character?
A_1:	Yes.				
Q_2:	Who?				
Q_1:	Did he say anything before leaving?	What did he do ?	What did he say?	What did he say before he left?	What did he say?
A_1:	Yes.				
Q_2:	What?				

Table 6: Generated output from our series of baselines, given a question-answer context, (Q_1, A_1) and follow-up one-word question. Examples are taken from the test split.

3.5 ANALYSIS

HUMAN JUDGMENT OF GENERATED RESOLUTIONS Knowing that our automatic evaluation metrics can be biased when applied at the sentence-level, we also include a human evaluation study on a random sample of 100 instances of sluices. We asked human evaluators to rank the resolutions generated by our best performing models, i.e. the LSTM-SEQ2SEQ architecture, the TRANSFORMER architecture, our fine-tuned GPT-2 model, as well as the human annotators' resolutions, by their quality and relevance in a QA context. We presented the four resolutions in random order and asked subjects to place them, from best to worst. If they deemed two or more candidates to be equally good or bad, we instructed them to simply order these randomly. We report performance using the Mean Reciprocal Rank (MRR), and what we refer to as r_1, which denotes the fraction of presented examples where the model was ranked as number 1. Our evaluation, shown in Table 5, reveals that the human judges tend to favour the resolutions provided by GPT-2 (FT) over the ones produced by the TRANSFORMER architecture. In fact, the GPT-2 resolutions are chosen over all other resolutions, including our gold standard, in 1/5 instances. Generally we see the same trend in the human evaluation experiment as with the automatic metrics, except that the GPT-2 model now significantly outperforms the other baselines. We believe this can be attributed to the fact that our human judges may be biased toward selecting well-formed resolutions, and the GPT-2 language model may simply be better at generating fluent language.

To illustrate an instance where GPT-2 can generate a more expressive resolution than our gold standard, consider the example in Figure 6. Here, the fine-tuned OpenAI GPT-2 model generates a resolution that the judges found to be better than the gold standard, not

because the gold-standard was wrong, but because the automatic resolution was more informative, easing interpretation.

Q_1:	Is anyone who works with them mentioned?
A_1:	Yes.
Q_2:	Who?
R_{GPT2}:	Who [else is mentioned]?
R_{Gold}:	Who [is mentioned]?

Figure 6: Conversational sluice resolution by the fine-tuned GPT-2 model that is judged better than the gold standard by our annotators.

VISUALIZATION OF ATTENTION WEIGHTS An advantage of the attention mechanism, is that it allows for high interpretability, when it comes to the showing where in the input sequence the model is attending at a given time-step. To get a better understanding of where the Transformer attends during decoding, we visualize the internal attention mechanisms of the model trained on our conversational sluicing corpus. Figure 5 shows the attention matrix heatmaps of a single attention-head in each layer and Figure 4 shows the attention matrix heatmaps for each of the 8 attention-heads in the last layer of the Transformer. When looking at Figure 5, we see that the various layers encode different levels of information, with the attention-head of the last layer seemingly being the most structured. From Figure 4, we can observe that the various attention-heads mostly present the same pattern. When generating the first word of the resolution, the attention is at the end of the input sequence, i.e. on the *wh*-fronted ellipsis. Generating the subsequent tokens then shifts the attention back to the beginning of the input sequence and learns to integrate the information of the question-answer context, as the resolution of the conversational sluice tends to repeat the structure of the antecedent of both the question and answer.

INSPECTION OF MODEL OUTPUT Table 6 present examples of conversational sluices from the test set along with the resolutions generated by our baselines as well as a gold annotated resolution. From the examples, we can observe that the LSTM-SEQ2SEQ often produces more nonsensical and less grammatically correct sentences, e.g. overusing question marks and inserting them in the middle of the sentences and it generally performs best when the input context and resolutions are short.The output of the TRANSFORMER does improve upon the results of the LSTM-SEQ2SEQ, producing more correct and coherent sentences, however, the lack of pre-training compared to GPT-2, still results in less expressive sentences. Most impressive are the results from the fine-tuned GPT-2 model. From its r_1 value we can see

that almost 20% of the instances, it actually generates a sluice resolution that our human judges ranked higher than the gold resolution. E.g., in the last sample generated by the GPT-2 model, demonstrates how it is able to incorporate all the information of the initial question Q_1, to a much higher degree than the what the annotator noted. The extensive pre-training does however allow the generated output to deviate a bit too much from the objective, as seen in the 3rd row.

APPLYING SLUICE RESOLUTIONS IN QA SYSTEMS As mentioned in §3.1, the ability to resolve occurrences of ellipsis, either implicitly or explicitly, is important for question-answering system. With our gold annotated sluice resolutions, we replace instances of conversational sluices in the CoQA development set with their resolved counterparts, and evaluate the quality of the answers their baseline model provides.[7] In Figure 7, we see how the resolution of the conversational sluice leads to a much better answer, $A_{no-sluice}$, compared to the case where the model has to automatically draw the connection between *'Why?'* and the context in Q_1 and A_1. Of course injecting our

Q_1:	What did Valetta think Mysie mustn't do?
A_1:	Stay out after dark.
Q_2:	Why [*does Valetta think that Mysie shouldn't stay out after dark*]?
$A_{no-sluice}$:	For fear she should cough.
A_{sluice}:	no.
A_{gold}:	Fear she should cough.

Figure 7: A case where resolving the sluice in the an instance of the CoQA dataset improves the performance of QA system. $A_{no-sluice}$ is the answer generated when information contained in the bracket is included.

annotations into the input at test time also biases the input data, making it less similar to the training data, and for this reason resolving sluices this way did not lead to significant improvements on average.

3.6 CONCLUSION

This paper addresses the challenge of resolving occurrences of conversational sluices; that is, correctly identifying the antecedent of a bare *wh*-fronted ellipsis in a dialogue setting. We frame the task as a language generation task, where we seek to generate the elided material. To this end, we crowd-sourced a new dataset of conversational sluices. We evaluate the performance of encoder-decoder ar-

7 Code for the pre-trained CoQA baseline model is provided by https://github.com/stanfordnlp/coqa-baselines

chitectures and language models on this data and show that human judges favour the resolutions generated by GPT-2, fine-tuned on our crowd-sourced annotations. Interestingly, resolutions rival the quality of human annotations.

4

GUIDELINE BIAS IN WIZARD-OF-OZ DIALOGUES

ABSTRACT

NLP models struggle with generalization due to sampling and annotator bias. This paper focuses on a different kind of bias that has received very little attention: *guideline bias*, i.e., the bias introduced by how our annotator guidelines are formulated. We examine two recently introduced dialogue datasets, CCPE-M and Taskmaster-1, both collected by trained assistants in a Wizard-of-Oz set-up. For CCPE-M, we show how a simple lexical bias for the word *like* in the guidelines biases the data collection. This bias, in effect, leads to poor performance on data without this bias: a preference elicitation architecture based on BERT suffers a 5.3% absolute drop in performance, when *like* is replaced with a synonymous phrase, and a 13.2% drop in performance when evaluated on out-of-sample data. For Taskmaster-1, we show how the order in which instructions are presented, biases the data collection.

4.1 INTRODUCTION

Sample bias is a well-known problem in NLP – discussed from Marcus (1982) to Barrett et al. (2019) – and annotator bias has been discussed as far back as Ratnaparkhi (1996). This paper focuses on a different kind of bias that has received very little attention: *guideline bias*, i.e., the bias introduced by how our annotator guidelines are formulated.

Annotation guidelines are used to train annotators, and guidelines are therefore in some sense intended to and designed to prime annotators. What we will refer to in our discussion of guideline bias, is rather the unintended biases that result from how guidelines are formulated, and the examples used in those guidelines. If a treebank annotation guideline focuses overly on parasitic gap constructions, for example, inter-annotator agreement may be higher on those, and annotators may be biased to annotate similar phenomena by analogy with parasitic gaps.

We focus on two recently introduced datasets, the Coached Conversational Preference Elicitation corpus (CCPE-M) from Radlinski et al. (2019), related to the task of conversational recommendation (Christakopoulou, Radlinski, and Hofmann, 2016; Li et al., 2018), and Taskmaster-1 (Byrne et al., 2019), which is a multi-purpose, multi-domain dialogue dataset. CCPE-M consists of conversations about

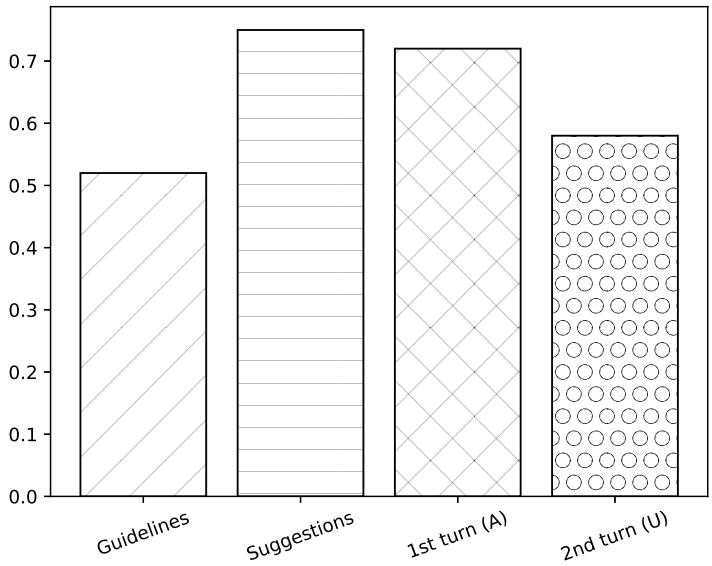


Figure 8: The percentage of sentences with the word *like* in the CCPE-M annotation guidelines (Guidelines), the suggested questions to ask users, in the guidelines (Suggestions), the *actual* first turns by the assistants (1st turn), and the actual replies by the users (2nd turn). In all cases, more than half of the sentences contain the word *like*.

movie preferences, and the part of Taskmaster-1, we focus on here, conversations about theatre ticket reservations. Both corpora were collected by having a team of assistants interact with users in a Wizard-of-Oz (WoZ) set-up, i.e. a human plays the role of a digital assistant which engages a user in a conversation about their movie preferences. The assistants were given a set of guidelines in advance, as part of their training, and it is these guidelines that induce biases. In CCPE-M, it is the overwhelming use of the verb *like* (see Figure 19 in the Appendix) and its trickle-down effects, we focus on; in Taskmaster-1, the order of the instructions. In fact, the CCPE-M guidelines consist of 324 words, of which 20 (**6%**) are inflections or derivations of the lemma *like*: As shown in Figure 19 in the Appendix, more than **50%** of the sentences in the guidelines include forms of *like*! This very strong bias in the guidelines has a clear downstream effect on the assistants that are collecting the data. In their first dialogue turn, the assistants use the word *like* in **72%** of the dialogues. This again biases the users responding to the assistants in the WoZ set-up: In **58%** of their first turns, given that the assistant uses a form of the word *like*, they also use the verb *like*. We show that this bias leads to overly optimistic estimates of performance. Additionally, we also demonstrate how the guideline affects the user responses through a controlled priming experiment. For Taskmaster-1, we show a similar effect of the guidelines on the collected dialogues.

CONTRIBUTIONS We introduce the notion of *guideline bias* and present a detailed analysis of guideline bias in two recently introduced dialogue corpora (CCPE-M and Taskmaster-1). Our main experiments focus on CCPE-M: We show how a simple bias toward the verb *like* easily leads us to overestimate performance in the wild by showing performance drops on semantically innocent perturbations of the test data, as well as on a new sample of movie preference elicitations that we collected from Reddit for the purpose of this paper. We also show that debiasing the data, improves performance. The CCPE-M provides a very clear example of *guideline bias*, but other examples can be found, e.g., in Taskmaster-1, which we discuss in §4.3. We discuss more examples in §4.4.

4.2 BIAS IN CCPE-M

We first examine the CCPE-M dataset of spoken dialogues about movie preferences. The dialogues in CCPE-M are generated in a Wizard-of-Oz set-up, where the assistants type their input, which is then translated into speech using text-to-speech technologies, at which point users respond by speech. The dialogues were transcribed and annotated by the authors of Radlinski et al. (2019).

SENTENCE CLASSIFICATION We frame the CCPE-M movie preference detection problem as a sentence-level classification task. If a sentence contains a labeled span, we let this label percolate to the sentence level and be a label of the entire sentence. If a sentence contains multiple unique label spans the sentence is assigned the leftmost label. A sentence-level label should therefore be interpreted as saying *in this sentence, the user elicits a movie or genre preference*. Our resulting sentence classification dataset contains five different preference labels, including a *NONE* label. We shuffle the data at the dialogue-level and divide the dialogues into training/development/test splits using a 80/10/10 ratio, ensuring sentences from the same dialogue will not end up in both training and test data. As the assistants utterances rarely express any preferences, we only include the user utterances to balance the number of negative labels. See Table 8 for statistics regarding the label distribution.

PERTURBATIONS OF TEST DATA In order to analyse the effects of guideline bias in the CCPE-M dataset, we introduce perturbations of the instances in the test set where *like* occurs, replacing *like* with a synonymous word, e.g. *love*, or paraphrase, e.g. *holds dearly*. We experiment with four different replacements for *like*: (i) *love*, (ii) *was incredibly affected by*, (iii) *have as my all time favorite movie* and (iv) *am out of this world passionate about*. See Figure 9 for an example sentence and its perturbed variants. The perturbations occasionally, but rarely, lead

to grammatically incorrect input.[1] We emphasize that even though we increase the length of the sentence, the phrases we replace *like* with should signal an even stronger statement of preference, which models should be able to pick up on. Since our data consists of informal speech it includes adverbial uses of *like*; we only replace verb occurrences, relying on SpaCy's POS tagger.[2] We replace 219 instances of the verb *like* throughout the test set.

Testing on (↓)/Training on (→)	CCPE-M		CCPE-M$_{thesaurus}$	
	BiLSTM	BERT	BiLSTM	BERT
CCPE-M	74.79	**79.07**	75.16	78.73
CCPE-M$_{love}$	74.39	78.82	75.43	**78.87**
CCPE-M$_{was\ incredibly\ affected\ by}$	70.32	75.03	73.36	**77.42**
CCPE-M$_{have\ as\ my\ all\ time\ favorite\ movie}$	70.75	74.37	67.85	**76.93**
CCPE-M$_{am\ out\ of\ this\ world\ passionate\ about}$	70.70	73.76	72.84	**78.24**
Reddit	44.55	65.86	46.48	**67.45**

Table 7: Comparison of in-sample F_1 performance, performance on the same data with *like* replaced with phrases with similar meaning, and performance on Reddit data. Results are reported for training models on biased CCPE-M as well as a debiased CCPE-M$_{thesaurus}$ which improves model performance in almost all cases.

Original

I [*like*] Terminator 2

Perturbed

I [*love*] Terminator 2

I [*was incredibly affected by*] Terminator 2

I [*have as my all time favorite movie*] Terminator 2

I [*am out of this world passionate about*] Terminator 2

Figure 9: Example of test sentence permutations.

PERTURBATIONS OF TRAIN DATA We also augment the training data to create a less biased resource. Here we adopt a slightly different strategy, also to evaluate a model trained on the debiased training data to the above perturbed test data: We use six paraphrases of the verb *like* listed in a publicly available thesaurus,[3] none of which overlap with the words used to perturb the test data, and randomly

1 Our models are generally robust to such variation, and, as we will see in our experiments below, the perturbations are less harmful than collecting a new sample of evaluation data and evaluating your model on this sample.

2 https://spacy.io/

3 http://thesaurus.com. The paraphrases consists of: (1) *derive pleasure from*, (2) *get a kick out of*, (3) *appreciate*, (4) *take an interest in*, (5) *cherish*, (6) *find appealing*.

Label	train	dev	test	Reddit
NONE	4508	535	545	60
MOVIE_OR_SERIES	2736	346	313	119
MOVIE_GENRE_OR_CATEGORY	1274	169	166	20
PERSON	66	6	9	11
SOMETHING_ELSE	21	0	0	1
total	8605	1056	1033	211

Table 8: CCPE-M and Reddit sentence-level statistics

replace verbal *like* with a probability of 20%. The paraphrases are sampled from a uniform distribution. A total of 401 instances are replaced in the training data using this approach. This is not intended as a solution to guideline bias, but in our experiments below, we show that a model trained on this simple, debiased dataset generalizes better to out of sample data, showing that the bias toward *like* was in fact one of the reasons that our baseline classifier performed poorly in this domain.

REDDIT MOVIE PREFERENCE DATASET In addition to the perturbed CCPE-M dataset, we also collect and annotate a challenge dataset from Reddit threads discussing movies for the purpose of preference elicitation. The comments are scraped from Reddit threads with titles such as '*Here's A Simple Question. What's Your Favorite Movie Genre And Why?*' or '*What's a movie that you love that everyone else hates?*' and mostly consist of top-level comments. These top-level comments typically respond directly the question posed by the thread, and explicitly state preferences. We also include some random samples from discussion trees that contain no preferences, to balance the label distribution slightly. In this data, we observe the word *like*, but less frequently: The verb *like* occurred in 15/211 examples. The data is annotated at the sentence level, as described previously, and we follow the methodology described by Radlinski et al. (2019) and identify anchor items such as names of movies or series, genres or categories and then label each sentence according to the preference statements describing said item, if any. The dataset contains roughly 100 comments, that when divided into individual sentences resulting in 211 datapoints. The statistics can be found in the final column of Table 8. We make the data publicly available.[4]

RESULTS We evaluate the performance on two different models on the original and perturbed CCPE-M, as well as on our Reddit data: (i) a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) sentence

4 https://github.com/vpetren/guideline_bias

classifier, trained only on CCPE-M, including the embeddings, and (ii) a fine-tuned BERT sentence classification model (Devlin et al., 2019). For (i), we use two BiLSTM layers ($d = 128$), randomly initialized embeddings ($d = 64$), and a dropout rate of 0.5. The model is trained for 45 epochs. For (ii), we use the base, uncased BERT model with the default parameters and finetune for 3 epochs. Model selection is conducted based on performance on the development set. Performance is measured using class-weighted F_1 score. We report results in Table 7 on the various perturbation test sets as well as the Reddit data, when (i) the models are trained on the unchanged CCPE-M data, and (ii) the models are trained on the debiased version CCPE-$M_{thesaurus}$.

On the original dataset, BERT performs slightly better than the BiLSTM architecture, but the differences are relatively small. Both BiLSTM and BERT suffer a drop in performance, when examples are perturbed and the word *like* is replaced with synonymous words or phrases. Note how longer substitutions result in a larger drop in performance, e.g. *love* vs. *am out of this world passionate about*. We see the drops follow the same pattern for both architectures, while BiLSTM seems a bit more sensitive to our test permutations. Both models do even worse on our newly collected Reddit data. Here, we clearly see the sensitivity of the BiLSTM architecture, which suffers a 30% absolute drop in F_1; but even BERT suffers a bit performance drop of more than 13%, when evaluated on a new sample of data. When training on CCPE-$M_{thesaurus}$, both models become more invariant to our perturbations,with up to 4.5 F_1 improvements for BERT model and 3 F_1 improvements for the BiLSTM, without any loss of performance on the original test set. We also observe improvements on our collected Reddit data, suggesting that *the initial drop in performance can be partially explained by guideline bias and not only domain differences*.

CONTROLLED PRIMING EXPERIMENT To establish the priming effect of guidelines in a more controlled setting, we set up a small crowdsourced experiment. We asked turkers to respond to a hypothetical question about movie preferences. For example, turkers were asked to imagine they are in a situation in which they 'are asked what movies' they 'like', and that they like a specific movie, say *Harry Potter*. The turker may then respond: *I've always liked Harry Potter*. We collected 40 user responses for each of the priming verbs *like*, *love* and *prefer*, 120 total, and for each of the verbs used to prime the turkers, we compute a probability distribution over most of the verbs in the response vocabulary that are likely to be used to describe a general preference towards something. Figure 10 shows the results of the crowdsourced priming experiments. We can observe that when a specific priming word, such as *like*, is used, there is a significantly higher probability that the response from the user will contain that

same word, illustrating that when keywords in guidelines are heavily over-represented, the collected data will also reflect this bias.

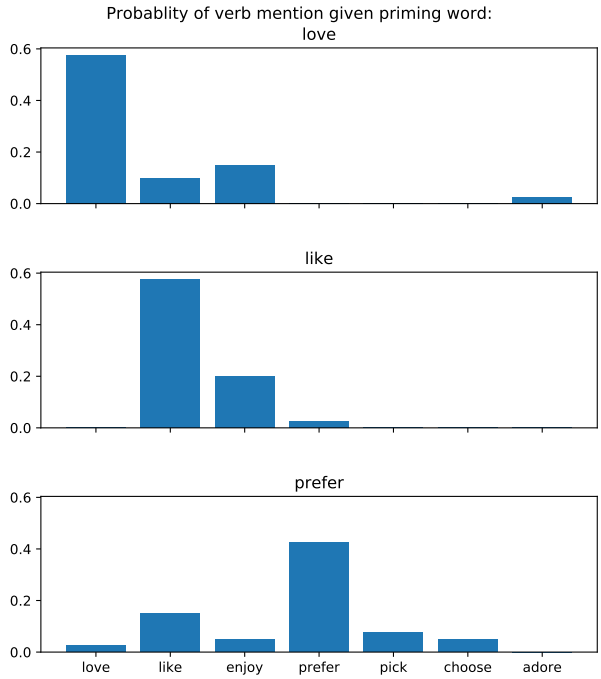


Figure 10: Probability that a verb that describes a preference towards a movie is mentioned, given a priming word by the annotator is mentioned.

4.3 BIAS IN TASKMASTER-1

The order in which the goals of the conversation is described to annotators in the guidelines can also bias the order in which these goals are pursued in conversation. Taskmaster-1 contains conversations between a user and an agent where the user seeks to accomplish a *goal* by, e.g., booking tickets to a movie, which is the domain we focus on. When booking tickets to go see a movie, we can specify the movie title before the theatre, or vice versa, but models may not become robust to such variation if exposed to very biased examples.

Unlike CCPE-M, the Taskmaster-1 dataset was (wisely) collected using two different sets of guidelines to reduce bias, and we can therefore investigate the downstream effects of of the bias induced by the two sets of guidelines. To quantify the guideline bias, we compute the probability that a goal x_1 is mentioned before another one x_2 in an dialogue, given that x_1 precedes x_2 in the guidelines. We only consider dialogues where all goals are mentioned at least once, i.e., ~ 900 in total; the conversations are then divided into two, based on the guideline that was used. Figure 11 shows the heat map of these rel-

ative probabilities. The guidelines have a clear influence on the final structure of the conversation, i.e. if the movie title (x_1) is mentioned before the city (x_2) in the guideline, there is a high probability (0.75) that the same is true in the dialogues. If they are not, the probability is much lower (0.57).

4.4 RELATED WORK

Plank, Hovy, and Søgaard (2014) present an approach to correcting for adjudicator biases. Bender and Friedman (2018) raise the possibility of (demographic) bias in annotation guidelines, but do not provide a means for detecting such biases or show any existing datasets to be biased in this way. Amidei, Piwek, and Willis (2018) also discuss the possibility, but in a footnote. Geva, Goldberg, and Berant (2019) investigates how crowdsourcing practices can introduce annotator biases in NLU datasets and therefore result in models overestimating confidence on samples from annotators that have contributed to both the training and test sets. Liu et al. (2018b), on the other hand, discuss a case in which annotation guidelines are biased by being developed for a particular domain and not easily applicable to another. Cohn and Specia (2013) explores how models can learn from annotator bias in a somewhat opposite scenario from ours, e.g. when annotators deviate from annotation guidelines and inject their own bias into the data, and by using multi-task learning to train annotator specific models, they improve performance by leveraging annotation (dis)agreements. There are, to the best of our knowledge, relatively few examples of researchers identifying concrete guideline-related bias in benchmark datasets: Dickinson and Meurers (2003) suggest that POS annotation in the English Penn Treebank is biased by the vagueness of the annotation guidelines in some respects. Friedrich et al. (2015) report a similar guideline-induced bias in the ACE datasets. Dandapat et al. (2009) discuss an interesting bias in a Bangla/Hindi POS-annotated corpus arising from a decision in the annotation guidelines to include two labels for when annotators were uncertain, but not specifying in detail how these labels were to be used. Goldberg and Elhadad (2010) define structural bias for dependency parsing and how it can be attributed to bias in individual datasets, among other factors, originating from their annotation schemes. Valverde Ibañez and Ohtani (2014) report a similar case, where ambiguity in how special categories were defined, led to bias in a corpus of Spanish learner errors.

In the social sciences, this is a phenomenon which has been studies for decades in relation to survey design (Fisher, 2009; Schuman and Presser, 1977). For example, Smith (1987) observe how survey respondents are affected by seemingly minor word alterations, that still maintain the same intent, can skew the response distribution drastically.

4.5 DISCUSSION & CONCLUSION

In this work, we examined *guideline bias* in two newly presented WoZ style dialogue corpora: We showed how a lexical bias for the word *like* in the annotation guidelines of CCPE-M, through a controlled priming experiment leads to a bias for this word in the dialogues, and that models trained on this corpus are sensitive to the absence of this verb. We provided a new test dataset for this task, collected from Reddit, and show how a debiased model performs better on this dataset, suggesting the 13% drop is in part the result of guideline bias. We showed a similar bias in Taskmaster-1.

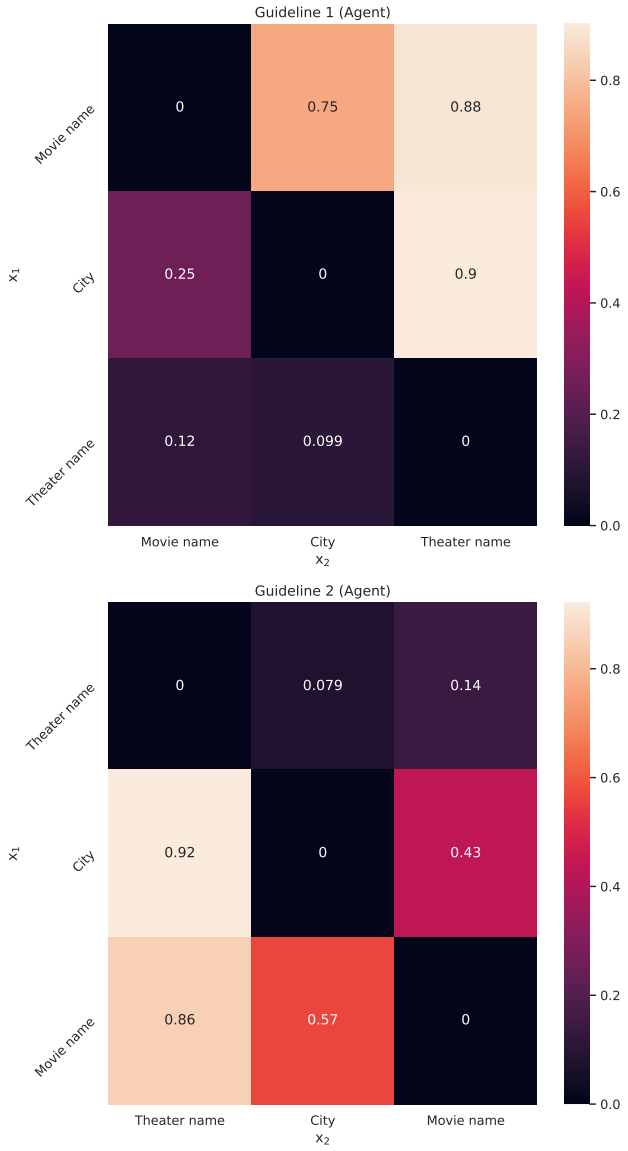


Figure 11: Probability that a guideline goal x_1 is mentioned before another one x_2 in an actual dialogue, given that x_1 comes before x_2 in the agent's guideline.

Part III

EXAMINING FAIRNESS IN NATURAL LANGUAGE PROCESSING

5

IS THE LOTTERY FAIR? EVALUATING WINNING TICKETS ACROSS DEMOGRAPHICS

ABSTRACT

Recent studies have suggested that weight pruning, e.g. using lottery ticket extraction techniques (Frankle and Carbin, 2019), comes at the risk of compromising the group fairness of machine learning models (Hooker et al., 2019, 2020; Paganini, 2020), but to the best of our knowledge, no one has empirically evaluated this hypothesis at scale in the context of natural language processing. We present experiments with two text classification datasets annotated with demographic information: the Trustpilot Corpus (sentiment) and Civil-Comments (toxicity). We evaluate the fairness of lottery ticket extraction through layer-wise and global weight pruning across three languages and two tasks. Our results suggest that there is a small increase in group disparity, which is most pronounced at high pruning rates and correlates with instability. The fairness of models trained with distributionally robust optimization objectives is sometimes less sensitive to pruning, but results are not consistent. The code for our experiments is available at `https://github.com/vpetren/fairness_lottery`.

5.1 INTRODUCTION

Heavily pruning deep neural network models is a way of reducing inference cost for resource-constrained environments, but does weight-pruning of deep neural networks increase their unfairness? Several recent papers suggest this (Hooker et al., 2019; Paganini, 2020), based on experiments from face and digit recognition, but does this also hold for natural language processing (NLP) models? Systematic biases may easily be exacerbated by pruning interventions in high-dimensional problems because of feature swamping effects (Sutton, Sindelar, and McCallum, 2006). Overparameterized deep neural networks generalize well, in part because they can hedge their bets and rely on multitudes of weak evidence rather than the most prominent independent variables. Sparse models do not have that luxury and are therefore more sensitive to shifts (Globerson and Roweis, 2006; Søgaard, 2013).

We introduce a *fairness sensitivity to pruning* metric that measures how Rawlsian min-max fairness across demographic groups changes with weight pruning. We estimate this sensitivity by taking the gradi-

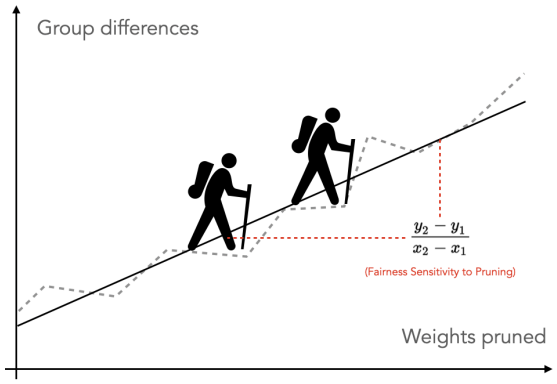


Figure 12: Fairness Sensitivity to Pruning (FSP): the gradient of the linear fit of (the logarithm of) the pruning ratio to min-max group-level disparity. We use this to quantify the sensitivity of Rawlsian min-max fairness to weight pruning across architectures, pruning strategies and datasets.

ent of the linear fit of the logarithm of the pruning ratio to min-max group-level disparity. We show that across four datasets, fairness sensitivity to pruning is similar for layer-wise and global pruning strategies (Frankle and Carbin, 2019), as well as for text classifiers based on feed-forward and recurrent neural networks. Subsequently, we consider the impact of a popular robust optimization strategy designed to improve the fairness of classification models (Hashimoto et al., 2018; Sagawa et al., 2020b), on the fairness sensitivity of feed-forward networks.

CONTRIBUTIONS We are, to the best of our knowledge, the first to study the impact of weight pruning on fairness in NLP at scale. We introduce a *fairness sensitivity to pruning* (FSP) metric that measures how Rawlsian min-max fairness across demographic groups decreases with weight pruning. We evaluate FSP across two architectures, two pruning strategies and two datasets, including multilingual sentiment classification and English toxicity classification. Our results suggest that pruning increases group-level performance disparities, but mostly at high pruning rates and with some variance across architectures and pruning strategies. Group-level disparities seem to be in part a result of the instability of weight pruning. We compare FSP between our baseline empirical risk models and robust models induced with Distributional Robust Optimization (DRO) (Hashimoto et al., 2018; Sagawa et al., 2020b). Our results show that weight pruning in combination with DRO can *sometimes* (8/16 cases here) be used to induce fairer, sparse classifiers, but the effect is not significant ($p \sim 0.18$) across our experiments.

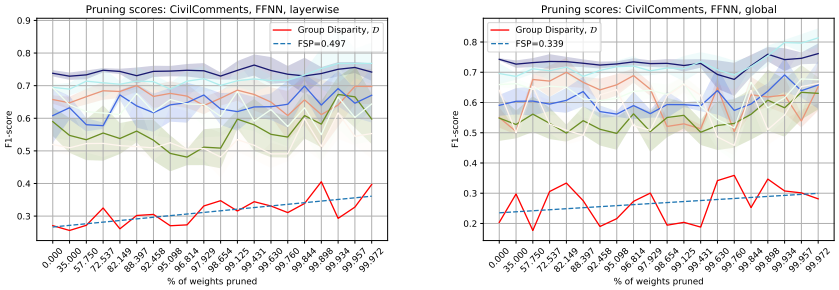


Figure 13: Macro-averaged performance of our feed-forward networks as a function of pruning ratio. Fairness Sensitivity to Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs. Results are for CivilComments. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. See the Appendix for similar plots for the Trustpilot Corpus.

5.2 RELATED WORK

PRUNING NEURAL NETWORKS The literature on pruning neural networks is decades old (Cun, Denker, and Solla, 1990; Hassibi and Stork, 1993; Mozer and Smolensky, 1989), but has recently seen a resurgence with the all-encompassing success of neural networks and the need for small and fast on-device model inference (Frankle and Carbin, 2019; Frankle et al., 2020; Han et al., 2015; Sze et al., 2017). In NLP, specifically, pruning methods have been applied to recurrent neural networks (Desai, Zhan, and Aly, 2019; Yu et al., 2020), as well as transformers (Brix, Bahar, and Ney, 2020; Chen et al., 2020; Gordon, Duh, and Andrews, 2020; Prasanna, Rogers, and Rumshisky, 2020; Sanh, Wolf, and Rush, 2020).

FAIRNESS IN PRUNED MODELS Measuring fairness in pruned models is an unexplored area. However, Paganini (2020) evaluates the fairness, i.e., the difference between the best- and worst-case groups, of lottery ticket-style weight pruning for digit recognition problems: Specifically, they retrain models for a fixed number of iterations using global unstructured pruning. In addition, they present a meta-regression study suggesting that underrepresented and more complex classes are most severely affected by pruning procedures. See Hooker et al. (2019) for related work and similar results in face recognition.[1]

1 Bartoldson et al. (2020) arguably present results from object recognition that show the opposite trend: Generalization increases with (layer-wise) pruning. This seems to be a side effect of overparameterization; interestingly, we see the opposite trend for feed-forward networks and layer-wise pruning.

IMPROVING FAIRNESS Fairness of overparameterized models can be improved by distributionally robust optimization (DRO) (Hashimoto et al., 2018; Levy et al., 2020), or to some extent by simpler post-hoc correction methods such as classifier retraining or group-specific classification thresholds (Menon, Rawat, and Kumar, 2021). DRO minimizes the worst-case expected loss over an uncertainty set of distributions. The uncertainty set represents the distributions we want our model to perform well on. In Sagawa et al. (2020a), the uncertainty set is all possible mixtures of a known set of groups, a variant referred to as Group DRO. Sagawa et al. (2020b) find that subsampling the majority groups can be a way for overparameterized models to achieve both low minority test error as well as low average test error.

5.3 PRUNING METHODOLOGY

We extract winning lottery tickets from our network according to the iterative procedure outlined in Frankle and Carbin (2019): Given a model $f(x;\theta)$ with initial network parameters θ_0 and mask m_0, for each pruning iteration i, we start by initializing a model $f(x;\theta)$ with initial parameter θ_0 and train it for N epochs, resulting in $f(x;\theta_N)$. After training, we prune a fixed fraction $p \in [0,1]$ from the remaining parameters in θ_N to obtain the mask m_i. The pruned weights are chosen using the L_1 norm, meaning the neurons with the lowest magnitude are masked out. Pruning can either be done w.r.t. individual layers or all of them combined, also referred to as *layer-wise* and *global* pruning. m_i is then carried over to the subsequent pruning iteration $i+1$ with the model $f(x, m_i \odot \theta_0)$ and retrained once again. At iteration i, the fraction of weights pruned is therefore $1-(1-p)^i$.

5.4 EXPERIMENTS

5.4.1 *Data*

DATASETS We examine fairness among heavily pruned models using two text classification datasets: i) The multilingual **Trustpilot** Corpus (Hovy, Johannsen, and Søgaard, 2015),[2] which contains user reviews from the Trustpilot website of various companies and services in five different countries (Germany, Denmark, France, United Kingdom and United States). The reviews are based on a one to five star rating scale and some are accompanied by demographic attributes about the author, such as gender, age and location. 2) The **CivilComments** dataset (Borkan et al., 2019),[3] which contains comments annotated for toxicity, for the purpose of hate speech detection. A subset

2 https://bitbucket.org/lowlands/release/src/master/WWW2015/data/

3 https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

of the comments are also annotated for the protected attributes they address, including gender, race, and religion.

PREPROCESSING For the Trustpilot Corpus, we divide the data into demographics based on a combination of *gender* (male/female), *age* (young/old) and *location* (NUTS regions). For age, young is defined as being 35 or less. We exclude the French and American parts of the datasets as they do not have properly annotated NUTS regions. For UK and Germany, we use NUTS-1 regions, and for Denmark, where more data is available, we use NUTS-2 regions. We convert the 5-star ratings to binary sentiment labels, grouping 4 and 5 stars as positive, and 1 and 2 as negative. Neutral reviews (three stars) are discarded.[4] Likewise for CivilComments, we threshold comments with a toxicity rating > 0.5 as toxic, and otherwise label them as a non-toxic. This is similar to the binarization performed in Koh et al. (2021). Comments can for each demographic sub-attribute contain multiple partial values (e.g. $\texttt{asian} = 0.3$, $\texttt{black} = 0.4$ for the *race* attribute), so for each annotated attribute we assign it the sub-attribute with the largest value. In our experiments we consider demographics based on combinations of the *race* and *gender* attributes. For each language and dataset we randomly sample 100, 200 or 500 of each demographic as test sets, based on the the amount of annotated datapoints in the dataset, and use a 80-20 split of the remaining data for training and validation. If a demographic contains less than the specified number of datapoints, we disregard it. Due to high class imbalance, the majority class for our train-val data is downsampled to match the minority class. Table 9 shows the statistics for the respective datasets we train and evaluate on.

Dataset	Train	Val	N	S
Trustpilot-DK	222229	55557	20	500
Trustpilot-DE	26146	6536	42	100
Trustpilot-UK	127965	31991	50	200
CivilComments	357602	89400	7	100

Table 9: Detailed dataset statistics. N refers to the number of discrete demographics in the dataset and S is the size of each demographic test set.

4 This binarization scheme is standard; see, e.g., Gupta, Thadani, and O'Hare (2020) and Desai, Zhan, and Aly (2019)

FFNN				
Dataset	E_{dim}	h_{dim}	B	N
Trustpilot-DK	128	256	15	32
Trustpilot-DE	128	256	15	8
Trustpilot-UK	128	256	15	16
CivilComments	128	256	15	32
LSTM				
Dataset	E_{dim}	h_{dim}	B	N
Trustpilot-DK	128	256	10	64
Trustpilot-DE	128	256	15	16
Trustpilot-UK	128	256	10	32
CivilComments	128	256	10	64

Table 10: FFNN and LSTM hyperparameters. E_{dim} is embedding layer size, h_{dim} is hidden layer size, B is batch size and N is number of epochs. Both the layer-wise and global pruning structures use the same set of hyperparameters.

5.4.2 *Models*

We consider simple **FFNN** (Rumelhart, Hinton, and Williams, 1986) and **LSTM** (Hochreiter and Schmidhuber, 1997) neural networks for text classification.

FFNN The FFNN consists of the following: The embedding layer, which maps every token id in the text to a fixed size vector as a bag-of-embeddings and sums them together, resulting in a single representation $e \in \mathbb{R}^{|E_{dim}|}$, followed by 3 fully connected layers of size $\mathbb{R}^{|E_{dim} \times h|}$, $\mathbb{R}^{|h \times h|}$ and $\mathbb{R}^{|h \times 2|}$ respectively. We use the hyperbolic tangent activation between layers and each linear layer is initialized using He initialization (He et al., 2015a).

LSTM The LSTM network is a 2-layer bidirectional LSTM (Hochreiter and Schmidhuber, 1997) which encodes our input text, followed by a fully connected layer for classification. The weights are initialized using $\mathcal{U}(-\sqrt{k}, \sqrt{k})$ where $k = \frac{1}{\text{hidden\_size}}$ and the final fully connected layer uses He initialization. See all model hyperparameters used in Table 10.

Both the FFNN and LSTM models are trained using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e-3$ and a weight decay of $1e-4$.

DISTRIBUTIONALLY ROBUST OPTIMIZATION LOSS Additionally, we also train our models with DRO loss (Levy et al., 2020). We use the implementation provided by Levy et al. (2020) [5]. For our experiments, a χ^2 uncertainty set of size 1 is used.

For all of our experiments, we extract our winning tickets over 20 pruning iterations and use a pruning rate of $p = 0.35$. We run a total of 5 independent runs for each model-dataset combination.

5.4.3 *Measuring group disparity*

At each pruning step we measure the group disparity $\mathcal{D}$, from a set of demographics D, between repeated runs R, by computing the maximum difference of F_1 scores as follows:[6]

$$\mathcal{D} = \max_{d_m \in D} \max_{d_{n \neq m} \in D} \max_{r_i \in R} \max_{r_{j \neq i} \in R} |F_{1\, r_i\, d_m} - F_{1\, r_j\, d_n}| \tag{4}$$

Intuitively, this corresponds to the difference between the highest scoring run for the highest scoring demographic and the lowest counterpart across all repeated runs. We compute FSP by taking the gradient of the linear fit of $\mathcal{D}$ over a P pruning steps multiplied by 100.

		Trustpilot			CC	Avg
		da	de	en	en	
FFNN	lw	−0.183	0.281	−0.230	0.497	0.091
	gl	0.227	1.375	1.054	0.339	0.749
FFNN-DRO	lw	−0.044	0.321	0.143	0.089	0.127
	gl	0.351	0.875	−0.040	0.368	0.388
LSTM	lw	0.221	0.411	0.206	0.823	0.415
	gl	1.099	0.198	0.352	0.252	0.475
LSTM-DRO	lw	0.263	−0.282	−0.082	1.335	0.309
	gl	0.262	−0.609	0.544	0.006	0.051

Table 11: FSP values across architectures, layer-wise (lw) and global (gl) pruning, and the four datasets. Our main observation is that FSP values are almost consistently positive, and slightly higher for global pruning. DRO does not consistently reduce FSP; we highlight cases where it does.

5.5 RESULTS

MAIN EXPERIMENTS Our first set of results evaluate FSP across architectures, datasets, and pruning techniques. In 14/16 combinations of FFNN and LSTM neural networks, the Trustpilot Corpus and

5 https://github.com/daniellevy/fast-dro/

6 Maximum discrepancy has also been used as a measure of fairness in Alabi, Immorlica, and Kalai (2018) and Calmon et al. (2017). See Williamson and Menon (2019) for discussion.

CivilComments, layer-wise and global pruning, we see positive FSP values. In other words, weight pruning leads to higher group-level performance disparities, i.e., less fairness. Comparing layer-wise and global pruning, we note that group disparity is generally higher for global pruning. In Figure 13, we present two plots - for layer-wise and global pruning of a feed-forward network trained on CivilComments. The remaining plots are presented in Appendix A.3. The FSP values are listed in Table 11. FFNNs exhibit very high FSP values with global pruning, but while global pruning increases unfairness, layer-wise pruning does not. For LSTMs, the effects of the two pruning strategies are similar: Both lead to moderate increases in group disparities.[7] In a couple of instances we witnessed model degeneration due to heavy pruning resulting in single-class prediction before 20 pruning iterations. The plots and FSP values exclude these datapoints as they are not relevant for our analysis.

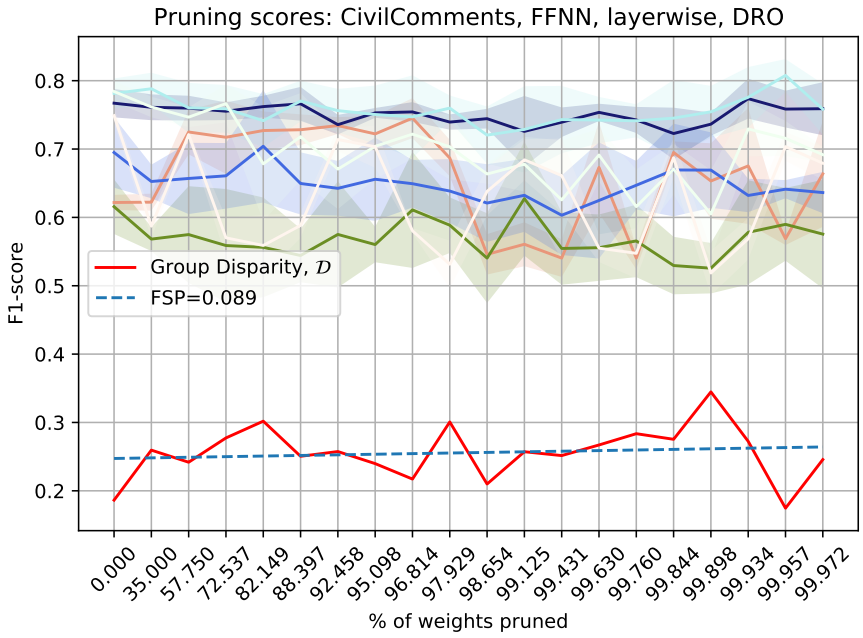


Figure 14: FSP for Distributional Robust Optimization

DISTRIBUTIONALLY ROBUST OPTIMIZATION We ran comparable experiments using DRO loss (Hashimoto et al., 2018) to see whether the adverse effects of weight pruning on min-max fairness could be reduced by training with a more robust objective. This seems to hold true in some instances. We present a single plot for DRO in Figure 14, for feed-forward networks, layer-wise pruning on CivilComments;

7 While fairness correlates with stability, the difference between FFNNs and LSTMs is not explained by stability differences (see plots in the Appendix), but should probably be attributed to the general performance differences between FFNNs and LSTMs, as well as relative overparameterization in FFNNs (see Footnote 1).

see the Appendix for more plots. Comparing with Figure 13 (left) the FSP metric is considerably lower than for baseline empirical risk minimization (0.089 vs. 0.497) while maintaining equal, or even better, performance at high pruning rates; but note from the red numbers in Table 11, that we only see this type of reduction in FSP in 3/8 cases for FFNNs, but DRO does reduce the average FSP for global pruning. In 5/8 cases for the LSTM, however, DRO does improves fairness, reducing the average FSP with both layer-wise and global pruning.

5.6 CONCLUSION

In this work, we take a first step in examining group disparity among heavily pruned models, using lottery ticket extraction, in NLP. We measure group disparity, using *fairness sensitivity to pruning*, on the Trustpilot Corpus, a sentiment classification dataset covering 3 languages, as well as CivilComments, a toxicity classification dataset, for both feed-forward and recurrent neural networks. We find that models subject to heavy pruning are more susceptible to higher levels of group disparity, but that this effect can to some degree be mitigated using distributionally robust optimization objectives.

6

THE IMPACT OF DIFFERENTIAL PRIVACY ON GROUP DISPARITY MITIGATION

ABSTRACT

The performance cost of differential privacy has, for some applications, been shown to be higher for minority groups; fairness, conversely, has been shown to disproportionally compromise the privacy of members of such groups. Most work in this area has been restricted to computer vision and risk assessment. In this paper, we evaluate the impact of differential privacy on fairness across four tasks, focusing on how attempts to mitigate privacy violations and between-group performance differences interact: Does privacy inhibit attempts to ensure fairness? To this end, we train (ε, δ)-differentially private models with empirical risk minimization and group distributionally robust training objectives. Consistent with previous findings, we find that differential privacy increases between-group performance differences in the baseline setting; but more interestingly, differential privacy *reduces* between-group performance differences in the robust setting. We explain this by reinterpreting differential privacy as regularization.

6.1 INTRODUCTION

Classification tasks in computer vision and natural language processing face the challenge of balancing performance with the need to prevent discrimination against protected demographic subgroups, satisfying fairness principles. In some tasks, we train our classifiers on private data and therefore also need our models to satisfy privacy guarantees.

Privacy-preserving algorithms, however, tend to disproportionally affect members of minority classes (Farrand et al., 2020). Bagdasaryan, Poursaeed, and Shmatikov (2019), for example, show the performance cost of differential privacy (Dwork et al., 2006) in face recognition is higher for minority groups, suggesting that privacy and fairness are fundamentally at odds (Agarwal, 2021; Chang and Shokri, 2021).

In this paper, we evaluate two hypotheses at scale: (a) that the performance cost of differential privacy is unevenly distributed across demographic groups (Bagdasaryan, Poursaeed, and Shmatikov, 2019; Cummings et al., 2019; Ekstrand, Joshaghani, and Mehrpouyan, 2018; Farrand et al., 2020), and (b) that such effects can in part be mitigated

by more robust learning objectives (Pezeshki et al., 2020; Sagawa et al., 2020a).

CONTRIBUTIONS We build upon previous work suggesting that differential privacy and fairness are at odds: Differential privacy hurts minority groups the most, and reducing the fairness gap by focusing on minority groups during training typically puts their privacy at risk. We evaluate this hypothesis at scale by measuring the impact of differential privacy in terms of fairness across (1) a baseline empirical risk minimization and (2) under a group distributionally robust optimization. We conduct our experiments across four tasks of different modalities, assuming the group membership information is available at training time, but not at test time: face recognition (CelebA), topic classification, volatility forecasting based on earning calls, and sentiment analysis of product reviews. Our results confirm that differential privacy compromises fairness in the baseline setting; however, we demonstrate that differential privacy not only mitigates the decrease but also *improves* fairness compared to non-private experiments for $4/5$ tasks in the distributionally robust setting. We explain this by reinterpreting differential privacy as an approximation of Gaussian noise injection, which is equivalent to strategies previously shown to determine the efficacy of group-robust learning.

6.2 FAIRNESS AND PRIVACY

Fair machine learning aims to ensure that induced models do not discriminate against individuals with specific values in their protected attributes (e.g., race, gender). We represent each data point as $z = (x, g, y) \in \mathcal{X} \times \mathcal{G} \times \mathcal{Y}$, with $g \in \mathcal{G}$ encoding its protected attribute(s).[1] Let $\mathcal{D}_y^g$ denote the distribution of data with protected attribute g and label y.

Several definitions of group fairness exist in the literature (Williamson and Menon, 2019), but here we focus on a generalization of approximately constant conditional (equalized) risk (Donini et al., 2018):[2]

Definition 1 (Δ-Fairness)**.** Let $\ell^{g_i}(\theta) = \mathbb{E}[\ell(\theta(x), y)|g = g_i]$ be the risk of the samples in the group defined by g_i, and $\Delta \in [0, 1]$. We say that a model θ is Δ-fair if for any two values of g, say g_i and g_j, $|\ell^{g_i}(\theta) - \ell^{g_j}(\theta)| < \Delta$.

Note that if ℓ coincides with the performance metric of a task, and $\delta = 0$, this is identical to performance or classification parity (Yuan

1 In practice our protected attributes in § 6.3 will be *age* and *gender*. Both are protected under the Equality Act 2010.

2 In the fairness literature, approximate fairness is referred to as δ-fairness, but below we will use lower case δ to refer to (ε, δ)-differential privacy, and we refer to Δ-fairness to avoid confusion.

et al., 2021).[3] Such a notion of fairness can be derived from John Rawls' theory on distributive justice and stability, treating model performance as a resource to be allocated. Rawls' *difference principle*, maximizing the welfare of the worst-off group, is argued to lead to stability and mobility in society at large (Rawls, 1971). Δ directly measures what is sometimes called Rawlsian *min-max fairness* (Bertsimas, Farias, and Trichakis, 2011). In our experiments, we measure Δ-fairness as the absolute difference between performance of the worst-off and best-off subgroups.

Recall the standard definition of (ε, δ)-privacy is as follows:

Definition 2. θ is (ε, δ)-private iff $\Pr[\theta(\mathcal{X})] \leqslant \exp(\varepsilon) \times \Pr[\theta(\mathcal{X}')] + \delta$ for any two distributions, $\mathcal{X}$ and $\mathcal{X}'$, different at most in one row.

Differential privacy thereby ensures that an algorithm will generate similar outputs on similar data sets. Note the multiplicative bound $\exp(\varepsilon)$ and the additive bound δ serve different roles: The δ term represents the possibility that a few data points are not governed by the multiplicative bound, which controls the level of privacy (rather than its scope). Note that it also follows directly that if $\varepsilon = 0$ and $\delta = 0$, absolute privacy is required, leading θ to be independent of the data.

Several authors have shown that differential privacy comes at different costs for minority subgroups (Bagdasaryan, Poursaeed, and Shmatikov, 2019; Cummings et al., 2019; Ekstrand, Joshaghani, and Mehrpouyan, 2018; Farrand et al., 2020). The more private the model is required to be, the larger group disparities it will exhibit.[4] This happens because differential privacy distributes noise where it is needed to reduce the influence of individual examples. Since outlier examples are likely to have disproportional influence on output distributions (Campbell, 1978; Chernick and Murthy, 1983), they are also disproportionally affected by noise injection in differential privacy.

Agarwal (2021) show that, in fact, a $(\varepsilon, 0)$-private and fully fair model – using equalized odds as the definition of fairness – will be unable to learn anything. To see this, remember that a fully private model is independent of the data and unable to learn from correlations between input and output. If θ is, in addition, required to be fair, it is thereby required to be fair for all distributions, which prevents θ from encoding any prior beliefs about the output distribution. Note this finding generalizes straight-forwardly to equalized risk, and even

3 Performance or classification parity has been argued to suffer from statistical limitations in (Corbett-Davies and Goel, 2018), which remind us that when risk distributions differ, standard error metrics are poor proxies of individual equity. This is known as the problem of infra-marginality. Note, however, that this argument does not apply to binary classification problems.

4 Note this is a different trade-off than the fairness-privacy trade-off which results from the need for collecting sensitive data to learn fair models; the latter is discussed at length in Veale and Binns (2017).

Figure 15: Examples of the different subgroups that appear in a subset of the datasets we train on. CelebA (left) contains images of celebrities, using hair-color as our target variable and gender as our protected attribute. Blog Authorship Corpus (right) contains text-based blogposts on two topics {Technology, Arts} our targets, using $\mathcal{G} : \{\text{Man}, \text{Woman}\} \times \{\text{Young}, \text{Old}\}$ as our protected subgroups.

to approximate fairness (since even for finite distributions, we can define a $\Delta > 0$, such that preserving absolute privacy would lead to a constant θ).

Theorem 1. *For sufficiently small values of Δ, a fully $(\varepsilon, 0)$-private model θ that is also Δ-fair, will have trivial performance.*

Proof. This follows directly from the above. □

While we do not strictly require an absolute privacy in our experiments (setting $\delta = 10^{-5}$), intuitively, privacy compromises fairness by adding more noise to data points of minority group members than to those of majority groups. Fairness, on the other hand, leads to oversampling or over-attending to data points of minority group members, more likely compromising their privacy.

Pannekoek and Spigler (2021) show, however, that it *is* possible to learn *somewhat private* and *somewhat fair* classifiers. They combine differential privacy with reject option classification. Their results nevertheless confirm that privacy and fairness objectives are fundamentally at odds, as fairness decreases with the introduction of differential privacy.

6.3 EXPERIMENTS

This section describes the algorithms and datasets involved in our experiments, and presents the results of these.

6.3.1 *Algorithms*

EMPIRICAL RISK MINIMIZATION For a model parameterized by θ, in our baseline Empirical Risk Minimization (ERM) setting, we

minimize the expected loss $\mathbb{E}[\ell(\theta(x), y)]$ with data $(x, g, y) \in \mathcal{X} \times \mathcal{G} \times \mathcal{Y}$ drawn from a dataset $\mathcal{D}$:

$$\hat{\theta}_{\text{ERM}} = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\hat{\mathcal{D}}}[\ell(\theta(x), y)] \tag{5}$$

Here $\hat{\mathcal{D}}$ denotes the empirical training distribution. Note that we disregard any group information in our data. In an overparameterized setting, ERM is prone to overfitting spurious correlations, which are more likely to hurt performance on minority groups (Sagawa et al., 2020b).

DISTRIBUTIONALLY ROBUST OPTIMIZATION Several authors have suggested to mitigate the effects of such overfitting by explicitly optimizing for out-of-distribution mixtures of sub-populations (Hu et al., 2018; Oren et al., 2019; Sagawa et al., 2020a). In this work we focus on Group-aware Distributionally Robust Optimization (Group DRO) (Sagawa et al., 2020a).

Under the assumption that the training distribution $\mathcal{D}$ is a mixture of a discrete number of groups, $\mathcal{D}_g$ for $g \in \mathcal{G}$, we define the worst-case loss as the maximum of the group-specific expected losses:

$$\ell(\theta)_{worst} = \max_{g \in \mathcal{G}} \mathbb{E}_{\hat{\mathcal{D}}_g}[\ell(\theta(x), y)] \tag{6}$$

In Group DRO – in contrast with ERM – we exploit our knowledge of the group membership of data points (x, g, y). The overall objective is for minimizing the empirical worst-case loss is therefore:

$$\hat{\theta}_{\text{DRO}} = \underset{\theta}{\operatorname{argmin}} \left[\ell(\hat{\theta})_{worst} := \max_{g \in G} \mathbb{E}_{\hat{\mathcal{D}}_g}[\ell(\theta(x), y)] \right] \tag{7}$$

Note, again, that the knowledge of group membership g is only available at training time, not at test time. Unlike Sagawa et al. (2020a), we do not employ heavy ℓ_2 regularization during our experiments, but rather use it with the same parameters as Koh et al. (2021).

DIFFERENTIALLY PRIVATE STOCHASTIC GRADIENT DESCENT (DP-SGD) We implement differential privacy (Dwork et al., 2006) using DP-SGD, as presented in Abadi et al. (2016). DP-SGD limits the influence of training samples by (i) clipping the per-batch gradient where its norm exceeds a pre-determined clipping bound C, and by (ii) adding Gaussian noise $\mathcal{N}$ characterized by a noise scale σ to the aggregated per-sample gradients. We control this influence with a privacy budget ε, where lower values for ε indicates a more strict level of privacy. DP-SGD has remained popular, among other things because it generalizes to iterative training procedures (McMahan et al., 2018), and supports tighter bounds using the Rényi method (Mironov, 2017).

Differential privacy generally comes at a performance cost, leading to privacy-preserving models performing worse compared to their

non-private counterparts (Alvim et al., 2011). However, we follow Kerrigan, Slack, and Tuyls (2020) and *finetune* the private models, which are first pre-trained (without differential privacy) on a large public dataset. This protocol generally seems to provide a better trade-off between accuracy and privacy (Kerrigan, Slack, and Tuyls, 2020), leading to better-performing, yet private models. The only exception to this setup is the volatility forecasting task, where our models were trained from scratch, as those rely on PRAAT audio features.

6.3.2 *Tasks and architectures*

To study the impact of differential privacy on fairness, in ERM and Group DRO, we evaluate increasing levels of differential privacy across five datasets that span four tasks and three different modalities: speech, text and vision.

FACIAL ATTRIBUTE DETECTION We study facial attribute recognition with the CelebFaces Attributes Dataset (CelebA) (Liu et al., 2015) [5]. It contains faces of celebrities annotated with attributes, such as hair color, gender and other facial features. Following Sagawa et al. (2020a), we use the hair color as our target variable, with gender being the demographic attribute (see Figure 15 (left)). The dataset contains ~ 163K datapoints, where the smallest group (blond males) only counts 1387. We finetune a publicly pre-trained ResNet50, a standard model for image classification tasks,[6] on the CelebA dataset and evaluate model performances as accuracies over 3 individual seeds.

TOPIC CLASSIFICATION For topic classification, we use the Blog Authorship Corpus (Schler et al., 2006).[7][8] The Blog Authorship Corpus contains weblogs written on 19 different topics, collected from the Internet before August 2004. The dataset contains self-reported demographic information about the gender and age of the authors. Gender information is binary, and we binarize age, distinguishing between young (=< 35) and older (> 35) authors[9], resulting in four different group combinations (see Figure 15 (right)). We chose two topics of roughly equal size (Technology and Arts), reducing the topic classification task to a binary classification task. For our experiments, we finetune a pre-trained English DistilBERT model (Sanh et al., 2019).[10] To reduce the overall added computational cost of DP-SGD, we freeze

5 The CelebA dataset is available for non-commercial research purposes only.
6 ResNet50 is a variant of the ResNet model (He et al., 2015b), which has 48 convolution layers along with 1 max pooling and 1 average pooling layer. It has 3.8 x 10^9 floating points operations.
7 https://www.kaggle.com/rtatman/blog-authorship-corpus
8 The Blog Authorship Corpus is available for non-commercial research purposes only.
9 Older authors tend to be underrepresented in web data (Nguyen et al., 2014)
10 DistilBERT is a small Transformer model trained by distilling BERT (Devlin et al., 2019) (bert-base-uncased). It has 3/5th of the parameters of bert-base-uncased, runs

		Performance at ε-Privacy							
		No DP		ε_1		ε_2		ε_3	
		Score	ε	Score	ε	Score	ε	Score	ε
Celeb	ERM	0.954 ± 0.000	-	0.943 ± 0.001	9.50	0.940 ± 0.002	5.17	0.932 ± 0.001	0.99
	DRO	0.953 ± 0.001	-	0.899 ± 0.006	9.50	0.891 ± 0.014	5.17	0.873 ± 0.007	0.99
Blog	ERM	0.699 ± 0.002	-	0.661 ± 0.003	9.25	0.661 ± 0.003	5.03	0.648 ± 0.005	1.02
	DRO	0.692 ± 0.001	-	0.651 ± 0.001	9.25	0.650 ± 0.005	5.03	0.630 ± 0.003	1.02
Vol.	ERM	0.756 ± 0.036	-	0.778 ± 0.073	9.32	0.794 ± 0.046	6.42	0.778 ± 0.039	0.96
	DRO	0.814 ± 0.061	-	0.798 ± 0.042	9.32	0.815 ± 0.056	6.42	0.833 ± 0.093	0.96
T-UK	ERM	0.933 ± 0.008	-	0.919 ± 0.002	9.39	0.916 ± 0.001	4.94	0.889 ± 0.009	1.02
	DRO	0.931 ± 0.004	-	0.893 ± 0.006	9.39	0.873 ± 0.015	4.94	0.820 ± 0.015	1.02
T-US	ERM	0.894 ± 0.007	-	0.817 ± 0.014	10.71	0.812 ± 0.009	5.10	0.666 ± 0.019	1.01
	DRO	0.899 ± 0.009	-	0.569 ± 0.132	10.71	0.437 ± 0.112	5.10	0.342 ± 0.012	1.01

		Group-disparity at ε-Privacy							
		No DP		ε_1		ε_2		ε_3	
		GD	ε	GD	ε	GD	ε	GD	ε
Celeb	ERM	0.556 ± 0.021	-	0.746 ± 0.032	9.50	0.734 ± 0.025	5.17	0.770 ± 0.013	0.99
	DRO	0.514 ± 0.042	-	0.039 ± 0.018	9.50	0.080 ± 0.031	5.17	**0.056** ± 0.027	0.99
Blog	ERM	0.108 ± 0.013	-	0.149 ± 0.006	9.25	0.140 ± 0.004	5.17	0.136 ± 0.011	0.99
	DRO	0.078 ± 0.009	-	0.056 ± 0.020	9.25	0.070 ± 0.013	5.17	**0.077** ± 0.027	0.99
Vol.	ERM	0.302 ± 0.042	-	0.328 ± 0.067	9.32	0.557 ± 0.050	6.42	0.573 ± 0.050	0.96
	DRO	0.221 ± 0.062	-	0.320 ± 0.085	9.32	0.371 ± 0.058	6.42	0.421 ± 0.083	0.96
T-UK.	ERM	0.018 ± 0.005	-	0.022 ± 0.006	9.39	0.020 ± 0.014	4.94	0.037 ± 0.006	1.02
	DRO	0.030 ± 0.008	-	0.030 ± 0.004	9.39	0.039 ± 0.023	4.94	**0.025** ± 0.010	1.02
T-US	ERM	0.055 ± 0.006	-	0.048 ± 0.019	10.71	0.054 ± 0.015	5.10	0.109 ± 0.017	1.01
	DRO	0.036 ± 0.007	-	0.118 ± 0.040	10.71	0.078 ± 0.030	5.10	**0.021** ± 0.030	1.01

Table 12: Performance (top) and Δ-Fairness (bottom) of ERM and Group DRO across different degrees of differential privacy (ε). ε_1, ε_2 and ε_3 corresponds to ε-values of roughly 10, 5 and 1 respectively (see table for exact values). We report F1 scores for sentiment and topic classification, accuracy for face recognition and MSE for volatility forecasting. Group disparity (GD) is measured by the absolute difference between the best and worst performing sub-group (Δ-Fairness; see Definition 2.1). The performance and corresponding uncertainties are based on several individual runs of each configuration, see § A.4 in the Appendix for further details. Differential privacy consistently hurts fairness for ERM. For Group DRO, we **bold-face** numbers where strict differential privacy (ε_3) *increases* fairness; this happens in 4/5 datasets. We see large increases for face recognition and small increases for topic classification and sentiment analysis.

our model, except for the outer-most Transformer (Vaswani et al., 2017) encoder layer as well as the classification layer. We report model performances as F1 scores over 3 individual seeds.

VOLATILITY FORECASTING For the stock volatility forecasting task, we use the Earnings Conference Calls dataset by Qin and Yang (2019). This consists of 559 public earnings calls audio recordings for 277 companies in the S&P 500 index, spanning over a year of earnings calls. We obtain the self-reported gender of the CEOs from Reuters,[11] Crunchbase,[12] and the WikiData API.[13] Gender information is binary, with 12.3% of speakers being female and 87.7% of speakers being male, a highly skewed distribution. Since our primary focus with this task is to explore the impact of differential privacy on speech, we use only audio features without the call transcripts. For each audio recording A of a given earning call E, the goal is to predict the company's stock volatility as a regression task. Following Kogan et al. (2009) and Qin and Yang (2019), we calculate the average log volatility τ days (temporal window) following the day of the earnings call. For each audio clip belonging to a given call, we extract 26-dimensional features with PRAAT (Boersma and Van Heuven, 2001). Each audio embedding of the call is fed sequentially to a bi-directional long short term memory network (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997), followed by an attention layer and two fully-connected layers. The model is trained by optimizing the Mean Square Error (MSE) between the predicted and true stock volatility. For all results, we report MSE on the test set for a 70:10:20 temporal split of the data (Qin and Yang, 2019). The results are averaged over 5 seeds.

SENTIMENT ANALYSIS For our sentiment analysis task, we use the Trustpilot Corpus (Hovy, Johannsen, and Søgaard, 2015)[14]. It consists of text-based user reviews from the Trustpilot website, rating companies and services on a 1 to 5 star scale. The reviews spans 5 different countries; Germany, Denmark, France, United Kingdom and USA, however, we only consider the English reviews, i.e. UK and US. The Trustpilot contains demographic information about the gender, age and geographic location of the users, but as with the topic classification task, we only concern ourselves with the gender and age of the users. As with the topic classification task, we finetune DistilBERT on the UK and US English parts of the Trustpilot Corpus, freezing all parameters but the final encoder layer, as well as the classification layer.

60% faster, while preserving over 95% of the performance of bert-base-uncased, as measured on the GLUE language understanding benchmark (Wang et al., 2019).

11 https://www.thomsonreuters.com/en/profiles.html

12 https://www.crunchbase.com/discover/people

13 https://query.wikidata.org/

14 The Trustpilot Corpus is available from https://bitbucket.org/lowlands/release/src/master/WWW2015/data/ for non-commercial research purposes only.

Classification performance is measured as F1 scores and the results are averaged over 3 seeds.

Our implementation is a PyTorch extension of the WILDS repository[15] (Koh et al., 2021) using the DP-SGD implementation provided by the Opacus Differential Privacy framework[16]. For further details about data and training, see §A.4.2 in the Appendix.

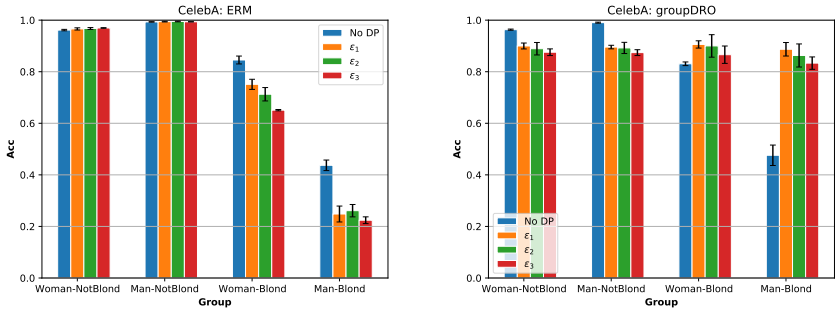


Figure 16: **Face Attribute Detection:** Performance of individual groups of increasing levels of ε. Comparing baseline ERM to Group DRO, we find that Group DRO performance on the minority group (blond males) perform much better under privacy constraints; we return to this in § 6.3.4.

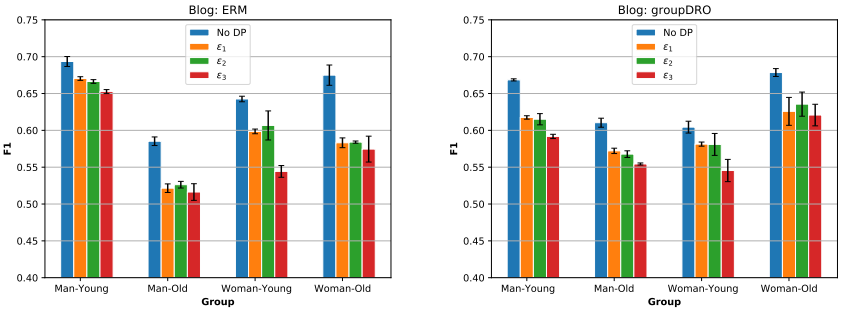


Figure 17: **Topic Classification:** Performance of individual groups of increasing levels of ε. Group DRO, compared to baseline ERM, results in a more balanced performance across all groups, even on a low privacy budget.

6.3.3 *Results*

Our results are presented in Table 12. The top half of the table presents standard (average) performance numbers across multiple runs of ERM and Group DRO at different privacy levels. Recall that performance for sentiment analysis as well as topic classification is measured in F1, volatility forecasting is measured in MSE and face recognition is mea-

15 https://github.com/p-lambda/wilds/
16 https://opacus.ai/

sured in accuracy. The accuracy of our ERM face attribute detection classifier is 0.954 in the non-private setting, for example.

Our first observation is that, as hypothesized earlier, differential privacy hurts model performance. For our smallest text-based dataset (T-US), performance becomes very poor at the strictest privacy level. This is however associated with a high amount of variance between seeds, see Figure 24 in the Appendix. The above face attribute detection classifier, which had an accuracy of 0.954 in the non-private setting, has a performance of 0.932 at this level.

DIFFERENTIAL PRIVACY HURTS FAIRNESS IN ERM The effect of differential privacy on fairness (bottom half of Table 12) is also quite consistent. The gap between the majority group and the minority group (or, more precisely, the best-performing and the worst-performing demographic subgroup) widens with increased privacy. In face recognition, for example, the accuracy gap between the two groups is 0.556 without differential privacy, but 0.770 at the strictest privacy level.

DIFFERENTIAL PRIVACY INCREASES FAIRNESS IN GROUP DRO For Group DRO, we see the opposite effect. For 4/5 datasets, we see that differential privacy leads to an increase in fairness. For face recognition, for example, the gap goes from 0.514 in the non-private setting to 0.056 in the strictest, basically disappearing. This is also illustrated in the bar plots in Figure 16. See Figure 17 for similar bar plots of the topic classification results; we include similar plots for other tasks in the Appendix. We do also observe that this increase in privacy can be expensive in terms of overall performance (e.g. Trustpilot-US). Note that the increase in fairness at higher privacy levels is seemingly at odds with previous results suggesting that privacy and fairness conflict, e.g., Agarwal (2021). We return to this question in § 6.3.4.

Note also that the only exception to the latter trend is for volatility forecasting, where differential privacy hurts fairness both in ERM and Group DRO (though Group DRO mitigates the disparity). This speech-based prediction is the only regression task, and the only task for which we do not rely on pre-trained models trained on public data.

For this task, we further analyze group disparity for varying temporal windows (τ) used to calculate target volatility values, along with increasingly strict privacy budgets (ε) in Figure 18. The disparity between subgroups widens with stricter privacy guarantees (Bagdasaryan, Poursaeed, and Shmatikov, 2019). This gap is significant for lower values of τ, strengthening the hypothesis that short-term volatility forecasting is much harder than long-term (Qin and Yang, 2019), especially for minority classes due to the disproportionate impact of noise. Comparing ERM and Group DRO, we find Group DRO mitigates this disparity gap. We observe disparity reduces with increasing

temporal window, since stock prices over a larger time frame are comparatively more stable (Qin and Yang, 2019). As a consequence, the influence of Group DRO for higher τ (6, 7) is reduced, despite facilitating faster convergence. Most importantly, we observe the power of Group DRO in mitigating the disparity caused by strict privacy safeguards ($\varepsilon = 0.96$) for crucial short term prediction ($\tau = 3$) tasks.

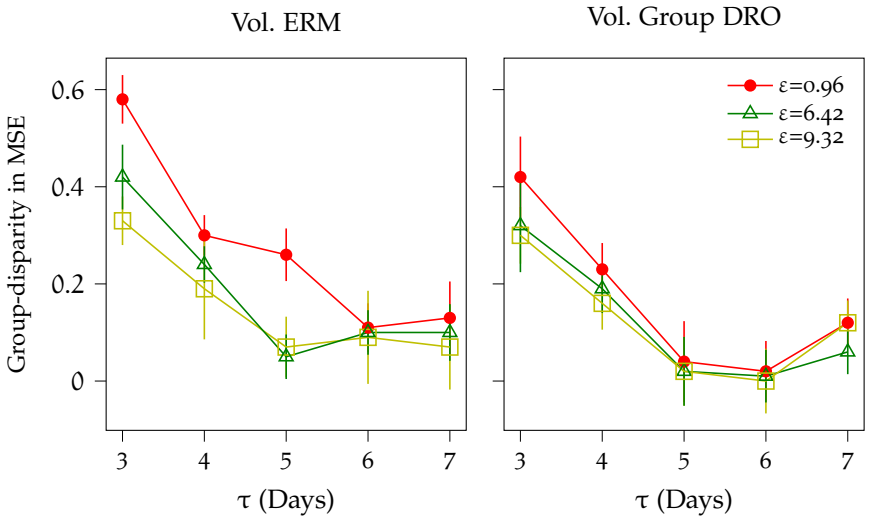


Figure 18: **Volatility Forecasting:** A comparison of group-disparity between subgroups for increasing temporal volatility windows (τ) and privacy budgets (ε), over 5 independent runs.

6.3.4 *Discussion*

It is well-known that differential privacy comes with a performance cost (Bassily, Smith, and Thakurta, 2014; Shokri and Shmatikov, 2015).[17] However, recent work has additionally shown that differential privacy is at odds with most, if not all, definitions of fairness, including equalized risk (Bagdasaryan, Poursaeed, and Shmatikov, 2019; Cummings et al., 2019; Ekstrand, Joshaghani, and Mehrpouyan, 2018; Farrand et al., 2020). Our work makes two important contributions: (a) We evaluate and confirm this hypothesis at a larger scale than previous studies for standard empirical risk minimization; and (b) we point out that the opposite holds true in the context of Group Distributionally Robust Optimization: Here, adding differential privacy improves fairness (equalized risk).

17 A multitude of algorithmic improvements have been proposed to mitigate the overall accuracy drop caused by the increased privacy protection — including private sampling from hyperbolic word representation spaces (Feyisetan, Diethe, and Drake, 2019), auto-encoder-based transformation (Krishna, Gupta, and Dupuy, 2021), Gaussian f-differential privacy (Bu et al. 2020), and gradient denoising (Nasr et al., 2020). It is yet to be examined, if the empirical application of such utility preservation techniques affects the disparate impact issue.

While (b) at first seems to contradict the very hypothesis that (a) confirms – namely that privacy is at odds with fairness – we believe the explanation is quite simple, namely that we are observing two opposite trends (at the same time): On one hand, differential privacy adds disproportionate noise to minority group examples; but on the other hand, it adds Gaussian noise which acts as a regularizer to improve robust optimization.

In their evaluation of Group Distributionally Robust Optimization, Sagawa et al. (2020a) observe that robustness is only achieved in the context of heavy regulation; specifically, they show fairness improvements when they add ℓ_2 regularization or early stopping. The ℓ_2 regularization and early stopping did not increase fairness under ERM, but seemed to 'activate' Group DRO. This makes intuitive sense: Since regularized models cannot perfectly fit the training data, heavily regularized Group DRO sacrifices average performance for worst-case performance and obtain better generalization. In the absence of regularization, however, Group DRO is less effective.

In our experiments (§ 6.3), we add minimal regularization to Group DRO, following the implementation in Koh et al. (2021), but differential privacy, we argue, provides that additional regularization. To see this, remember that DP-SGD works by Gaussian noise injection. Gaussian noise injection is known to be near-equivalent to ℓ_2-regularization and early stopping (Bishop, 1995). DP-SGD simply makes the trade-off more urgent.

6.4 RELATED WORK

FAIR MACHINE LEARNING Early work on mitigating group-level disparities included oversampling (Guo and Viktor, 2004; Shen, Lin, and Huang, 2016) and undersampling (Barandela et al., 2003; Drummond, 2003), as well as instance weighting (Shimodaira, 2000). Other proposals modify existing training algorithms or cost functions to obtain fairness (Chung, Lin, and Yang, 2015; Havaei et al., 2017; Khan et al., 2017). In the context of large-scale deep neural networks, Group DRO is a particularly interesting approach to mitigating group-level disparities (Creager, Jacobsen, and Zemel, 2021; Michel, Hashimoto, and Neubig, 2021). See Williamson and Menon (2019) and Corbett-Davies and Goel (2018) for interesting discussions of how fairness has been measured. More recent alternatives to Group DRO include Invariant Risk Minimization (Arjovsky et al., 2020), Spectral Decoupling (Pezeshki et al., 2020) and Adaptive Risk Minimization (Zhang et al., 2021). We ran experiments with both Invariant Risk Minimization and Spectral Decoupling, but they performed much worse than Group DRO.

FAIRNESS AND PRIVACY Recent studies suggest that privacy-preserving methods such as differential privacy tend to disproportionately affect minority class samples (Bagdasaryan, Poursaeed, and Shmatikov, 2019; Cummings et al., 2019; Ekstrand, Joshaghani, and Mehrpouyan, 2018; Farrand et al., 2020). Pannekoek and Spigler (2021) show that it *is* possible to learn *somewhat private* and *somewhat fair* classifiers, in their case by combining differential privacy and reject option classification. Jagielski et al. (2019) introduced the so-called DP-oracle-learner, derived from an *oracle-efficient* algorithm (Agarwal et al., 2018), which satisfies equalized odds, an alternative notion of fairness (Williamson and Menon, 2019). Lyu et al. (2020) introduced Differentially Private GANs (DPGANs), while Tran, Fioretto, and Van Hentenryck (2021) utilize Lagrangian duality to integrate fairness constraints to protected attributes. Group DRO has, to the best of our knowledge, not been studied under differential privacy before.

6.5 ETHICS STATEMENT

Training fair machine learning models often relies on training data with private demographic information, and while techniques have been introduced to minimize the risk of leakage (Hu et al., 2019), this is a valid concern. Veale and Binns (2017) discuss this problem at length in the context of businesses with commercial interests in model predictions, and present three proposals for mitigating the risk of leakage, including using third parties to store data and incorporate fairness constraints into model-building in a privacy-preserving manner, using collaborative online platforms to share knowledge and to promote transparency and fairness in machine learning systems, and to consider unsupervised learning of fairness (Hashimoto et al., 2018). The protected attributes that we rely on in the above experiments were all self-reported, in a manner detailed in the corresponding publications, and they are insufficient to identify people. We hope the above findings can contribute to the development of methods for scenarios in which both privacy and fairness are required.

6.6 CONCLUSIONS

In § 6.2, we summarized previous work suggesting that differential privacy and fairness are at odds. In §6.3, we then confirmed this hypothesis at scale, across five datasets, spanning four tasks and three modalities, showing that for Empirical Risk Minimization, stricter levels of privacy consistently *hurt* fairness. This holds true even after pre-training on large-scale public datasets (Kerrigan, Slack, and Tuyls, 2020). In the context of Group-aware Distributionally Robust Optimization (Group DRO) (Sagawa et al., 2020a), however, which is designed to mitigate group-level performance disparities (optimizing

for equalized risk), we saw the opposite effect: Strict levels of differential privacy were associated with an *increase* in fairness. In § 6.3.4, we discuss how this aligns well with the observation that Group DRO works best in the context of heavy ℓ_2 regularization, keeping in mind that Gaussian noise injection is near-equivalent to ℓ_2 regularization (Bishop, 1995).

Part IV

CONCLUSION

7

DISCUSSION AND CONCLUSION

The previous chapters of this thesis present new work in the fields of dialogue systems and fairness in NLP. In Part ii, we looked into several challenges that current dialogue systems face, such as domain adaptation, retrieving semantic frames from a conversational context and how potential biases might arise in the data collection process. To this end, we revisit the initial research questions posed initially, the first being:

How do we leverage user feedback to more efficiently improve the generalization capabilities of our dialogue systems?

Collecting highly annotated conversational corpora for data-hungry dialogue systems is an involved and costly process. In Chapter 2 we investigated how to adapt task-oriented dialogue systems to new domains, motivated by user feedback in a real-world setting. By leveraging reward signals collected at the end of a dialogue, as opposed to every turn, we showed how reinforcement learning can be used to transfer knowledge of already trained models efficiently to new domains and even further improve in-domain performance.

The next challenge we addressed concerned how conversational QA systems sometimes fails to capture challenging aspects of dialogue, such as elliptical constructions. We approach this from the viewpoint of our next research question:

How can we resolve implicit content from a conversational context to improve the quality of our dialogue systems?

In Chapter 3, we studied the task of resolving conversational sluices, i.e. identifying the elided material of one-word questions from a conversational context. We introduced a new resource of annotated conversational sluices and presented a series of baselines using heuristics, encoder-decoder frameworks and pre-trained LMs. Our results show that framing the task as a language generation task allows transformer-based models to produce high-quality sluice resolutions. A human evaluation study revealed that resolutions generated by a fine-tuned GPT-2 model sometimes rival human-generated resolutions.

Dialogue systems are heavily dependent on the labelled dataset that we manually annotate. When we train our models on such datasets, we implicitly assume that the annotators are without bias. This assumption leads us to our next research question:

To what extent does the formulation of conversational data collection guidelines influence the resulting corpora?

In Chapter 4 we introduced the concept of guideline bias. We studied the downstream effect that unintended priming of annotators through guidelines has on our models when trained on such biased resources. Using two recent datasets curated using the Wizard-of-Oz setup, we showed two things: First, how a lexical bias in the guidelines, which we confirm through a controlled priming experiment, can lead to overestimated model performance and how we can mitigate it. Second, how the order of the described conversation goals can lead to a bias in the order in which the annotators pursue them in the dialogue. Due to the rising number of new datasets released every year, it is increasingly important to know how the data we base our models on is curated. Integrating a check for guideline bias in frameworks such as datasheets (Gebru et al., 2018) or data statements (Bender and Friedman, 2018) could be a reasonable step to create more awareness around this issue.

Another type of bias we examined in this thesis was demographic bias, i.e. when the models induced from the data learn spurious correlations based on the protected demographics. In Part iii, we examined model fairness in NLP under different settings. The first was motivated by the deployment of models in a resource-constrained environment, e.g. mobile devices, where parameter pruning methods are often used for reduced inference and storage cost. Our research question here was:

How well does our NLP models satisfy fairness principles when subject to compression techniques?

In Chapter 5, we analyzed the lottery ticket extraction from the angle of algorithmic fairness in NLP. We hypothesized that systematic biases are exacerbated when models are forced not to rely on weak evidence, which pruned models to a greater extent are unable to. We introduced a new metric that measures Rawlsian min-max group disparity across demographics as a function of pruning level. We showed that heavily pruned models are associated with higher levels of group performance disparity. Additionally, we show that robust optimization techniques can at times increase model fairness among winning tickets. Somewhat contrary to our findings, recent work by Diffenderfer et al. (2021) argues that compressed models *can* be robust to distributional shifts when using rewind-based pruning techniques, such as lottery ticket extraction. However, their experiments only analyse robustness across surface-level corruptions in image recognition systems and at less extreme pruning levels. This indicates that this area of research still needs to be explored further to fully understand how different compression methods affect fairness.

Our last research question pertains to another aspect of fairness, namely how model privacy guarantees complies with popular group disparity mitigation methods:

How is fairness affected by group robust objectives when under the influence of privacy preserving methods?

In Chapter 6 we tackled this question by evaluating model fairness on a set of different tasks, spanning NLP, facial recognition, and speech, in both a baseline Empirical Risk Minimization (ERM) setting as well as a Group-aware Distributionally Robust Optimization (Group DRO) setting. Like in Chapter 5, our notion of fairness was derived from a Rawlsian min-max perspective. In line with previous work, we found that strict levels of privacy hurt fairness; however, we also observe that differentially private models trained with Group DRO reduces group disparity, sometimes even to a great extent. As heavy regularization is essential for reducing worst-group error rates with Group DRO, we hypothesized that DP can be interpreted as regularization.

An issue we face in our studies of Chapter 5 and 6 is the need for annotated demographics for every data point. This is a severely limiting factor when it comes to mitigating bias as many techniques directly rely on these annotations (Sagawa et al., 2020a), but it is also a problem when it comes to just identifying the underlying issue in the first place. Hooker et al. (2020) also highlights the issue of limited access to demographic information and proposes a method for surfacing demographic groups that models find challenging, using model compression. They do so by measuring where performance diverges between full and compressed networks. These data points can then be submitted to domain experts for further annotation as a human-in-the-loop auditing tool. This type of human-in-the-loop auditing tool would not only be beneficial for mitigating bias in conversational systems but also ML in general. Examining how differentially private models could also be used in the same manner, or in combination with compressed models, to screen our datasets for potential biases is a direction that warrants further attention.

Part V

APPENDIX

A

SUPPLEMENTARY MATERIAL FOR INDIVIDUAL STUDIES

A.1 CHAPTER 2

System utterance	User utterance	Baseline prediction	PG fine-tune prediction
N/A	I'm looking for a cheap place to dine, preferably in the centre of town.	inform(area=center) inform(pricerange=expensive)	inform(area=center) inform(pricerange=cheap)
Yes, I have 4 results matching your request, is there a price range you're looking for?	I would like moderate price range please.	inform(pricerange=expensive)	inform(pricerange=moderate)
There are a number of options for Indian restaurants in the centre of town.What price range would you like ?	I would prefer cheap restaurants.	inform(pricerange=expensive)	inform(pricerange=cheap)

Table 13: Comparison of example turn predictions from the MultiWOZ dataset between the baseline model trained on the HOTEL domains, and the policy gradient fine-tuned model. Green indicates a correct prediction whereas red indicates a wrong prediction.

A.2 CHAPTER 4

GENERAL INSTRUCTIONS The goal of this type of dialogue is for you to get the users to explain their movie preferences: The KIND of movies they like and dislike and WHY. We really want to end up finding out WHY they like what they like movie AND why the DON'T like what they don't like. We want them to take lots of turns to explain these things to you.
IMPORTANT We want users to discuss likes and dislikes for kinds of movies rather than just about specific movies. (But we trigger these more general preferences based on remembering certain titles.) You may bring up particular movie titles in order to get them thinking about why they like or dislike that kind of thing. Do not bring up particular directors, actors, or genres. For each session do the following steps:

1. Start with a normal introduction: Hello. I'd like to discuss your movie preferences.
2. Ask them what kind of movies they like and why they generally like that kind of movie.
3. Ask them for a particular movie name they liked.
4. Ask them what about that KIND of movie they liked. (get a couple of reasons at least – let them go on if they choose)
5. Ask them to name a particular movie they did not like.
6. Ask them what about that movie they did not like. (get a couple of reasons at least or let them go on if they choose)
7. Now choose a movies using the movie generator link below. Ask them if they liked that movie (if they haven't seen it: (a) ask if they have heard of it. If so, ask if they would see it (b) then choose another that they have seen to ask about). Once you find a movie from the list they have seen, ask them why they liked or disliked that kind of movie (get a couple of reasons).
8. Finally, end the conversation gracefully

Figure 19: CCPE-M Guidelines to Assistants

A.3 CHAPTER 5

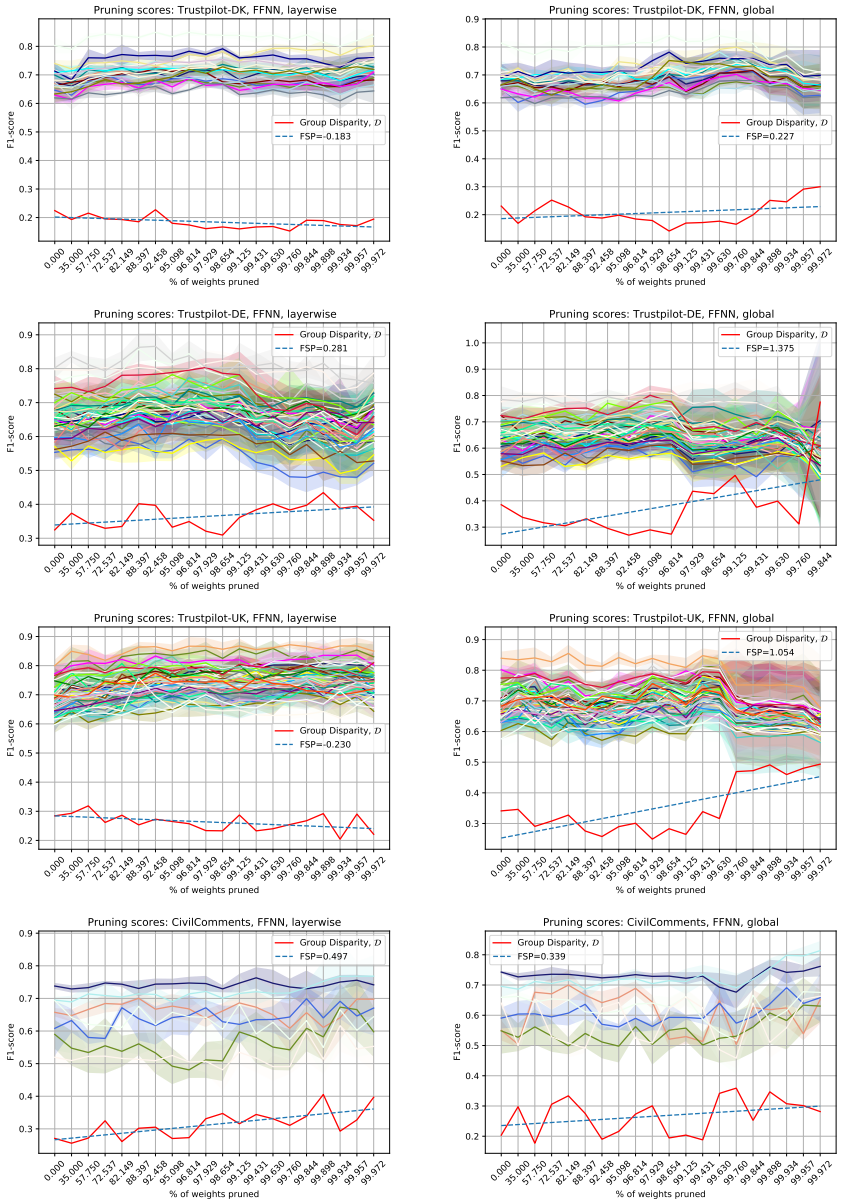


Figure 20: Macro-averaged performance of our feed-forward networks as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.

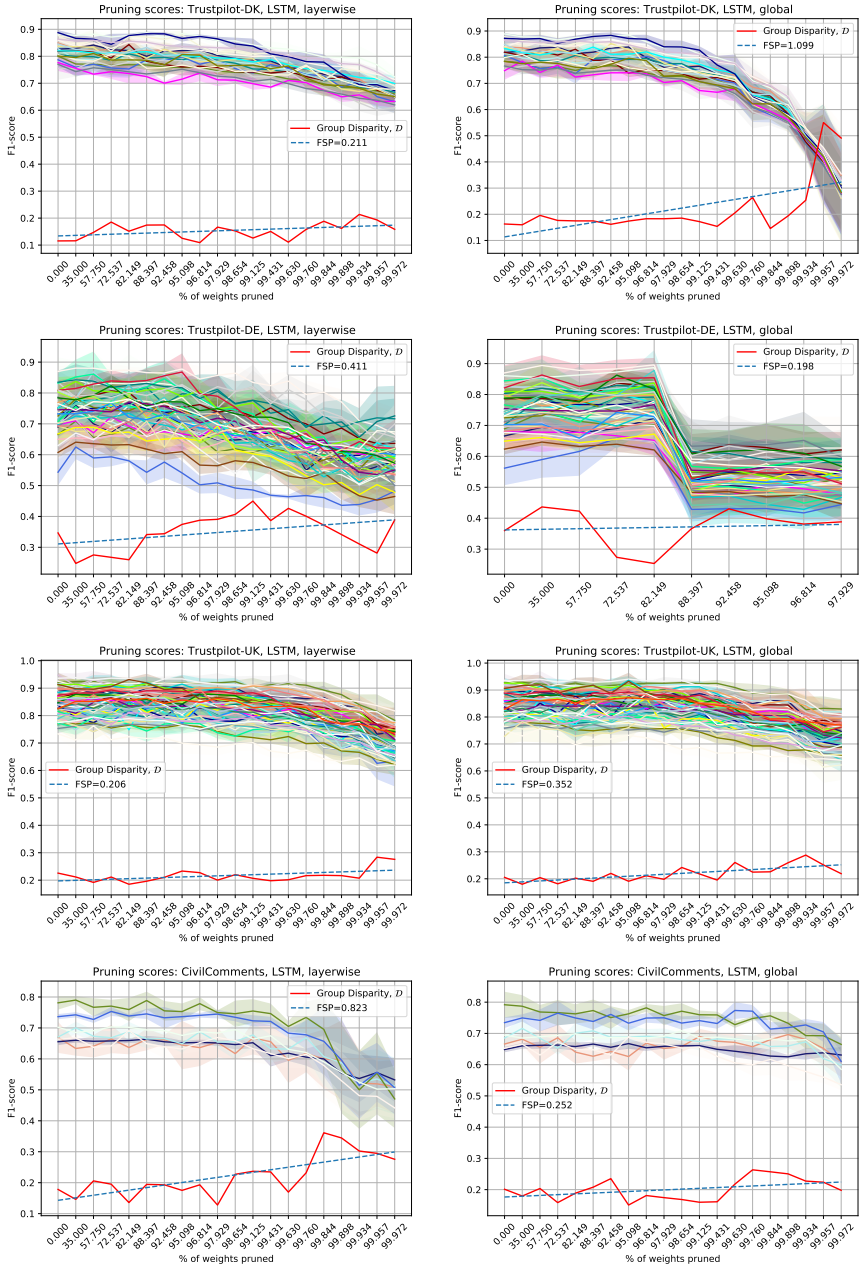


Figure 21: Macro-averaged performance of our LSTMs as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.

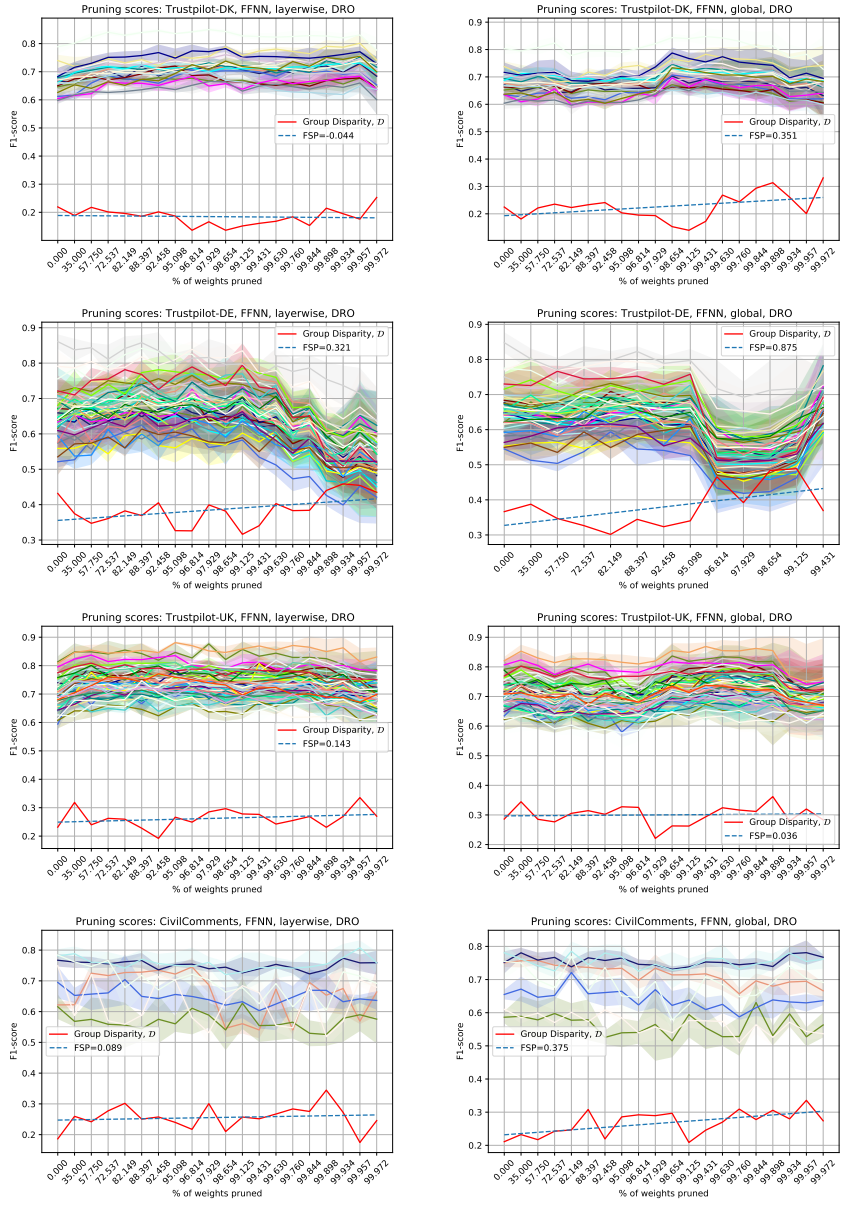


Figure 22: Macro-averaged performance of our layer-wise and globally pruned feed-forward networks trained with DRO as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.

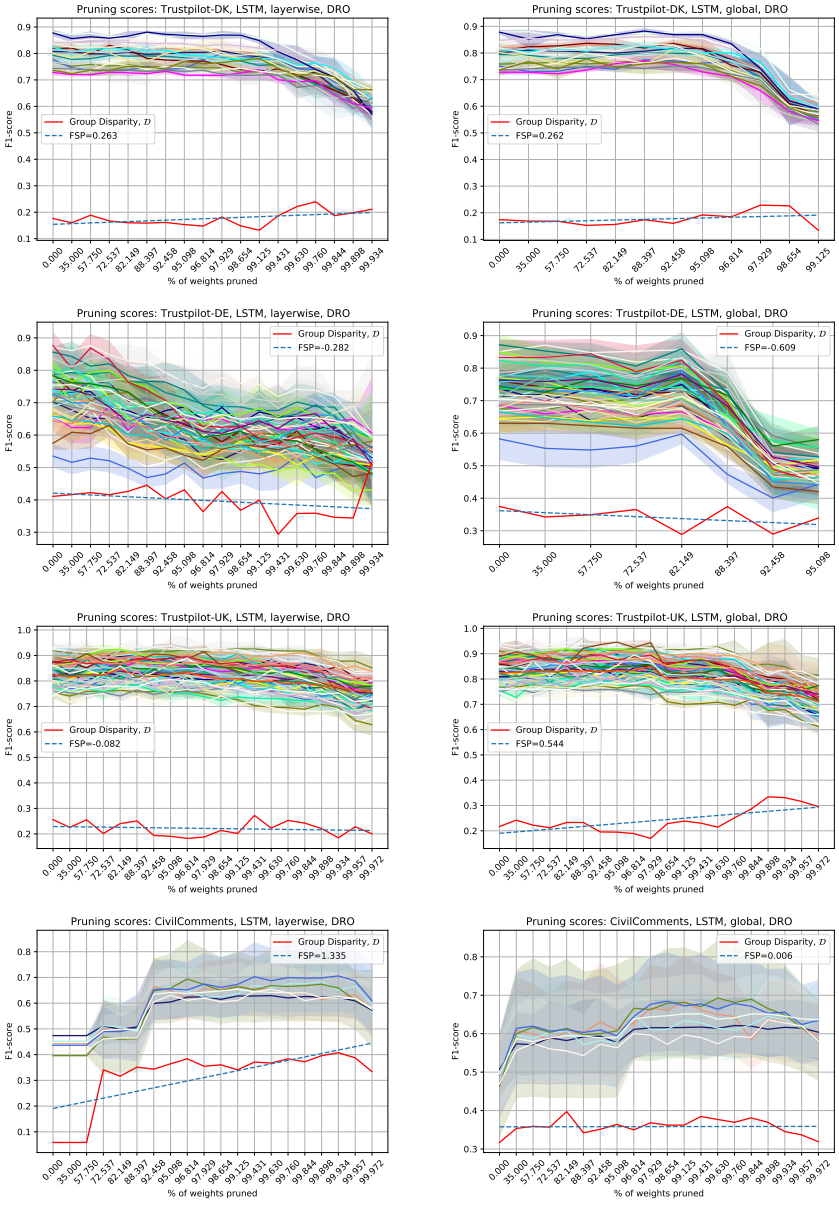


Figure 23: Macro-averaged performance of our layer-wise and globally pruned LSTM networks trained with DRO as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.

A.4 CHAPTER 6

A.4.1 *Additional Figures*

This section contains group-specific bar-plots for the performance on individual groups in the Trustpilot Corpus. For barplots on CelebA and Blog Authorship, see Figure 16 and 17.

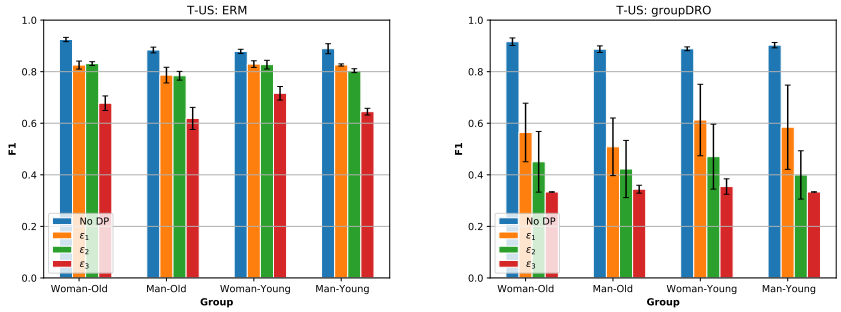


Figure 24: Performance of individual groups of increasing levels of ε for the Trustpilot-US corpus. Error bars show standard deviation over 3 individual seeds.

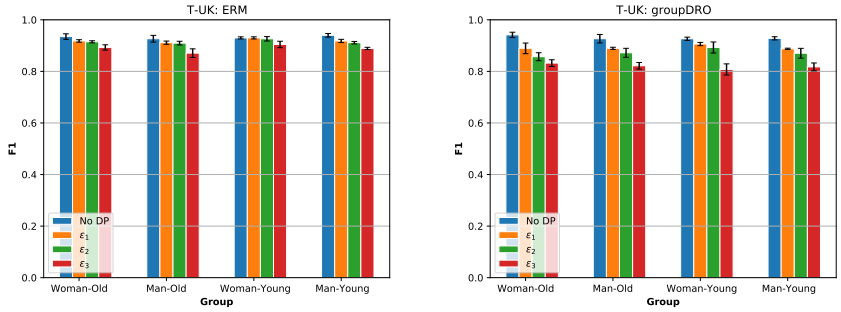


Figure 25: Performance of individual groups of increasing levels of ε for the Trustpilot-UK corpus. Error bars show standard deviation over 3 individual seeds.

A.4.2 *Experimental Details*

This section contains additional details surrounding the experiments described in § 6.3.

CELEBA We use the same processed version of the CelebA dataset as Sagawa et al. (2020a) and Koh et al. (2021), that is, we use the same train/val/test splits as Liu et al. (2015) with the *Blond Hair* attribute as the target with the *Male* attribute being the spuriously correlated variable. See group distribution in the training data in Table 14.

Group	Non-Blond, Man	Blond, Man	Non-Blond, Woman	Blond, Woman
Count	66874	1387	71629	22880

Table 14: Group distribution in the training set of CelebA

BLOG AUTHORSHIP CORPUS In addition to the preprocessing described in § 6.3, we split the data into a 60/20/20 train/val/test split (you can find the exact seed that generates the splits in our code). See group distribution in the training data in Table 15.

Group	Young, Man	Old, Man	Young, Woman	Old, Woman
Count	27222	2295	12750	2435

Table 15: Group distribution in the training set of Blog Authorship corpus

EARNINGS CONFERENCE CALLS Out of the 559 calls, we only include 535 datapoints that contain self-reported demographic attributes about gender. See Table 16 for group distributions for the training data. The target stock volatility variable is calculated following Kogan et al., 2009; Qin and Yang, 2019, defined by:

$$v_{[t-\tau,t]} = \ln\left(\sqrt{\frac{\sum_{i=0}^{\tau}(r_{t-i}-\bar{r})^2}{\tau}}\right) \tag{8}$$

Here r_t is the return price at day t and $\bar{r}$ the mean of return prices over the period of $t-\tau$ to t. We refer to τ as the temporal volatility window in our experiments. The return price r_t is defined as $r_t = \frac{P_t}{P_{t-1}} - 1$ where P_t is the closing price on day t.

Group	Man	Woman
Count	333	42

Table 16: Group distribution in the training set of Earnings Conference Calls

TRUSTPILOT We only include the datapoints that contains complete demographic attributes, i.e. the gender, age and location, but as with our topic classification experiments, we only study the group that we can define based on age and gender. All attributes are self-reported. For training we divide the reviews into the four resulting groups (*Old-Man*, *Young-Woman*, etc.) and downsample the largest groups to match the size of the smallest group. For validation as well as testing, we withhold 200 samples from each demographic with an even distribution among the ratings (1 to 5). The review scores are then binarized by grouping positive (4 and 5 stars) and negative (1 and 2 stars) and discarding neutral ones (3 stars). For a similar use of this binarization scheme, see Gupta, Thadani, and O'Hare (2020) and Desai, Zhan, and Aly (2019). See the group distributions for the training data in Table 17 and 18 for the US and UK tasks respectively.

Group	Young, Man	Old, Man	Young, Woman	Old, Woman
Count	7242	7210	7222	7255

Table 17: Group distribution in the training set of Trustpilot-US

Group	Young, Man	Old, Man	Young, Woman	Old, Woman
Count	18464	18693	18554	18693

Table 18: Group distribution in the training set of Trustpilot-UK

BILSTM The BiLSTM model was trained using a Nvidia Tesla K80 GPU. We use a learning rate of $1e^{-2}$ and train using DP-SGD for 30 epochs using a virtual batch size of 32. The average sequence length of the audio embeddings is 159. We set the maximum sequence length to 150 as we did not observe a performance increase for higher values. We run 5 individual seeds for each configuration.

In our differentially private experiments with the BiLSTM (i.e Earnings Conference Calls), we fix the gradient clipping C to 0.8. By specifying various approximate target levels of $\varepsilon \in \{1, 5, 10\}$ a corresponding noise multiplier σ is computed with the Opacus framework, based on the batch size and number of training epochs.

DISTILBERT We finetune DistilBERT on the Trustpilot corpus and Blog Authorship corpus for 20 epochs each, using a batch size of 8, accumulating gradient for a total virtual batch size of 16 using the built in Opcaus functionality. We limit the number of tokens in a sequence to 256 and use a learning rate of $5e^{-4}$ with the AdamW optimizer in addition to a weight decay of 0.01. Otherwise we use the default parameters defined in the Huggingface Transformers python package (version 4.4.2). The models are trained using a single Nvidia TitanRTX GPU and each configuration takes between 5 and 14 hours to run, depending on the size of that dataset and if DP is used or not. We run 3 individual seeds for each configuration.

In our differentially private experiments with DistilBERT (i.e. Blog Authorship and Trustpilot), we fix the gradient clipping C to 1.2 and by specifying various target levels of $\varepsilon \in \{1, 5, 10\}$ a corresponding noise multiplier σ is computed with the Opacus framework, based on the batch size and number of training epochs.

RESNET50 We finetune our Resnet50 model on the CelebA dataset for 20 epochs using a batch size of 64. We optimize the model using standard stochastic gradient descent (SGD) with a learning rate of $1e^{-3}$, momentum of 0.9 and no weight decay. We train our models using a single Nvidia TitanRTX GPU and each configuration takes between 6 and 8 hours to run, depending on if DP is used or not. We run 3 individual seeds for each configuration.

As with the differentially private DistilBERT experiments, we also here fix the gradient clipping C to 1.2 and by specifying various target

levels of $\varepsilon \in \{1, 5, 10\}$ a corresponding noise multiplier σ is computed with the Opacus framework, based on the batch size and number of training epochs.

BIBLIOGRAPHY

AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018). New Orleans, LA, USA: Association for Computing Machinery.

Abadi, Martin, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang (2016). "Deep learning with differential privacy." In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318.

Abid, Abubakar, Maheen Farooqi, and James Zou (2021). "Large language models associate Muslims with violence." In: *Nature Machine Intelligence* 3.6, pp. 461–463.

Adiwardana, Daniel et al. (2020). "Towards a Human-like Open-Domain Chatbot." In: *CoRR* abs/2001.09977.

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach (2018). "A reductions approach to fair classification." In: *International Conference on Machine Learning*. PMLR, pp. 60–69.

Agarwal, Sushant (2021). "Trade-Offs between Fairness and Privacy in Machine Learning." In: *IJCAI 2021 Workshop on AI for Social Good*.

Alabi, Daniel, Nicole Immorlica, and Adam Kalai (2018). "Unleashing Linear Optimizers for Group-Fair Learning and Optimization." In: *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 2043–2066.

Alvim, Mário S, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi (2011). "Differential privacy: on the trade-off between utility and information leakage." In: *International Workshop on Formal Aspects in Security and Trust*. Springer, pp. 39–54.

Amidei, Jacopo, Paul Piwek, and Alistair Willis (Aug. 2018). "Rethinking the Agreement in Human Evaluation Tasks." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 3318–3329.

Anand, Pranav and Daniel Hardt (Nov. 2016). "Antecedent Selection for Sluicing: Structure and Content." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1234–1243.

Anand, Pranav and Jim McCloskey (June 2015). "Annotating the Implicit Content of Sluices." In: *Proceedings of The 9th Linguistic Annotation Workshop*. Denver, Colorado, USA: Association for Computational Linguistics, pp. 178–187.

Arjovsky, Martin, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz (2020). *Invariant Risk Minimization*.

Austin, John Langshaw (1962). *How to Do Things with Words*. Clarendon Press.

Bagdasaryan, Eugene, Omid Poursaeed, and Vitaly Shmatikov (2019). "Differential Privacy Has Disparate Impact on Model Accuracy." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Baird, Austin, Anissa Hamza, and Daniel Hardt (May 2018). "Classifying Sluice Occurrences in Dialogue." In: *Proceedings of the 11th Language Resources and Evaluation Conference*. Miyazaki, Japan: European Language Resource Association.

Barandela, Ricardo, E Rangel, José Salvador Sánchez, and Francesc J Ferri (2003). "Restricted decontamination for the imbalanced training sample problem." In: *Iberoamerican congress on pattern recognition*. Springer, pp. 424–431.

Barrett, Maria, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard (Nov. 2019). "Adversarial Removal of Demographic Attributes Revisited." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 6330–6335.

Bartoldson, Brian, Ari S. Morcos, Adrian Barbu, and Gordon Erlebacher (2020). "The Generalization-Stability Tradeoff In Neural Network Pruning." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin.

Bassily, Raef, Adam Smith, and Abhradeep Thakurta (2014). "Private empirical risk minimization: Efficient algorithms and tight error bounds." In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE, pp. 464–473.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 610–623.

Bender, Emily and Batya Friedman (2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." In: *TACL*.

Bertsimas, Dimitris, Vivek F. Farias, and Nikolaos Trichakis (Jan. 2011). "The Price of Fairness." In: *Oper. Res.* 59.1, 17–31.

Bishop, Chris M. (1995). "Training with Noise is Equivalent to Tikhonov Regularization." In: *Neural Computation* 7.1, pp. 108–116.

Blitzer, John, Ryan McDonald, and Fernando Pereira (2006). "Domain adaptation with structural correspondence learning." In: *Proceedings of EMNLP*.

Boersma, Paul and Vincent Van Heuven (2001). "Speak and unSpeak with PRAAT." In: *Glot International* 5.9/10, pp. 341–347.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 4356–4364.

Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*.

Bordia, Shikha and Samuel R. Bowman (June 2019). "Identifying and Reducing Gender Bias in Word-Level Language Models." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 7–15.

Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman (2019). "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification." In: *CoRR* abs/1903.04561.

Bowling, Michael and Manuela Veloso (2001). "Rational and Convergent Learning in Stochastic Games." In: *IJCAI*.

Brix, Christopher, Parnia Bahar, and Hermann Ney (2020). *Successfully Applying the Stabilized Lottery Ticket Hypothesis to the Transformer Architecture*.

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners." In: *CoRR* abs/2005.14165.

Budzianowski, Paweł and Ivan Vulić (Nov. 2019). "Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems." In: *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong: Association for Computational Linguistics, pp. 15–22.

Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic (2018). "MultiWOZ- A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented

Dialogue Modelling." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026.

Buolamwini, Joy and Timnit Gebru (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, pp. 77–91.

Byrne, Bill, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik (Nov. 2019). "Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 4516–4525.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases." In: *Science* 356.6334, pp. 183–186.

Calmon, Flavio, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney (2017). "Optimized Pre-Processing for Discrimination Prevention." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Campbell, Norm A. (1978). "The Influence Function as an Aid in Outlier Detection in Discriminant Analysis." In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 27.3, pp. 251–258.

Caswell, Isaac et al. (2021). "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets." In: *CoRR* abs/2103.12028.

Cercas Curry, Amanda and Verena Rieser (June 2018). "#MeToo Alexa: How Conversational Systems Respond to Sexual Harassment." In: *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 7–14.

Chang, Hongyan and Reza Shokri (2021). *On the Privacy Risks of Algorithmic Fairness*.

Chen, Hongshen, Xiaorui Liu, Dawei Yin, and Jiliang Tang (Nov. 2017). "A Survey on Dialogue Systems: Recent Advances and New Frontiers." In: *SIGKDD Explor. Newsl.* 19.2, 25–35.

Chen, Lu, Cheng Chang, Zhi Chen, Bowen Tan, Milica Gašić, and Kai Yu (2018). "Policy adaptation for deep reinforcement learning-based dialogue management." In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6074–6078.

Chen, Tianlong, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin (2020). "The Lottery Ticket Hypothesis for Pre-trained BERT Networks." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 15834–15846.

Chernick, M. and V. K. Murthy (1983). "The Use of Influence Functions for Outlier Detection and Data Editing." In: *American Journal of Mathematical and Management Sciences* 3, pp. 47–61.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer (2018). "QuAC : Question Answering in Context." In: *CoRR* abs/1808.07036.

Christakopoulou, Konstantina, Filip Radlinski, and Katja Hofmann (2016). "Towards Conversational Recommender Systems." In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 815–824.

Chung, Yu-An, Hsuan-Tien Lin, and Shao-Wen Yang (2015). "Cost-aware pre-training for multiclass cost-sensitive deep learning." In: *arXiv preprint arXiv:1511.09337*.

Cohn, Trevor and Lucia Specia (Aug. 2013). "Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 32–42.

Colby, K., S. Weber, and F. D. Hilf (1971). "Artificial Paranoia." In: *Artif. Intell.* 2, pp. 1–25.

Colman, Marcus, Arash Eshghi, and Pat Healey (June 2008). "Quantifying Ellipsis in Dialogue: an index of mutual understanding." In: *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Columbus, Ohio: Association for Computational Linguistics, pp. 96–99.

Corbett-Davies, Sam and Sharad Goel (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*.

Creager, Elliot, Jörn-Henrik Jacobsen, and Richard Zemel (2021). *Environment Inference for Invariant Learning*.

Cummings, Rachel, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern (2019). "On the compatibility of privacy and fairness." In: *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 309–315.

Cun, Yann Le, John S. Denker, and Sara A. Solla (1990). "Optimal Brain Damage." In: *Advances in Neural Information Processing Systems 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 598–605.

Dandapat, Sandipan, Priyanka Biswas, Monojit Choudhury, and Kalika Bali (Aug. 2009). "Complex Linguistic Annotation – No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks." In: *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*. Suntec, Singapore: Association for Computational Linguistics, pp. 10–18.

Dastin, Jeffrey (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*.

Daume III, Hal and Daniel Marcu (2006). "Domain adaptation for statistical classifiers." In: *Journal of Artificial Intelligence Research* 26, pp. 101–126.

Desai, Shrey, Hongyuan Zhan, and Ahmed Aly (2019). "Evaluating Lottery Tickets Under Distributional Shifts." In: *CoRR* abs/1910.12708.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Dhingra, Bhuwan, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng (2017). "Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 484–495.

Dickinson, Markus and W. Detmar Meurers (Apr. 2003). "Detecting Errors in Part-of-Speech Annotation." In: *10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary: Association for Computational Linguistics.

Diffenderfer, James, Brian R. Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura (2021). "A Winning Hand: Compressing Deep Networks Can Improve Out-Of-Distribution Robustness." In: *CoRR* abs/2106.09129.

Dinan, Emily, Gavin Abercrombie, A. Stevie Bergman, Shannon L. Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser (2021). "Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling." In: *CoRR* abs/2107.03451.

Dinan, Emily, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston (Nov. 2020). "Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8173–8188.

Dinan, Emily et al. (2019). "The Second Conversational Intelligence Challenge (ConvAI2)." In: *CoRR* abs/1902.00098.

Donini, Michele, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil (2018). "Empirical Risk Minimization Under Fairness Constraints." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc.

Drumnond, Chris (2003). "Class Imbalance and Cost Sensitivity: Why Undersampling beats Oversampling." In: *ICML-KDD 2003 Workshop: Learning from Imbalanced Datasets*.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006). "Calibrating Noise to Sensitivity in Private Data Analysis." In: *Proceedings of the Third Conference on Theory of Cryptography*. TCC'06. New York, NY: Springer-Verlag, 265–284.

Ekstrand, Michael D., Rezvan Joshaghani, and Hoda Mehrpouyan (2018). "Privacy for All: Ensuring Fair and Equitable Privacy Protections." In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, pp. 35–47.

El Asri, Layla, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman (2017). "Frames: a corpus for adding memory to goal-oriented dialogue systems." In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 207–219.

Eric, Mihail, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur (May 2020). "MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 422–428.

Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst (July 2019). "Understanding Undesirable Word Embedding Associations." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1696–1705.

Farrand, Tom, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask (2020). "Neither Private Nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy." In: *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pp. 15–19.

Fernández, Raquel, Jonathan Ginzburg, and Shalom Lappin (2007). "Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach." In: *American Journal of Computational Linguistics* 33.3, pp. 397–427.

Feyisetan, Oluwaseyi, Tom Diethe, and Thomas Drake (2019). "Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text." In: *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, pp. 210–219.

Fisher, Bonnie S. (2009). "The Effects of Survey Question Wording on Rape Estimates: Evidence From a Quasi-Experimental Design." In: *Violence Against Women* 15.2. PMID: 19126832, pp. 133–147.

Frankle, Jonathan and Michael Carbin (2019). "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Frankle, Jonathan, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin (2020). "Linear Mode Connectivity and the Lottery Ticket Hypothesis." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 3259–3269.

Friedrich, Annemarie, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal (June 2015). "Annotating genericity: a survey, a scheme, and a corpus." In: *Proceedings of The 9th Linguistic Annotation Workshop*. Denver, Colorado, USA: Association for Computational Linguistics, pp. 21–30.

Gajane, Pratik (2017). "On formalizing fairness in prediction with machine learning." In: *CoRR* abs/1710.03184.

Gao, Jianfeng, Michel Galley, and Lihong Li (July 2018). "Neural Approaches to Conversational AI." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Melbourne, Australia: Association for Computational Linguistics, pp. 2–7.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes." In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644.

Garrido-Muñoz, Ismael, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López (2021). "A Survey on Bias in Deep NLP." In: *Applied Sciences* 11.7.

Gasic, Milica, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young (2013). "POMDP-based dialogue manager adaptation to extended domains." In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 214–222.

Gašić, Milica, Nikola Mrkšić, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young (2017). "Dialogue manager domain adaptation using Gaussian process reinforcement learning." In: *Computer Speech & Language* 45, pp. 552–569.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford (2018). *Datasheets for Datasets*.

Geva, Mor, Yoav Goldberg, and Jonathan Berant (Nov. 2019). "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)*. Hong Kong, China: Association for Computational Linguistics, pp. 1161–1166.

Globerson, Amir and Sam Roweis (2006). "Nightmare at Test Time: Robust Learning by Feature Deletion." In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 353–360.

Goldberg, Yoav and Michael Elhadad (July 2010). "Inspecting the Structural Biases of Dependency Parsing Algorithms." In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Uppsala, Sweden: Association for Computational Linguistics, pp. 234–242.

Gonen, Hila and Yoav Goldberg (2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them." In: *Proceedings of NAACL-HLT*.

Gordon, Mitchell, Kevin Duh, and Nicholas Andrews (July 2020). "Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning." In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. Online: Association for Computational Linguistics, pp. 143–155.

Green, Bert F., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery (1961). "Baseball: An Automatic Question-Answerer." In: *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*. IRE-AIEE-ACM '61 (Western). Los Angeles, California: Association for Computing Machinery, 219–224.

Guo, Daya, Yibo Sun, Duyu Tang, Nan Duan, Jian Yin, Hong Chi, James Cao, Peng Chen, and Ming Zhou (2018). "Question Generation from SQL Queries Improves Neural Semantic Parsing." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1597–1607.

Guo, Hongyu and Herna L Viktor (2004). "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach." In: *ACM Sigkdd Explorations Newsletter* 6.1, pp. 30–39.

Gupta, Aakriti, Kapil Thadani, and Neil O'Hare (Dec. 2020). "Effective Few-Shot Classification with Transfer Learning." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1061–1066.

Ham, Donghoon, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim (July 2020). "End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 583–592.

Han, Song, Jeff Pool, John Tran, and William J. Dally (2015). "Learning Both Weights and Connections for Efficient Neural Networks." In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 1135–1143.

Hashimoto, Tatsunori B., Megha Srivastava, Hongseok Namkoong, and Percy Liang (2018). "Fairness Without Demographics in Repeated Loss Minimization." In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1934–1943.

Hassibi, Babak and David Stork (1993). "Second order derivatives for network pruning: Optimal Brain Surgeon." In: *Advances in Neural Information Processing Systems*. Ed. by S. Hanson, J. Cowan, and C. Giles. Vol. 5. Morgan-Kaufmann, pp. 164–171.

Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle (2017). "Brain tumor segmentation with deep neural networks." In: *Medical image analysis* 35, pp. 18–31.

He, Kaiming, X. Zhang, Shaoqing Ren, and Jian Sun (2015a). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015b). *Deep Residual Learning for Image Recognition*.

Henderson, Matthew (2015). "Machine learning for dialog state tracking: A review." In: *Proc. of The First International Workshop on Machine Learning in Spoken Language Processing*.

Henderson, Matthew, Blaise Thomson, and Jason D Williams (2014). "The second dialog state tracking challenge." In: *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 263–272.

Henderson, Matthew, Blaise Thomson, and Steve Young (2014). "Word-based dialog state tracking with recurrent neural networks." In: *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 292–299.

Henderson, Peter, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau (2018). "Ethical Challenges in Data-Driven Dialogue Systems." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Soci-*

ety. AIES '18. New Orleans, LA, USA: Association for Computing Machinery, 123–129.

Hill, Felix, Kyunghyun Cho, and Anna Korhonen (2016). "Learning Distributed Representations of Sentences from Unlabelled Data." In: *CoRR* abs/1602.03483.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

Hooker, Sara (2021). "Moving beyond "algorithmic bias is a data problem"." In: *Patterns* 2.4, p. 100241.

Hooker, Sara, Aaron C. Courville, Yann N. Dauphin, and Andrea Frome (2019). "Selective Brain Damage: Measuring the Disparate Impact of Model Pruning." In: *CoRR* abs/1911.05248.

Hooker, Sara, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton (2020). "Characterising Bias in Compressed Models." In: *CoRR* abs/2010.03058.

Hosseini-Asl, Ehsan, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher (2020). "A Simple Language Model for Task-Oriented Dialogue." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 20179–20191.

Hovy, Dirk, Anders Johannsen, and Anders Søgaard (2015). "User Review Sites as a Resource for Large-Scale Sociolinguistic Studies." In: *Proceedings of the 24th International Conference on World Wide Web*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 452–461.

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

Hu, Hui, Yijun Liu, Zhen Wang, and Chao Lan (2019). *A Distributed Fair Machine Learning Framework with Private Demographic Data Protection*.

Hu, Weihua, Gang Niu, Issei Sato, and Masashi Sugiyama (2018). "Does Distributionally Robust Supervised Learning Give Robust Classifiers?" In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 2029–2037.

Jagielski, Matthew, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman (2019). "Differentially private fair learning." In: *International Conference on Machine Learning*. PMLR, pp. 3000–3008.

Jiang, Jing and ChengXiang Zhai (2007). "Instance weighting for domain adaptation in NLP." In: *Proceedings of ACL*.

Jiang, Ziheng, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer (2020). "Exploring the Memorization-Generalization Continuum in Deep Learning." In: *CoRR* abs/2002.03206.

Jurafsky, Daniel and James H. Martin (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd. USA: Prentice Hall PTR.

Kakade, Sham M (2002). "A natural policy gradient." In: *Advances in neural information processing systems*, pp. 1531–1538.

Kazuhide, Yamamoto and Sumita Eiichiro (1998). "Feasibility Study for Ellipsis Resolution in Dialogues by Machine-learning Technique." In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*. ACL '98/COLING '98. Montreal, Quebec, Canada: Association for Computational Linguistics, pp. 1428–1435.

Kerrigan, Gavin, Dylan Slack, and Jens Tuyls (Nov. 2020). "Differentially Private Language Models Benefit from Public Pre-training." In: *Proceedings of the Second Workshop on Privacy in NLP*. Online: Association for Computational Linguistics, pp. 39–45.

Khan, Salman H, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri (2017). "Cost-sensitive learning of deep feature representations from imbalanced data." In: *IEEE transactions on neural networks and learning systems* 29.8, pp. 3573–3587.

Khullar, Payal, Konigari Rachna, Mukul Hase, and Manish Shrivastava (July 2018). "Automatic Question Generation using Relative Pronouns and Adverbs." In: *Proceedings of ACL 2018, Student Research Workshop*. Melbourne, Australia: Association for Computational Linguistics, pp. 153–158.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun.

Kiritchenko, Svetlana and Saif M. Mohammad (2018). "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems." In: **SEMEVAL*.

Kogan, Shimon, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith (2009). "Predicting risk from financial reports with regression." In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280.

Koh, Pang Wei et al. (2021). "WILDS: A Benchmark of in-the-Wild Distribution Shifts." In: *International Conference on Machine Learning (ICML)*.

Krishna, Satyapriya, Rahul Gupta, and Christophe Dupuy (Apr. 2021). "ADePT: Auto-encoder based Differentially Private Text Transformation." In: *Proceedings of the 16th Conference of the European Chap-*

ter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, pp. 2435–2439.

Lample, Guillaume, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato (2018). "Unsupervised Machine Translation Using Monolingual Corpora Only." In: *International Conference on Learning Representations*.

Le, Hung, Doyen Sahoo, Chenghao Liu, Nancy Chen, and Steven C.H. Hoi (Nov. 2020). "UniConv: A Unified Conversational Neural Architecture for Multi-domain Task-oriented Dialogues." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 1860–1877.

Lei, Wenqiang, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin (July 2018). "Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1437–1447.

Levy, Daniel, Yair Carmon, John C Duchi, and Aaron Sidford (2020). "Large-Scale Methods for Distributionally Robust Optimization." In: *Advances in Neural Information Processing Systems*.

Li, Lihong, Jason D. Williams, and Suhrid Balakrishnan (2009). "Reinforcement learning for dialog management using least-squares Policy iteration and fast feature selection." In: *INTERSPEECH*.

Li, Raymond, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal (2018). "Towards Deep Conversational Recommendations." In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 9725–9735.

Li, Xiujun, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz (Nov. 2017). "End-to-End Task-Completion Neural Dialogue Systems." In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 733–743.

Lin, Chin-Yew and Franz Josef Och (2004). "Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics." In: *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. ACL '04. Barcelona, Spain: Association for Computational Linguistics.

Liu, Bing, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck (2018a). "Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems." In: *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Vol. 1, pp. 2060–2069.

Liu, Yijia, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith (June 2018b). "Parsing Tweets into Universal Dependencies." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 965–975.

Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang (2015). "Deep Learning Face Attributes in the Wild." In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738.

Lyu, Lingjuan, Yitong Li, Karthik Nandakumar, Jiangshan Yu, and Xingjun Ma (2020). "How to democratise and protect AI: fair and differentially private decentralised deep learning." In: *IEEE Transactions on Dependable and Secure Computing*.

Marcus, Mitchell P. (June 1982). "Building Non-Normative Systems - The Search for Robustness: An Overview." In: *20th Annual Meeting of the Association for Computational Linguistics*. Toronto, Ontario, Canada: Association for Computational Linguistics, pp. 152–152.

McMahan, Brendan, Galen Andrew, Ilya Mironov, Nicolas Papernot, Peter Kairouz, Steve Chien, and Úlfar Erlingsson (2018). "A General Approach to Adding Differential Privacy to Iterative Training Procedures." In: Workshop on Privacy Preserving Machine Learning (NeurIPS 2018).

Menon, Aditya Krishna, Ankit Singh Rawat, and Sanjiv Kumar (2021). "Overparameterisation and worst-case generalisation: friend or foe?" In: *ICLR*.

Michel, Paul, Tatsunori Hashimoto, and Graham Neubig (2021). "Modeling the Second Player in Distributionally Robust Optimization." In: *International Conference on Learning Representations*.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space." In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun.

Mironov, Ilya (2017). "Rényi Differential Privacy." In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru (2019). "Model Cards for Model Reporting." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 220–229.

Mozer, Michael C. and Paul Smolensky (1989). "Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment." In: *Advances in Neural Information Processing Systems 1*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 107–115.

Mrkšić, N, DO Séaghdha, B Thomson, M Gašić, PH Su, D Vandyke, TH Wen, and S Young (2015). "Multi-domain dialog state tracking using recurrent neural networks." In: *ACL-IJCNLP 2015-53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. Vol. 2, pp. 794–799.

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young (2017a). "Neural Belief Tracker: Data-Driven Dialogue State Tracking." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1777–1788.

Mrkšić, Nikola and Ivan Vulić (July 2018). "Fully Statistical Neural Belief Tracking." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 108–113.

Mrkšić, Nikola, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young (2017b). "Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints." In: *Transactions of the Association of Computational Linguistics* 5.1, pp. 309–324.

Nguyen, Dong, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong (Aug. 2014). "Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment." In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 1950–1961.

Ni, Jinjie, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria (2021). "Recent Advances in Deep Learning Based Dialogue Systems: A Systematic Survey." In: *CoRR* abs/2105.04387.

Nouri, Elnaz and Ehsan Hosseini-Asl (2018). "Toward Scalable Neural Dialogue State Tracking." In: *NeurIPS 2018, 2nd Conversational AI workshop*.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan (2019). "Dissecting racial bias in an algorithm used to manage the health of populations." In: *Science* 366.6464, pp. 447–453.

Oren, Yonatan, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang (Nov. 2019). "Distributionally Robust Language Modeling."

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)*. Hong Kong, China: Association for Computational Linguistics, pp. 4227–4237.

Paganini, Michela (2020). *Prune Responsibly*.

Pannekoek, Marlotte and Giacomo Spigler (2021). *Investigating Trade-offs in Utility, Fairness and Differential Privacy in Neural Networks*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation." In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 311–318.

Papini, Matteo, Matteo Pirotta, and Marcello Restelli (2017). "Adaptive batch size for safe policy gradients." In: *Advances in Neural Information Processing Systems*, pp. 3591–3600.

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer (2017). "Automatic differentiation in PyTorch." In: *NIPS-W*.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna (2020). "Data and its (dis)contents: A survey of dataset development and use in machine learning research." In: *CoRR* abs/2012.05345.

Peng, Baolin, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao (2020). "SOLOIST: Few-shot Task-Oriented Dialog with A Single Pre-trained Auto-regressive Model." In: *CoRR* abs/2005.05298.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation." In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). "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)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.

Pezeshki, Mohammad, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie (2020). *Gradient Starvation: A Learning Proclivity in Neural Networks*.

Phaneuf, Alicia (2020). *Artificial intelligence in financial Services: Applications and benefits of AI in finance*.

Plank, Barbara, Dirk Hovy, and Anders Søgaard (Apr. 2014). "Learning part-of-speech taggers with inter-annotator agreement loss." In: *Proceedings of the 14th Conference of the European Chapter of the*

Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, pp. 742–751.

Popović, Maja (Sept. 2015). "chrF: character n-gram F-score for automatic MT evaluation." In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395.

Prasanna, Sai, Anna Rogers, and Anna Rumshisky (Nov. 2020). "When BERT Plays the Lottery, All Tickets Are Winning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3208–3229.

Qin, Yu and Yi Yang (July 2019). "What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 390–401.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). *Improving Language Understanding by Generative Pre-Training*.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). *Language Models are Unsupervised Multitask Learners*.

Radlinski, Filip, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi (Sept. 2019). "Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences." In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association for Computational Linguistics, pp. 353–360.

Ramadan, Osman, Paweł Budzianowski, and Milica Gasic (2018). "Large-Scale Multi-Domain Belief Tracking with Knowledge Sharing." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2, pp. 432–437.

Rapp, Reinhard (2009). "The Back-translation Score: Automatic MT Evaluation at the Sentence Level without Reference Translations." In: *ACL*.

Rastogi, Abhinav, Dilek Hakkani-Tür, and Larry Heck (2017). "Scalable multi-domain dialogue state tracking." In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 561–568.

Ratnaparkhi, Adwait (1996). "A Maximum Entropy Model for Part-Of-Speech Tagging." In: *Conference on Empirical Methods in Natural Language Processing*.

Rawls, John (1971). *A Theory of Justice*. 1st ed. Cambridge, Massachussets: Belknap Press of Harvard University Press.

Reddy, Siva, Danqi Chen, and Christopher D. Manning (2018). "CoQA: A Conversational Question Answering Challenge." In: *CoRR* abs/1808.07042.

Ren, Liliang, Kaige Xie, Lu Chen, and Kai Yu (2018). "Towards Universal Dialogue State Tracking." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2780–2786.

Rigano, Christopher (2018). *Using Artificial Intelligence to Address Criminal Justice Needs*.

Roller, Stephen et al. (Apr. 2021). "Recipes for Building an Open-Domain Chatbot." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 300–325.

Rønning, Ola, Daniel Hardt, and Anders Søgaard (June 2018). "Sluice Resolution without Hand-Crafted Features over Brittle Syntax Trees." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 236–241.

Ross, John R (1969). "Guess Who?" In: *CLS 5: Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme (June 2018). "Gender Bias in Coreference Resolution." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 8–14.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning Internal Representations by Error Propagation." In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. Ed. by David E. Rumelhart and James L. Mcclelland. Cambridge, MA: MIT Press, pp. 318–362.

Sagawa, Shiori, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang (2020a). "Distributionally Robust Neural Networks." In: *International Conference on Learning Representations*.

Sagawa, Shiori, Aditi Raghunathan, Pang Wei Koh, and Percy Liang (2020b). "An Investigation of Why Overparameterization Exacerbates Spurious Correlations." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 8346–8356.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *CoRR* abs/1910.01108.

Sanh, Victor, Thomas Wolf, and Alexander M. Rush (2020). *Movement Pruning: Adaptive Sparsity by Fine-Tuning*.

Santhanam, Sashank and Samira Shaikh (2019). "A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions." In: *CoRR* abs/1906.00500.

Schler, Jonathan, Moshe Koppel, S. Argamon, and J. Pennebaker (2006). "Effects of Age and Gender on Blogging." In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Schuman, Howard and Stanley Presser (1977). "Question Wording as an Independent Variable in Survey Analysis." In: *Sociological Methods & Research* 6.2, pp. 151–170.

Schuster, M. and K.K. Paliwal (Nov. 1997). "Bidirectional Recurrent Neural Networks." In: *IEEE Transactions on Signal Processing* 45.11, 2673–2681.

Serban, Iulian Vlad, Alberto Garcia-Duran, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio (2016). "Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus." In: *EMNLP*.

Serban, Iulian Vlad, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau (2018). "A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version." In: *Dialogue Discourse* 9.1, pp. 1–49.

Shen, Li, Zhouchen Lin, and Qingming Huang (2016). "Relay backpropagation for effective learning of deep convolutional neural networks." In: *European conference on computer vision*. Springer, pp. 467–482.

Shimodaira, Hidetoshi (Oct. 2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function." In: *Journal of Statistical Planning and Inference* 90.2, pp. 227–244.

Shokri, Reza and Vitaly Shmatikov (2015). "Privacy-preserving deep learning." In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321.

Singh, Satinder, Diane Litman, Michael Kearns, and Marilyn Walker (2002). "Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system." In: *Journal of Artificial Intelligence Research* 16, pp. 105–133.

Smith, Tom W. (1987). "That Which We Call Welfare by Any Other Name Would Smell Sweeter an Analysis of the Impact of Question Wording on Response Patterns." In: *The Public Opinion Quarterly* 51.1, pp. 75–83.

Søgaard, Anders (Aug. 2013). "Part-of-speech tagging with antagonistic adversaries." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 640–644.

Sun, Kai, Lu Chen, Su Zhu, and Kai Yu (2014). "A generalized rule based tracker for dialogue state tracking." In: *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 330–335.

Sun, Kai, Su Zhu, Lu Chen, Siqiu Yao, Xueyang Wu, and Kai Yu (2016). "Hybrid Dialogue State Tracking for Real World Human-to-Human Dialogues." In: *INTERSPEECH*, pp. 2060–2064.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks." In: *CoRR* abs/1409.3215.

Sutton, Charles, Michael Sindelar, and Andrew McCallum (June 2006). "Reducing Weight Undertraining in Structured Discriminative Learning." In: *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City, USA: Association for Computational Linguistics, pp. 89–95.

Sutton, Richard S. and Andrew G. Barto (1998). *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press.

Sze, V., Y. Chen, T. Yang, and J. S. Emer (2017). "Efficient Processing of Deep Neural Networks: A Tutorial and Survey." In: *Proceedings of the IEEE* 105.12, pp. 2295–2329.

Tran, Cuong, Ferdinando Fioretto, and Pascal Van Hentenryck (2021). "Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.11, pp. 9932–9939.

Valverde Ibañez, M. Pilar and Akira Ohtani (Dec. 2014). "Annotating Article Errors in Spanish Learner Texts: Design and Evaluation of an Annotation Scheme." In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. Phuket,Thailand: Department of Linguistics, Chulalongkorn University, pp. 234–243.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way (2018). "Getting Gender Right in Neural Machine Translation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3003–3008.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Veale, Michael and Reuben Binns (2017). "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data." In: *Big Data & Society* 4.2, p. 2053951717743530.

Verma, Sahil and Julia Rubin (2018). "Fairness Definitions Explained." In: *Proceedings of the International Workshop on Software Fairness*. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, 1–7.

Vinyals, Oriol and Quoc V. Le (2015). "A Neural Conversational Model." In: *CoRR* abs/1506.05869.

Vlachos, Andreas and Stephen Clark (2014). "A New Corpus and Imitation Learning Framework for Context-Dependent Semantic Parsing." In: *TACL*.

Voigt, Rob, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov (May 2018). "RtGender: A Corpus for Studying Differential Responses to Gender." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In: *International Conference on Learning Representations*.

Wang, Zhuoran and Oliver Lemon (2013). "A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information." In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 423–432.

Weaver, Lex and Nigel Tao (2001). "The Optimal Reward Baseline for Gradient-Based Reinforcement Learning." In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01. Seattle, Washington: Morgan Kaufmann Publishers Inc., 538–545.

Weizenbaum, Joseph (Jan. 1966). "ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine." In: *Commun. ACM* 9.1, 36–45.

Williams, Jason D. and Steve Young (2007). "Partially observable Markov decision processes for spoken dialog systems." In: *Computer Speech & Language* 21.2, pp. 393 –422.

Williams, Jason (2013). "Multi-domain learning and generalization in dialog state tracking." In: *Proceedings of the SIGDIAL 2013 Conference*. Metz, France: Association for Computational Linguistics, pp. 433–441.

Williams, Jason, Antoine Raux, Deepak Ramachandran, and Alan Black (2013). "The dialog state tracking challenge." In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 404–413.

Williams, Ronald J and Jing Peng (1991). "Function optimization using connectionist reinforcement learning algorithms." In: *Connection Science* 3.3, pp. 241–268.

Williamson, Robert and Aditya Menon (2019). "Fairness risk measures." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 6786–6797.

Wolf, Thomas, Victor Sanh, Julien Chaumond, and Clement Delangue (2019). "TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents." In: *CoRR* abs/1901.08149.

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." In: *CoRR* abs/1609.08144.

Yu, Haonan, Sergey Edunov, Yuandong Tian, and Ari S. Morcos (2020). "Playing the lottery with rewards and multiple languages: lottery tickets in RL and NLP." In: *International Conference on Learning Representations*.

Yuan, Ming, Vikas Kumar, Muhammad Aurangzeb Ahmad, and Ankur Teredesai (2021). *Assessing Fairness in Classification Parity of Machine Learning Models in Healthcare*.

Zang, Xiaoxue, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen (2020). "MultiWOZ 2.2: A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines." In: *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pp. 109–117.

Zhang, Marvin, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn (2021). *Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Distribution Shift*.

Zhang, Yizhe, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan (July 2020). "DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 270–278.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (June 2018a). "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)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 15–20.

Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang (2018b). "Learning Gender-Neutral Word Embeddings." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4847–4853.

Zhao, Tiancheng and Maxine Eskenazi (Sept. 2016). "Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning." In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles: Association for Computational Linguistics, pp. 1–10.

Zhao, Yao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke (2018c). "Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks." In: *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, pp. 3901–3910.

Zhong, Victor, Caiming Xiong, and Richard Socher (2018). "Global-Locally Self-Attentive Encoder for Dialogue State Tracking." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1458–1467.