

ANA VALERIA GONZÁLEZ

TOWARDS HUMAN-CENTERED NATURAL  
LANGUAGE PROCESSING

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



# TOWARDS HUMAN-CENTERED NATURAL LANGUAGE PROCESSING

ANA VALERIA GONZÁLEZ



UNIVERSITY OF  
COPENHAGEN

PhD Thesis

February 2021

Ana Valeria González: *Towards Human-centered Natural Language Processing*

THESIS SUPERVISOR:  
Anders Søgaard

ASSESSMENT COMMITTEE:  
Christina Lioma, University of Copenhagen  
Adina Williams, Facebook AI Research  
Tamar Solorio, University of Houston

AFFILIATION:  
Department of Computer Science  
Faculty of Science  
University of Copenhagen

THESIS SUBMITTED:  
February 27th, 2021

## ABSTRACT

---

With recent advancements in deep learning and the infrastructure to support training models on large amounts of data, there has been an increasing emphasis on developing data-driven Natural Language Processing (NLP) systems which contain billions of parameters and optimize for language understanding benchmark datasets. While many of these systems now exceed human performance in such benchmarks, this progress has been at the expense of other desirable system qualities such as *user satisfaction, fairness and transparency*. Due to their black-box nature, the full extent of model capabilities is still not completely clear, yet, there is increasing evidence showing that systems learn undesirable and socially unacceptable patterns and can make correct predictions for the wrong reasons. These challenges make adoption of systems by users controversial, corrode user trust in the system and make it unethical to deploy systems in the wild without understanding their impact on society.

As a response to this progression in the field, the studies in this dissertation adopt an interdisciplinary *human-centered* approach for studying and improving NLP systems. This perspective emphasizes that NLP technology must be built with an understanding of humans, society and the impact it has on both.

Specifically, this dissertation investigates ways of (1) improving performance of NLP systems by leveraging user interactions and (2) ensuring fairness and transparency in NLP systems. The first part of the thesis demonstrates how to incorporate user interactions and user feedback signals that better align to human expectations in the real world, to improve the predictive performance of dialogue systems and improve their ability to adapt to new domains. As ethical concerns have emerged in recent years, the second part of this dissertation shifts focus, acknowledging the need for *better evaluation*. The incorporating knowledge from NLP, Human Computer Interaction (HCI), linguistics, and cognitive science, more meaningful evaluation protocols can be created to assess the fairness and transparency of NLP models.

## ABSTRACT IN DANISH

---

Nylige fremskridt inden for både deep learning og infrastrukturen til at træne modeller på store datamængder har muliggjort udviklingen af datadrevne Natural Language Processing (NLP)-modeller. Sådanne modeller indeholder milliarder af parametre og har forbedret benchmarks på natural language understanding-datasæt. Selvom flere automatiske modeller nu er bedre end mennesker på disse benchmarks, er landvindingerne sket på bekostning af andre kvaliteter ved de automatiske systemer såsom *brugertilfredshed, fairness og transparens*. Fordi modellerne er black boxes, er det fulde omfang af deres egenskaber ikke afdækket, men der er stigende evidens for, at systemerne lærer uønskede og socialt uacceptable mønstre og klassificerer korrekt, men ud fra forkerte grunde. Disse udfordringer ødelægger tilliden til systemerne og gør det kontroversielt og uetisk at anvende dem i praksis uden at forstå deres indvirkning på samfundet.

Som en reaktion på denne udvikling anvender studierne i denne afhandling i stadig højere grad en tværfaglig *brugercen-**tret* tilgang for at undersøge og forbedre NLP-systemer. Denne tilgang understreger at teknologien skal udvikles med en forståelse for mennesker og samfund samt teknologiens indvirkning på disse.

Denne afhandling fokuserer specifikt på måder (1) at forbedre NLP-systemer ved at udnytte brugerinteraktioner og (2) at sikre fairness og transparens i NLP-systemer. Den første del af afhandlingen demonstrerer hvordan inkorporering af brugerinteraktioner og diskret feedback kan øge funktionaliteten af dialogsystemer og deres evne til at tilpasse sig nye tekstdomæner. De seneste år er etiske overvejelser blevet mere aktuelle og anden del af afhandlingen skifter fokus idet jeg anerkender et øget behov for *bedre evaluering*. Jeg demonstrerer i denne afhandling at kombineret viden fra NLP, Human Computer Interaction, lingvistik og kognitionsvidenskab kan skabe skarpere evalueringprotokoller for at evaluere fairness og transparens i NLP-modeller.

## PUBLICATIONS

---

This is an article-based dissertation. Below is a list of the publications (and manuscripts) that are included in this thesis. The research conducted during my studies, has been a collaboration with international as well as Danish researchers in [NLP](#) and Human-centered AI.

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

**González, Ana Valeria**, Isabelle Augenstein, and Anders Søgaard (2019). "Retrieval-Based Goal-Oriented Dialogue Generation." In: *3rd NeurIPS Conversational AI Workshop: "Today's Practice and Tomorrow's Potential."*

**González, Ana Valeria**, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srini Iyer (2021). "Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations." In: *Currently under review.*

**González, Ana Valeria**, Maria Barret, Rasmus Hvingelby, Kelly Webster, and Anders Søgaard (2020). "Type B Reflexivization as an Unambiguous Testbed for Multilingual Multi-Task Gender Bias." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

**González, Ana Valeria**, Anna Rogers, and Anders Søgaard (2021). "On the Interaction of Belief Bias and Explanations." In: *Currently under review.*

**González, Ana Valeria** and Anders Søgaard (2020). "The Reverse Turing Test for Evaluating Interpretability Methods on Unknown Tasks." In: *NeurIPS Workshop on Human And Machine in-the-Loop Evaluation and Learning Strategies.*

Below are a list of papers which I have authored and co-authored during my time as a PhD student, which are **not** part of this thesis.

Abdou, Mostafa, Ana Valeria González, Mariya Toneva, Daniel Herschcovich, and Anders Søgaard (2021). “Does injecting linguistic structure into language models lead to better alignment with brain recordings?” In: *Currently under review*.

Aralikatte, Rahul, Heather Lent, Ana Valeria González, Daniel Herschcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard (2019). “Rewarding Coreference Resolvers for Being Consistent with World Knowledge.” 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)*, pp. 1229–1235.

Beloucif, Meriem, Ana Valeria González, Marcel Bollmann, and Anders Søgaard (2019). “Naive Regularizers for Low-Resource Neural Machine Translation.” In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 102–111.

González, Ana Valeria (2021). “Towards Human-Centered NLP: An Interdisciplinary Perspective.” In: *Proceedings of the 1st workshop on Bridging HCI and NLP at EACL*. Association for Computational Linguistics.

González, Ana Valeria, Isabelle Augenstein, and Anders Søgaard (Oct. 2018). “A strong baseline for question relevancy ranking.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium: Association for Computational Linguistics, pp. 4810–4815.

González, Ana Valeria, Victor Petrén Bach Hansen, Joachim Bingel, and Anders Søgaard (2019). “Coastal at semeval-2019 task 3: Affect classification in dialogue using attentive bilstms.” In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 169–174.



## ACKNOWLEDGMENTS

---

Throughout my studies, I have been extremely lucky to be surrounded by people (at work and outside) who have always made me feel supported. To *all* the members of CoAStAL, former and current, thank you for making these last three years more fun! Mostafa, Victor, Heather, Meriem, Joachim, Maria, Marcel, Mareike, and many I don't have space to mention, thanks for being great friends and colleagues.

The great working environment at CoAStAL is in large part due to my supervisor, Anders. Thanks for all the optimism and motivation you provided when I most needed it! I am also incredibly grateful to my colleagues during my time at Facebook AI. Srini, Robin, Angela, Yashar, and Gagan thanks for a really enjoyable internship despite the current unprecedented world events. I look forward to meeting you all in person in a post-pandemic time.

While writing this dissertation, I had many people read and provide extremely valuable feedback on different versions; some versions were considerably more rough than others. Thanks to Maria Barrett, David Vilares, Mostafa, Joachim, Mareike, Simon, Anders, Daniel, Mathias, and my dad (yes, my dad read and gave me very valuable feedback on the structure of my thesis!). Special shout out to Joachim and Maria for reading different parts of my thesis multiple times!

To all my friends, near and far, thank you for always cheering me on and helping me have a much needed balance in my life. Mathias, thank you for your endless support, the last year and a half has particularly been filled with amazing company and delicious food.

Last but not least, my family. My sisters, abuelas, tíos, primos; always cheering for me no matter how far apart we are. And of course **my parents**, I am infinitely grateful for all your hard work and sacrifices. I am utterly aware that me being where I am, would never have happened if you had not taken a *series* of very tough decisions throughout the years; I do not take that for granted. I am forever grateful, in a way that cannot be expressed in a paragraph (or page) of an acknowledgements section!



# CONTENTS

---

## I BACKGROUND

1	INTRODUCTION	3
1.1	Towards Human-centered NLP . . . . .	4
1.2	Human-centric Ways of Improving Human-AI interaction . . . . .	6
1.2.1	Learning from user interactions . . . . .	6
1.2.2	Fairness and transparency in the system . . . . .	8
1.3	Research Questions . . . . .	10
1.4	Thesis outline and contributions . . . . .	11
1.4.1	Part ii: Learning from user interactions . . . . .	11
1.4.2	Part iii: Investigating fairness and trans- parency in NLP systems . . . . .	12
2	BACKGROUND	15
2.1	Advances in Transfer Learning for NLP . . . . .	15
2.1.1	Sequential transfer learning . . . . .	17
2.2	Transfer learning in dialogue systems . . . . .	20
2.2.1	Goal-oriented dialogue . . . . .	21
2.2.2	Open-ended chit chat . . . . .	21
2.2.3	Initial goals . . . . .	22
2.2.4	Recent advancements in dialogue using large pretrained models . . . . .	23
2.3	Research Shift: Ethical challenges . . . . .	24
2.4	Fairness and Transparency in NLP systems . . . . .	26
2.4.1	Detection of social bias in NLP . . . . .	26
2.4.2	Explainability methods . . . . .	28
2.4.3	Evaluation of explainability . . . . .	29

## II LEARNING FROM USER INTERACTIONS

3	RETRIEVAL-BASED GOAL-ORIENTED DIALOGUE GEN- ERATION	35
3.1	Abstract . . . . .	35
3.2	Introduction . . . . .	35
3.3	Model Description . . . . .	37
3.3.1	HRED . . . . .	37
3.3.2	Exemplar-HRED . . . . .	38
3.4	Experiments . . . . .	40
3.4.1	Dataset and preprocessing . . . . .	40
3.4.2	Metrics . . . . .	40
3.5	Results and Discussion . . . . .	42
3.6	Related Work . . . . .	44
3.7	Conclusion . . . . .	45

4	DOMAIN TRANSFER WITHOUT TURN-LEVEL SUPERVISION	47
4.1	Abstract . . . . .	47
4.2	Introduction . . . . .	47
4.3	Baseline Architecture . . . . .	48
4.4	Domain Transfer Using Reinforcement Learning . . . . .	50
4.5	Experiments . . . . .	52
4.5.1	Data . . . . .	52
4.5.2	Implementation Details . . . . .	53
4.5.3	Experimental Protocol . . . . .	54
4.6	Results . . . . .	55
4.7	Analysis . . . . .	55
4.7.1	Error Analysis . . . . .	57
4.7.2	Comparisons to Weak Supervision . . . . .	57
4.8	Related Work . . . . .	58
4.9	Conclusion . . . . .	60
III INVESTIGATING FAIRNESS AND TRANSPARENCY IN NLP		
5	TYPE B REFLEXIVIZATION AS A TESTBED FOR GENDER BIAS	63
5.1	Abstract . . . . .	63
5.2	Introduction . . . . .	63
5.3	The Anti-reflexive Bias Challenge . . . . .	66
5.4	Experiments . . . . .	69
5.5	Results . . . . .	72
5.6	Analysis: Biased statistics? . . . . .	73
5.7	Related Work . . . . .	75
5.8	Conclusion . . . . .	77
6	THE REVERSE TURING TEST FOR EVALUATING INTERPRETABILITY	79
6.1	Abstract . . . . .	79
6.2	Introduction . . . . .	79
6.3	Human Bias in Forward Prediction . . . . .	81
6.4	LIME – and its Limitations . . . . .	82
6.5	Human Forward Prediction Experiments . . . . .	83
6.5.1	Tasks and Data . . . . .	84
6.5.2	Classification Model . . . . .	85
6.5.3	Stimulus Presentation . . . . .	86
6.5.4	Pre-Experiment: The Effect of Training on Forward Prediction . . . . .	87
6.5.5	Main Experiment: The Effect of Local Interpretable Model-agnostic Explanations (LIME) on Forward Prediction . . . . .	88
6.6	Related Works . . . . .	90
6.7	Conclusion . . . . .	93

7	ON THE INTERACTION OF BELIEF BIAS AND EXPLANATIONS	95
7.1	Abstract . . . . .	95
7.2	Introduction . . . . .	95
7.3	Belief Bias . . . . .	97
7.4	Related Work . . . . .	98
7.5	Experimental Setup . . . . .	100
7.5.1	Models . . . . .	100
7.5.2	Data . . . . .	101
7.5.3	Explainability Methods . . . . .	101
7.6	Experiment 1: Human Forward Prediction . . . . .	102
7.7	Experiment 2: Best Model Selection . . . . .	106
7.8	Discussion: Mitigating Belief Bias . . . . .	110
7.9	Conclusion . . . . .	111
8	AN EVALUATION OF SPOKEN VS. VISUAL EXPLANATIONS	113
8.1	Abstract . . . . .	113
8.2	Introduction . . . . .	113
8.3	Related Work . . . . .	115
8.4	Visual vs. Spoken Modalities . . . . .	117
8.5	Experimental Setup . . . . .	118
8.5.1	Explanation Types and Conditions . . . . .	118
8.5.2	Hypotheses . . . . .	119
8.5.3	Implementation Details for Conditions . . . . .	119
8.5.4	User study & Interface . . . . .	121
8.6	Results . . . . .	123
8.6.1	Quantitative Results . . . . .	124
8.6.2	Qualitative results . . . . .	126
8.6.3	What misleads users? . . . . .	128
8.7	Discussion . . . . .	130
8.7.1	Why Explanations Worked for Open-domain Question Answering (ODQA)? . . . . .	130
8.7.2	Implications and Recommendations . . . . .	131
8.8	Conclusion . . . . .	132
<b>IV DISCUSSION AND CONCLUSION</b>		
9	DISCUSSION OF THE CONTRIBUTIONS	137
10	FUTURE DIRECTIONS	141
<b>V APPENDIX</b>		
A	APPENDIX	145
A.1	Chapter 5 . . . . .	145
A.1.1	Example Data . . . . .	145
A.1.2	Coreference Dataset Statistics . . . . .	147
A.2	Chapter 6 . . . . .	147
A.3	Chapter 7 . . . . .	148
A.3.1	Experiment 1: Human Forward Prediction	148

- A.3.2 Experiment 2: Best Model Selection . . . . 149
- A.4 Chapter 8 . . . . . 151
  - A.4.1 Temperature Scaling . . . . . 151
  - A.4.2 Additional Preprocessing . . . . . 152
  - A.4.3 Task Setup: Additional details . . . . . 153
  - A.4.4 Post-task survey . . . . . 154
  - A.4.5 Results . . . . . 154
  - A.4.6 Explanation Examples . . . . . 158

BIBLIOGRAPHY 159

## LIST OF FIGURES

---

Figure 3.1	Our model is similar to HRED (Sordoni et al., 2015a), we include an utterance encoder, a context encoder and a decoder, however, unlike HRED, our model include a simple, yet effective retrieval step used to condition the decoder to generate responses that are more appropriate for a specific domain and context. . . . .	39
Figure 4.1	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 finetuning policy $\pi_\theta(s a)$ on domain $d^F$ . . . . .	49
Figure 4.2	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 finetuning using the supervised model as starting point. . . . .	56
Figure 4.3	The turn level accuracy of our weakly supervised finetuning compared to finetuning using PG. Performance plateaus after about 50 samples for both methods. . . . .	56
Figure 5.1	Correlations between collected labor statistics. Numbers $> 0.7$ are significant ( $p < 0.01$ ). . . . .	75
Figure 6.1	Our experimental protocol. For each task, we train our models using standard datasets and evaluate the model on held out training data and testing data to be used for the training and evaluation sessions involving humans. We also extract LIME explanations. In the human experiments phase, the humans train and evaluate in these 2 conditions (LIME explanation or no explanation). Finally, we compare the results. . . . .	83

Figure 6.2 Example **LIME** explanation stripped of model decisions and class probabilities. We turn the images into gray scale to only highlight overall importance and avoid hinting the model’s final decision. . . . . 86

Figure 6.3 COMPARING KNOWN AND UNKNOWN TASKS. i) Left bars show mean inference time (secs) *with* **LIME** explanations; ii) middle bars show mean inference time *without*; and iii) right bars show mean inference time across *all* tasks, with and without **LIME**. 91

Figure 7.1 Evaluation protocols considered in this work . . . . . 96

Figure 7.2 Interface for Experiment 1 for LOW condition. To select model predictions, participants clicked on tokens to select the start and end of the span. Then they would see the actual model prediction. . . . . 103

Figure 7.3 Experiment 1 UI: Low(bottom) vs HIGH(top) condition. . . . . 107

Figure 7.4 Feedback categories and their distribution. We observed that the HIGH vs MEDIUM condition results are considerably different from the HIGH vs Low condition, with more participants giving generic answers for vanilla gradients, and emphasizing the irrelevant terms highlighted in the Integrated Gradients (**IG**) condition. . . . . 109

Figure 8.1 Using end-to-end user studies, we evaluate whether explanation strategies of open-domain QA assistants help users decide when to trust (or reject) predicted answers. . . . . 114

Figure 8.2 UI for visual (left) and spoken modalities (right) for EXT-SENT explanation type. Users either read or hear an explanation and decide whether to trust or discard the Question Answering (**QA**) system’s prediction. . . . . 120



Figure 8.3 Accuracy of users at error detectability (75 workers per condition). In the *spoken modality*, EXT-SENT explanations yield the best results and is significantly better than CONF. In contrast, in the *visual modality*, EXT-LONG explanations perform best. We observe a statistically significant ( $p < 0.01$ ) difference between EXT-LONG in visual vs spoken, perhaps due to differences in user’s cognitive limitations across modalities. . . . . 123

Figure 8.4 (Left) Explanations significantly increased participant ability to detect *correct* answers compared to simply displaying confidence. (Right) However, only EXT-SENT in the spoken modality and both explanations in the visual modality decreased the rate at which users are misled. . . . . 125

Figure 8.5 **Voice clarity:** Most participants found the voice of the assistant to be good or excellent. . . . . 126

Figure 8.6 **Top:** Users perceive the same explanation to be longer in the spoken modality. **Bottom:** While EXT-SENT and ABS were the same length, participants rate the latter as longer more often perhaps because of they contain more content. . . . . 127

Figure A.1 **(a)** Example of item in the training session for sentence length prediction. Note that the participants are able to check the model answer **(b)** Example of item in the evaluation session for sentence length prediction. Here the participants are no longer able to check the model answer . . . 148

Figure A.2 Confidence before and after calibration. . . . . 151

Figure A.3 **Reward:** The scores presented here are out of \$ 2.70. Although all explanations are better than CONFIDENCE, the explanations leading to the highest rewards change across modalities. . . . . 155

Figure A.4	<b>Helpfulness:</b> Participants indicated how helpful responses were. These results reflect the large differences we see in performance ( <code>BASELINE</code> vs the rest of the settings), but are not able to capture the more subtle differences among explanation strategies and <code>CONFIDENCE</code> . . . . .	156
------------	---	-----

## LIST OF TABLES

---

Table 3.1	Statistics of the MultiWOZ training data .	40
Table 3.2	The results of our dialogue generation experiments comparing HRED Serban et al. (2016) and Sordoni et al. (2015a) to our proposed exemplar-based model. We present results for standard metrics used in dialogue generation. For all the metrics we observe improvements over the strong baseline, with our best improvement of 6 percent in the vector extrema metric . . . .	43
Table 3.3	Examples of responses generated by both the baseline and our proposed model. By examining the outputs, it becomes noticeable that the baseline model tends to generate responses that are not precise about the current domain of the conversation (hotel, taxi booking, trains, restaurant, etc).	44
Table 4.1	Statistics of the MultiWOZ dataset. The reported numbers are from our processed dataset. . . . .	51

Table 4.2	Accuracy scores for our pretrained baseline (BL) and the policy gradient finetuning (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 finetuned results, and each pair of columns compare the 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. . . . .	53
Table 4.3	Comparison of example turn predictions from the MultiWOZ dataset between the baseline model trained on the HOTEL domains, and the policy gradient finetuned model. Green indicates a correct prediction whereas red indicates a wrong prediction. . . . .	57
Table 5.1	In Type B reflexivization (Heine, 2005), 3rd person pronouns cannot be used reflexively. We are interested in Type B languages with gendered pronouns, and where the non-gendered special (3rd person) reflexive marker has a possessive form.	64
Table 5.2	Gender Bias Results. Performance on benchmarks and Anti-reflexive Bias Challenge (ABC). $\checkmark$ : Pearson’s $\rho$ of error $\Delta$ on sentences with feminine pronouns and % of women in corresponding occupations significant ( $p < 0.01$ ); see S for a discussion of the statistics. †: Systems insensitive to variation in pronouns. . . .	71

Table 6.1	RESULTS FROM MAIN EXPERIMENT. Columns 1–2: accuracy of human forward prediction results on plain input ( $x$ ) or augmented with LIME interpretations (LIME( $x$ )). *: Significance of $\alpha < .05$ computed with Mann-Whitney $U$ test. Columns 3–4: average duration of evaluation sessions (human inference time). Column 5 lists the model accuracies with respect to human gold annotation; which we compare with human accuracies with respect to human gold annotation. . . . . 89
Table 7.1	Human forward prediction results (HUMAN( $\hat{y}$ )) for LOW and HIGH models, compared to no explanations (BASELINE). Each experiment is run on vanilla SQuAD 2.0 data (ORIG) and adversarial SQuAD 2.0 data (ADV). HUMAN( $y$ ) is the dataset ground truth and an indicator of belief bias. Statistically significant results are indicated with an asterisk. Time is the average time per question. The best $\hat{y}$ results in each condition are bolded. . . . 104
Table 7.2	Both methods do well in (HIGH vs LOW). In HIGH vs MEDIUM, performance drops dramatically for IG. * = statistical significant difference ( $\rho < 0.001$ ) . . . . . 107
Table 8.1	Mechanical Turk (MTurk) worker’s bonus as a function of the correctness of ODQA model’s prediction and the user’s decision to accept or reject the predicted answer. 122
Table A.1	Example data for NLI. For NLI, we only generate entailments and neutral statements. The English translation is shown for reference only. . . . . 145
Table A.2	Example data for machine translation. . . 146
Table A.3	Example data for coreference resolution. In brackets, we have the mentions that the system could cluster as coreferent. We include an English translation only for reference. . . . . 146
Table A.4	Example data for the language modeling task . . . . . 147
Table A.5	Statistics for the coreference data used for training. . . . . 147

Table A.6	Raw scores, before removing data points on training session . . . . .	149
Table A.7	Examples of some of the feedback categorized into these classes . . . . .	150
Table A.8	Time differences across modalities. Time differences in the right column have been adjusted by removing the duration of the audio files. We observe that with additional information, users can make faster decisions than the <code>BASELINE</code> condition. .	155
Table A.9	The codes used to uncover areas of improvement from the post-experimental user feedback. . . . .	156
Table A.10	Distribution of codes across all conditions. Codes are <b>not</b> mutually exclusive. .	157
Table A.11	<b>Explanation examples:</b> Example of how system responses looked for each explanation type and baseline, for the question <i>How many seasons of Marco Polo are there?</i>	158

## ACRONYMS

---

NLP	Natural Language Processing
HCI	Human Computer Interaction
ML	Machine Learning
RNN	Recurrent Neural Network
AI	Artificial Intelligence
CBOw	Continuous Bag-of-Words
LSTM	Long Short Term Memory
ELMo	Embedding from Language Model
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformer
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
QA	Question Answering
ODQA	Open-domain Question Answering
NLI	Natural Language Inference

IR	Information Retrieval
BLEU	BiLINGUAL Evaluation Understudy
IOU	Intersection Over Union
LM	Language Modeling
MTL	Multi-Task Learning
NLM	Neural Language Modeling
NER	Named Entity Recognition
DSTC	Dialogue State Tracking Challenge
NLU	Natural Language Understanding
DST	Dialogue State Tracking
NLG	Natural Language Generation
RL	Reinforcement Learning
LRP	Layer-wise Relevance Propagation
IG	Integrated Gradients
LIME	Local Interpretable Model-agnostic Explanations
QLIME	Quadratic LIME
SLIME	Sound-LIME
SHAP	SHapley Additive exPlanations
MT	Machine Translation
HRED	Hierarchical Recurrent Encoder-Decoder
GRU	Gated Recurrent Unit
ANNS	Approximate Nearest Neighbor Search
NNS	Nearest Neighbor Search
GloVe	Global Vectors for word representation
SQuAD	Stanford Question Answering Dataset
MAP	Mean Average Precision
AUTPC	Area Under the Threshold-Precision Curve
AOPC	Area Over the Perturbation Curve
AUPRC	Area Under the Precision-Recall Curve
GLUE	General Language Understanding Evaluation
SQuAD	Stanford Question Answering Dataset
HSD	Honest Significant Difference
RC	Reading Comprehension
NL	Natural Language
NQ	Natural Questions
DPR	Dense Passage Retrieval

MTurk	Mechanical Turk
MT	Machine Translation
ABC	Anti-reflexive Bias Challenge
XNLI	Crosslingual NLI
SST	Stanford Sentiment Treebank





Part I

BACKGROUND



## INTRODUCTION

---

The fields of Human Computer Interaction ([HCI](#)) and Artificial Intelligence ([AI](#)) emerged at different points in the history of computer science and with seemingly different goals. For instance, the core idea of [AI](#)—that human cognitive processes can be mechanized—has a long history dating back to the 12th-17th centuries (and even earlier) with philosophers who speculated that human reasoning can be reduced to mechanical calculation (Carreras and Carreras, 1939; McCorduck, 2004). The modern history of [AI](#) as a field in computer science studying intelligent agents exploded in the 1950s following many historic events including the emergence of the first modern computers (Goldstine and Goldstine, 1946), the landmark paper devising the famous Turing Test (Turing, 1950), and the Dartmouth conference where the term Artificial Intelligence was coined by John McCarthy and Marvin Minsky.

In contrast to the long history of that term, the field of [HCI](#), which studies the design and use of computing systems by *human users*, is very young. As the early computers were only available to computer scientists, engineers, or people who had a particular interest in such technologies, research in computer usability was not as widespread. The birth of [HCI](#) came with the era of personal computing in the 1980s<sup>1</sup> (MacKenzie, 2012), with key events such as the first SIGCHI conference, the publication of the seminal book *The Psychology of Human-Computer Interaction* (Card, 1983) coining the term [HCI](#) and the arrival of the Apple Macintosh, the first successful mass-market personal computer. In the years to come, [HCI](#) began to expand rapidly as an interdisciplinary research area which included the fields of computer science, cognitive science, psychology and human factors engineering, among many others.

Despite their different roots, today, [HCI](#) and [AI](#) are interacting and converging more than ever before. While for many years [AI](#) systems that were part of everyday life were mostly limited to science fiction, today the reach of [AI](#) has expanded dramatically with applications in transportation, finance, healthcare and commerce. As a result, in the last 30 years researchers in [HCI](#) have provided valuable insights for *human-centered* research (Amershi et al., 2019a; Bannon, 2011; Höök, 2000; Horvitz, 1999;

---

<sup>1</sup> Although important advances such as the computer mouse go back to the 1960s: see MacKenzie (2012) for additional historical context.

Lee and See, 2004; Norman, 1994; Wickramasinghe et al., 2020) which can improve human-AI interaction. A cross-pollination of AI, HCI and other fields, combined with the rapid advancements and impact of AI, has resulted in *human-centered* AI; a general perspective to building AI technology, which emphasizes that **intelligent systems must be designed with awareness of the larger ecosystem they are a part of and the humans who are in contact with or are affected by the technology** (Fiebrink and Gillies, 2018; Riedl, 2019). This dissertation, which comprises work done against the backdrop of the rapid development of intelligent systems, makes strides towards adopting an interdisciplinary *human-centered* approach for studying NLP systems and improving human-AI interaction. The next section further expands on the *human-centered* perspective and why there is a need for adopting such framework in NLP.

### 1.1 TOWARDS HUMAN-CENTERED NLP

The *human-centered* perspective prioritizes the creation of technology which aims to enhance and augment human abilities, rather than replace them (Auernhammer, 2020; Fiebrink and Gillies, 2018; Shneiderman, 2020; Xu, 2019) and emphasizes technology which is built with an understanding of humans, society and the impact technology has on both<sup>2</sup>. In this thesis, I argue that taking a human-centered approach to NLP requires interdisciplinary efforts, with NLP practitioners taking strides to understand humans from a cognitive and social perspective, incorporating insights from psychology, HCI, NLP, ethics and many other disciplines, and introducing more humans at different stages of the development process. The studies presented in this dissertation increasingly adopt this framework over a period of three years.

Some possible directions within this framework include: (1) studying the usability and usefulness (Hornbæk and Oulasvirta, 2017; Rasmussen, Pejtersen, and Goodstein, 1994; Rouse, 1986; Rouse and Rouse, 1991) of NLP systems by adopting ethnographic methods and user studies typically used in HCI to inform us of how to further improve NLP systems in a way that matches human social and cognitive expectations, (2) adopting insights from both AI and HCI for improving the predictive power of systems while offering more human control (e.g., human-in-the-loop and collaborative systems) and (3) adopting

<sup>2</sup> See the goals outlined by the recently created Human-centered AI Research center at Stanford University: <https://hai.stanford.edu/blog/introducing-stanfords-human-centered-ai-initiative>

techniques from the social sciences for the ongoing study of the impact that AI has on humans and society.

**But why is it important to adopt a human-centered approach to studying NLP systems?** Well, with advancements in deep learning and the infrastructure to support training models on large amounts of data, there has been an emphasis on larger models containing billions of parameters which optimize for benchmarks such as General Language Understanding Evaluation (GLUE)<sup>3</sup> (Wang et al., 2018). This focus has made it faster to evaluate systems and has paved the way to success for data-driven approaches. Such approaches may exceed human performance on such benchmarks, however, this has been at the expense of other desirable system qualities, e.g. *user satisfaction, fairness and transparency* (Ethayarajh and Jurafsky, 2020).

The complexity of the learned representations of such models makes it increasingly difficult to understand their inner workings. Many studies continue to investigate their learning patterns (Clark et al., 2019; Karthikeyan et al., 2019; Rogers, Kovaleva, and Rumshisky, 2021) with the full extent of model capabilities still not completely clear. However, clear evidence exists showing that these models encode many undesirable and discriminatory patterns (Caliskan, Bryson, and Narayanan, 2017; Khosla et al., 2012; Manzini et al., 2019; Tan and Celis, 2019) and can provide correct answers for the wrong reasons (McCoy, Pavlick, and Linzen, 2019). These challenges make the adoption of systems by users controversial, corrode user trust in the system, and make it simply *unethical* to deploy systems in the wild without understanding their impact on society.

While research in *human-centered* areas such as fairness and model transparency (Amershi et al., 2019b; Riedl, 2019) has increased in NLP in recent years, there is still a disconnect (which I will highlight through this thesis) between the methods being developed and the humans who interact with them. I argue that there is a greater need to consider human needs and capabilities through interdisciplinary *human-centered* research.

This dissertation follows the rapid progression in the field of NLP. My earlier PhD work focuses on improving human-AI interaction via goal-oriented dialogue systems, primarily concerned with improving predictive performance. Goal-oriented dialogue systems are naturally user-centric; they are meant to help *users* achieve a goal through *natural language interaction*. The first two studies presented in this thesis deal with such systems and incorporate transfer learning methods which are

---

<sup>3</sup> Benchmark dataset comprising different tasks such as semantic similarity and Natural Language Inference (NLI), meant to assess whether a model understands language

now ubiquitous in NLP. However, the rapid advancements in the field have also come with ethical challenges. Therefore, my later work starts to adopt a more interdisciplinary human-centered approach to studying NLP systems with a focus on fairness and model transparency and increasingly placing more emphasis on understanding the *human* aspect of human-AI interaction. The dimensions explored in this thesis are expanded in the next section.

## 1.2 HUMAN-CENTRIC WAYS OF IMPROVING HUMAN-AI INTERACTION

The topics presented in this dissertation all focus on improving human-AI interaction along two important dimensions: (1) improving the performance of interactive NLP systems by learning from user interactions and (2) improving interaction by ensuring fairness and transparency in NLP systems. These dimensions also fall within 18 *human-centric* guidelines for human-AI interaction compiled by Amershi et al. (2019a), which have been identified by researchers in the HCI community in the last 30 years. This section describes these dimensions in more detail and ties back to the guidelines outlined by Amershi et al. (2019a).

### 1.2.1 *Learning from user interactions*

Systems which allow a certain level of user control and can learn from user interactions over time can result in better user experience, satisfaction, user trust and more effective systems (Amershi et al., 2014). Amershi et al. (2019a) additionally mention some ways to use user signals, such as (1) remembering recent user interactions and (2) encouraging and learning from user feedback, among others. The earlier work in this thesis is predominantly focused on improving the predictive performance of dialogue systems along these dimensions.

**REMEMBER RECENT INTERACTIONS** Remembering recent user interactions to improve user experience and future interactions has many commercial applications such as music, movie or product recommendations (Amershi et al., 2019a; Webb, Pazzani, and Billsus, 2001; Yang, 2017). NLP systems can also improve performance and provide better user interaction and experience if they leverage previous user exchanges. In dialogue systems, which are the subject of the early work in this thesis, previous behaviors can serve as a prior for improving desirable qualities such as relevancy, coherence, and fluency of natural language responses, among many others.

However, systems learning from user behavior without control pose risks. In recent years, controversial applications such as Tay, Microsoft's open-ended chatbot, have exposed how NLP systems can learn negative behaviors from users if there is no control on what is being learned. In this case, Tay was vulnerable to adversarial attacks and quickly learned racist behaviors from human interactions <sup>4</sup>. In their guidelines, Amershi et al. (2019a) also mention that updating and learning from such behaviors should be done cautiously; part of the human-centered approach is to ensure responsible deployment and socially appropriate behavior. It is therefore important to leverage ways of learning from *successful and safe* previous interactions. Unlike open-ended dialogue systems, goal-oriented dialogue systems tend to rely on *annotated* dialogues, therefore, it may be a good and safe candidate for leveraging past user interactions as they must be annotated and checked by humans. Chapter 3 explores a method for improving goal-oriented dialogue generation by retrieving previous user interactions.

**LEARNING FROM USER FEEDBACK** Research in Machine Learning (ML) has shown that models benefit from incorporating user feedback in tasks such as image retrieval and recommender systems (Rashid et al., 2002; Vasconcelos and Lippman, 1999). Such signals have typically portrayed the human as an oracle who provides ground-truth information at any point as is needed by the system.

Work in HCI has found that feedback strategies should match the human expectations. Cakmak, Chao, and Thomaz (2010) found that during dialogues, a robot that asks for too much feedback is perceived by users as imbalanced and annoying. Guillory and Bilmes (2011) replicate this finding for movie recommender systems. Such results show that human users are not oracles willing to repeatedly tell the system whether it is wrong or right (Cakmak, Chao, and Thomaz, 2010). Amershi et al. (2014) suggest that ML practitioners should make efforts to account for human factors such as interruptibility and frustration when employing strategies that rely on user feedback.

In this thesis, (modeled) user feedback is leveraged to improve dialogue systems. Previous work in ML and NLP has presented Reinforcement Learning (RL) methods that model user feedback throughout the course of a dialogue or that optimize for dialogue length, which has shown to improve model robustness (Henderson, Lemon, and Georgila, 2008; Liu et al., 2017; Williams, Asadi, and Zweig, 2017; Williams and Zweig, 2016). However, many of such methods relied on annotations at ev-

---

<sup>4</sup> [https://en.wikipedia.org/wiki/Tay\\_\(bot\)](https://en.wikipedia.org/wiki/Tay_(bot))

ery turn or proxy measures which are not applicable in the wild: asking for too much feedback may erode user satisfaction and optimizing for dialogue length is not a realistic measure of success. Chapter 4 explores a method for modeling end-of-dialogue user feedback, as opposed to feedback at every turn, with the goal of leveraging less costly user signals which match real-world human expectations while remaining robust and generalizable.

### 1.2.2 Fairness and transparency in the system

As previously mentioned, with the emergence of large data-driven pretrained systems and the widespread use of NLP technologies, it has become increasingly important to study the risks of such systems, offer more control to humans and involve humans in the development and evaluation processes. This realization marks a shift in the research presented in this dissertation towards investigating fairness and transparency of models with an emphasis on *evaluation* and *user studies*.

This focus is in line with the priorities outlined by researchers advocating for human-centered AI. Riedl (2019), for example, argues that human-centered AI research, considers two broad aspects: (1) AI systems that understand humans from a socio-cultural perspective and (2) AI systems that help humans understand them. Additionally, Amershi et al. (2019a) mention that successful human-AI interaction requires (1) mitigation of social biases and ensuring socially appropriate behaviors, and (2) providing mechanisms for explaining model decisions to end-users. These are introduced briefly below.

**MITIGATION OF SOCIAL BIASES** As NLP systems become more widespread, studying their effects on society and trying to minimize risky behaviors is crucial. Such risky behaviors can be, among other things, the propagation of implicit social biases.

*Implicit bias* (as defined in psychology) refers to attitudes or stereotypes that affect our understanding and decisions in an unconscious manner (Greenwald and Krieger, 2006; Kelly and Roedder, 2008). Such biases can occur based on many characteristics such as age, race, ethnicity, gender, etc. Implicit biases create barriers that affect marginalized groups. Such barriers are harder to point out and dismantle than explicit biases. Many social inequities (e.g., housing, education, health, and criminal justice) can be tied back to implicit biases manifested structurally (Kelly and Roedder, 2008).

While this is a phenomenon observed in society, data-driven methods learning from *human-generated sources* also have such



biases. Models exhibiting social biases can also create barriers and exacerbate inequities for already marginalized groups. Some examples which have received attention in the media are (1) the Amazon hiring tool which disproportionately disregarded applications from women in already male dominated fields and (2) police profiling, where individuals of a certain race or ethnic background were targeted by the system more often due to historical data reflecting expressed social biases<sup>5</sup>. Detecting biases in models is of paramount importance for ensuring user trust in the system and for providing overall fair, inclusive and ethically compliant technologies (Olhede and Wolfe, 2018).

In NLP, work to ensure that systems' language and behavior does not reinforce negative and unfair stereotypes can include (but isn't limited to) the detection of biases in trained models and debiasing at the data and algorithmic levels. The work presented in this thesis (chapter 5) focuses on the diagnosis of biases in large data-driven models which are popular and often deployed in the wild, such as Google Translate.

**PROVIDING SYSTEM TRANSPARENCY** Lastly, models that are able to explain their decisions can allow for better detection of unfair model behaviors, the prevention of adversarial attacks, and improved human decision making (Amershi et al., 2019a). Human curiosity and the need to assess the reliability of a claim when making decisions (among other factors), also leads users to wonder why a recommendation was given (Miller, 2019). This is especially crucial in critical domains where poorly made recommendations to a query can have significant negative consequences on humans and society (e.g. in domains such as law and health care).

Recently, work has emerged within the field of explainable AI, in both AI and HCI (Camburu et al., 2018; Goebel et al., 2018; Holzinger et al., 2017; Narang et al., 2020a). However, the focus in each field has been different. While the ML community has typically been concerned with developing explainable models and post-hoc explainability methods which have mathematical rigor (Ribeiro, Singh, and Guestrin, 2016a; Sundararajan, Taly, and Yan, 2017), the HCI community has focused on end-user's trust and understanding of the explanations. In HCI, the criteria for what makes a good explanation has been presented in terms of human interaction with explanations (Hoffman et al., 2018), for instance: (1) whether *users* are satisfied with the explanations, (2) whether *users* are able to understand the system's reasoning, (3) whether the *users'* trust and reliance on the system is appro-

<sup>5</sup> <https://www.cbsnews.com/news/artificial-intelligence-racial-profiling-2-0-cbsn-originals-https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

priate and (4) whether the human-AI team performs well. Such human-centered criteria have not been readily adopted in NLP.

While many studies in ML and NLP have introduced automatic metrics for measuring the quality of *explanations* of model predictions (Atanasova et al., 2020a; DeYoung et al., 2020; Goh et al., 2020), explainability is aimed at making model decisions understandable to *human* end-users. The work in this thesis argues that NLP practitioners should take greater strides towards understanding humans from a cognitive and social perspective and incorporating them in the evaluation and development of explainability methods. The later studies presented in this dissertation (chapters 6-8) adopt insights from HCI, psychology and cognitive science in order to *evaluate the effectiveness of explainability by involving humans in meaningful ways*.

### 1.3 RESEARCH QUESTIONS

This dissertation presents novel work aimed at improving NLP and human-AI interaction, by expanding knowledge in the areas discussed above. The main questions answered in this thesis are presented below separated into the two main dimensions discussed earlier:

#### LEARNING FROM USER INTERACTIONS

- How can goal-oriented dialogue systems leverage *previous user interactions* to improve the relevancy and fluency of answers? (chapter 3)
- How can more realistic *dialogue-level user feedback* help dialogue systems further improve and adapt to new domains? (chapter 4)

#### INVESTIGATING FAIRNESS AND TRANSPARENCY IN NLP

- How can we diagnose negative *social biases* in multilingual systems, which currently make the *adoption of systems by end-users* difficult? (chapter 5)
- To what extent do *humans' cognitive biases* and previous world knowledge affect the explainability of NLP systems and does controlling for such biases change the conclusions we make about the best performing methods? (chapters 6 and 7)
- Which natural language explanations *help users in a real world downstream decision making task* such as deciding when to trust a model prediction and does the effectiveness of

explanations depend on presentation modality (e.g., voice assistants vs. visual displays)? (chapter 8)

- What explanation strategies would *users prefer* to be presented with? (chapter 8)

#### 1.4 THESIS OUTLINE AND CONTRIBUTIONS

In summary, the contribution of this dissertation is the study of NLP and human-AI interaction through a *human-centered* perspective which focuses on: (1) learning from user interactions and (2) investigating fairness and transparency in NLP systems through more interdisciplinary research.

Chapter 2 provides more background on the work which has influenced the studies in this dissertation, part ii presents work from the first half of my PhD, dealing with learning from user interactions, part iii comprises my later work on fairness and explainability and chapters 9 and 10 provide a discussion of the contributions of this thesis and concluding remarks. Below, I expand on the main parts (ii and iii) and describe how the individual studies in each part contribute to the field of NLP.

##### 1.4.1 Part ii: Learning from user interactions

The studies in part ii are the first in my PhD and investigate how to incorporate previous user interactions and more realistic user feedback signals for improving the relevancy of system responses and for adapting to new domains.

- Chapter 3 introduces a method for improving the relevancy of responses of a dialogue system by leveraging *previous user interactions*. In practice, goal-oriented dialogue systems provide template-based responses which are relevant but limited. Neural models allow more flexibility in responses, but are typically unfocused, yielding fluent-yet-irrelevant responses. This chapter bridges the gap by conditioning a neural model on responses to similar previous user questions and shows that responses perform better on automatic metrics but more importantly, are rated as more relevant and informative by *human* raters.
- Methods which require supervision at every turn or rely on proxy rewards are not realistic or desirable in practice since asking users for feedback after every utterance may corrode user satisfaction and system effectiveness. For this reason, chapter 4 introduces a method for modeling end-of-dialogue feedback which is more in line with human

expectations in real-world scenarios and can be used to finetune a dialogue system to new domains.

#### 1.4.2 *Part iii: Investigating fairness and transparency in NLP systems*

Part [iii](#) shifts focus to the investigation of fairness and transparency in [NLP](#). As mentioned earlier, popular models have been found to exhibit negative social biases and prejudice. Deploying such models can have extremely negative consequences in society; it is crucial to devise mechanisms for model diagnosis. For this reason, [chapter 5](#) investigates gender bias in multilingual state-of-the-art [NLP](#) systems.

- [Chapter 5](#) presents the first challenge dataset for the diagnosis of gender bias not in English (Danish, Swedish, Russian, and Chinese) and, unlike prior work, comprises four [NLP](#) tasks. This study shows the importance of looking at linguistic phenomena which do not exist in English to help us devise better evaluations. In addition, an evaluation of commonly used multilingual models is presented, which shows that these models encode gender stereotypes in the four languages that are part of the evaluation, pointing to the importance of diagnosis and mitigation before deployment.

Due to controversial findings about models' discriminatory behaviors such as the ones previously mentioned, as well as the black-box nature of the current models, there has been an increased interest in equipping models with mechanisms to better detect such weaknesses and to increase transparency. This has resulted in explainable [AI](#). The last three studies presented in this dissertation involve *human evaluation* of explainability. Proper human evaluation requires taking insights from disciplines other than [NLP](#). In this dissertation, the fields of cognitive science, learning psychology and [HCI](#) greatly influence the approaches taken for the evaluation of explainability. Below, three chapters expanding in this direction and their contributions are described:

- [Chapter 6](#) presents pilot experiments investigating the interaction of cognitive biases such as belief bias, with human evaluation of explainability. This work poses an important question which has not been addressed in the past: *do explainability methods work when we reduce participant's previous beliefs about the task?* Answering this question would allow us to better understand to what extent explanations offer benefits. This study evaluates explainability

for several classification tasks and shows that the positive effects of explanations are reduced when users do not have prior knowledge of the task.

- In chapter 7, the work from the previous chapter is extended. An overview of belief bias and its interaction with human evaluation is presented with the aim of helping NLP practitioners improve evaluation protocols. The role of belief bias is highlighted for two paradigms previously used to evaluate explainability in NLP. The study demonstrates that when introducing conditions which account for participants' previous beliefs, some methods are considerably less effective and the conclusions made about the best performing methods change.

Finally, despite the surge of explainable NLP, there has been a lack of end-to-end evaluation involving users. Recent work has also cast doubt on the overall effectiveness of explanations in helping users assess the reliability of model decisions, finding that for some tasks such as sentiment analysis and answering LSAT<sup>6</sup> questions, explanations work just about the same as showing model confidence (Bansal et al., 2020). Such studies echo some of the conclusions from previous chapters, namely, that *the effectiveness of explanations should not be taken for granted and better human evaluation protocols are needed.*

- Chapter 8 investigates whether natural language explanation strategies help users assess the reliability of model predictions in Open-domain Question Answering (ODQA), against calibrated model confidence. Additionally, as ODQA systems are mostly used through visual displays and spoken interfaces, this study is the first to investigate *differences across modalities*. The results show that some strategies are more effective than simply showing model confidence, however, most explanation methods evaluated still significantly mislead users into accepting incorrect model predictions, showing there is plenty of room for improvement. Furthermore, the effectiveness of explanations does change with mode of presentation, largely due to differing cognitive limitations imposed on users in each modality. This chapter presents extensive analysis on user errors, user needs and provides valuable recommendations for developing better explanations and evaluations which are of interest to both the NLP and HCI communities.

---

<sup>6</sup> Law School Admission Test, required in the U.S.



## BACKGROUND

---

This chapter presents the background needed to understand the flow of this dissertation. As mentioned in the last chapter, the work in this thesis has followed a progression in the field of [NLP](#). My initial work deals with improving the predictive performance of dialogue systems, leveraging transfer learning methods and human interaction signals. My later work is influenced by the ethical challenges emerging in recent years and places emphasis on the evaluation of state-of-the-art systems (systems which also leverage transfer learning). As transfer learning has influenced both early and later work in this thesis, a large portion of this background section is dedicated to outlining the rapid advancements in transfer learning in [NLP](#).

This chapter is divided into four sections covering the following topics: (1) advances in transfer learning in [NLP](#), (2) dialogue systems (the initial research focus) and how transfer learning techniques are incorporated, (3) ethical challenges that emerge from the current state-of-the-art models leading to a shift in research direction, and (4) recent work in fairness and transparency in [NLP](#) which play a role in the later work in this thesis.

This outline is meant to provide the motivation behind both my earlier and later work, and to connect the two parts of my thesis in the larger context of the field of [NLP](#). More specific background to further motivate the studies in this dissertation can be found in the individual chapters.

### 2.1 ADVANCES IN TRANSFER LEARNING FOR NLP

Transfer of learning is an integral part of the learning process in humans; it involves transferring skills and knowledge from one situation to another (Judd, 1908; Sousa, 2002; Thorndyke and Woodworth, 1901). Transfer learning in [ML](#) assumes that models, like humans, can benefit from sharing knowledge across learning problems.

In the traditional supervised learning scenario in [ML](#), a sufficient amount of labeled data is required to train a model to solve one problem. When a model is needed for solving a new problem, new data needs to be annotated and the model is trained from scratch. Such setups are limited; they work well under the assumption that annotating large collections of data is feasible and that the train and test data are drawn from the



same distribution (Pan and Yang, 2010). These assumptions are often broken in the real world. Transfer learning is useful as it allows us to train models that better generalize across data distributions and that reduce the need and effort to collect large amounts of training data. Two important concepts in transfer learning (*domain* and *task*) are briefly defined next. The terms and notation introduced follow Ruder (2019).

**DOMAIN** A *domain*  $\mathcal{D}$  consists of a feature space  $\mathcal{X}$  and a marginal probability distribution  $P(X)$ , where  $X = \{x_1, \dots, x_n\} \in \mathcal{X}$ . For the binary classification of documents using bag-of-words features,  $\mathcal{X}$  would correspond to the space of all document representations,  $x_i$  is the  $i$ -th term vector corresponding to some document and  $X$  is a specific learning sample.

In relation to this thesis, for the problem of QA and dialogue systems (part of the early and later work of this dissertation), different domains could correspond to specific areas that the system has expertise on. For instance, the system can only answer questions about local restaurants, or local attractions, etc.

**TASK** A *task*  $\mathcal{T}$  consists of a label space  $\mathcal{Y}$ , a prior distribution  $P(Y)$  and a conditional probability distribution  $P(Y|X)$  which is learned from the training data consisting of pairs  $\{x_i, y_i\}$  where  $x_i \in \mathcal{X}$  and  $y_i \in \mathcal{Y}$ . For the binary classification of documents,  $\mathcal{Y}$  is the set of all labels e.g.  $\{0, 1\}$  and  $y_i$  is either 0 or 1.

In the context of this thesis, the tasks which are explored include Dialogue State Tracking (DST), Open-domain Question Answering (ODQA), NLI, sentiment analysis, among others.

Given a *source* domain  $\mathcal{D}_S$ , a corresponding source task  $\mathcal{T}_S$ , and a *target* domain  $\mathcal{D}_T$  with its corresponding target task  $\mathcal{T}_T$ , transfer learning aims to learn the target conditional probability distribution  $P_T(Y_T|X_T)$  by transferring information from  $\mathcal{D}_S$  and  $\mathcal{T}_S$ , where  $\mathcal{D}_S \neq \mathcal{D}_T$  and  $\mathcal{T}_S \neq \mathcal{T}_T$ .

Recent work has placed several transfer learning techniques which are used in NLP under a taxonomy which includes: (1) transductive transfer learning consisting of domain adaptation and cross-lingual learning and (2) Inductive transfer learning, consisting of Multi-Task Learning (MTL) and sequential transfer learning (Ruder, 2019). In terms of this thesis, one of the most influential types of transfer learning techniques is *sequential transfer learning*. The early studies in this thesis employ such methods to improve the performance of models and the later studies focus on the evaluation of models based on sequential transfer learning.

This section is meant to describe seminal work in sequential transfer learning but also highlight how the training objectives



have remained simple and have not changed dramatically. On the other hand, models have progressively increased in terms of the amount of training data used, the number of layers, and parameters. This point becomes increasingly important for the later work discussed in this dissertation.

### 2.1.1 *Sequential transfer learning*

In recent years, *sequential transfer learning* has become the most frequently used method for transfer learning in NLP. This method is typically used when adaptation to many target tasks is necessary, when the source task contains much more data than the target task, or when data for different tasks is not available at the same time (to allow training jointly) (Ruder, 2019). Generally, the method consists of two phases: (1) *pretraining* and (2) *adaptation*. During pretraining, the model is trained on the source task  $\mathcal{T}_S$ , and in the adaptation phase the knowledge from the trained model is transferred to  $\mathcal{T}_T$ .

The main benefit of pretraining is that it **reduces the need for annotated data in the adaptation phase**. Pretraining can be achieved in many ways, for example through supervised, unsupervised, or multi-task learning techniques (Ruder, 2019). Described next are: (1) Unsupervised *pretraining* in NLP (up to 2018) (some methods which are incorporated in part ii are described here), (2) a progression into larger Transformer-based (Vaswani et al., 2017) pretraining methods (some which are evaluated in part iii) and (3) a brief section on the *adaptation* phase.

#### 2.1.1.1 *Unsupervised pretraining: up to 2018*

Bengio et al. (2003) introduced Neural Language Modeling (NLM) for learning *distributed representations* for words or *word embeddings* which are based on distributional semantics, a research area which aims to quantify semantic similarities of linguistic items based on their distributional properties in large corpora. This NLM trained on about **800,000 words** used a cross-entropy criterion which maximized the probability of the next word given the previous words.<sup>1</sup> Word embeddings are now standard in NLP for adapting to downstream tasks, with variants of the Language Modeling (LM) objective becoming one of the most used for unsupervised pretraining.

---

<sup>1</sup> A language model is a statistical model of language, which aims to learn the joint probability function of sequences of words in a language. Such models have been useful in various natural language applications, such as speech recognition, translation, grammatical error correction and Information Retrieval (IR) (Bengio et al., 2003; Qadar and Mago, 2020)

Computational cost was a problem faced by Bengio et al. (2003). Collobert et al. (2011) focused on a method for pretraining word embeddings on a large dataset of more than **800 million tokens** by using a *pairwise ranking* task which was more efficient than the LM objective. In pairwise ranking, a higher score is assigned to a correct or probable word sequence than for an incorrect one. The intermediate learned representations encoded important semantic and syntactic relations and were effectively used as features for several NLP tasks, pushing state-of-the-art. This work was essentially one of the first to show the usefulness of pretraining on a *general* language task on a large corpus and adapting representations to many downstream tasks.

Soon after, Mikolov et al. (2013) showed how to effectively obtain distributed representations by training a simple neural network architecture on a **1.6 billion word corpus**, with the task of predicting a word based on its context before and after (Continuous Bag-of-Words (CBOW)) as well as predicting surrounding words based on a center word (Skip-gram). Pennington, Socher, and Manning (2014) then introduced Global Vectors for word representation (GloVe), which are trained on over **55 billion tokens** obtained from various sources. This method relied on word co-occurrences and matrix factorization, and proved to be another effective unsupervised method for yielding pretrained language representations which captures global statistics, rather than information about the local context.

More recently, Peters et al. (2018) presented *deep contextualized word representations* derived from a bidirectional Long Short Term Memory (LSTM) network trained with a coupled LM objective on a large corpus. These embeddings also called Embedding from Language Model (ELMo), were able to better capture word sense information and pushed state-of-the-art in many tasks including NLI, coreference resolution, Named Entity Recognition (NER), among others. Models of varying sizes were introduced, with the largest model trained on **5.5 billion tokens and containing 94 million parameters**.

In the last few years, very deep pretrained models have been introduced using *Transformer* architectures (Vaswani et al., 2017). Transformers, like a Recurrent Neural Network (RNN), are designed to handle sequential data. However, unlike RNNs, they do not require that data be processed in order. While most previous works typically trained networks consisting of a couple of layers, Transformer models are now consisting of 24 or more Transformer blocks in most cases. The improved parallelization of such models, coupled with an increasing amount of unlabeled text which is openly available on the web and the hardware

to support training huge models, has made Transformers the standard architecture in NLP.

#### 2.1.1.2 Transformer-based pretraining: 2018 until now

Generative Pre-trained Transformer (GPT) is a large language model consisting of 12 transformer blocks (Radford et al., 2018). It is trained to predict the next word given all previous words within some context window. At the time of its release it achieved state-of-the-art results on 9 out of 12 tasks it was finetuned on. GPT-2 (Radford et al., 2019), its successor, is trained with the same LM objective. However, GPT-2 learns from 8 million web pages, with the final model containing **1.5 billion parameters**.

Bidirectional Encoder Representations from Transformers (BERT) is a model pretrained on two objectives jointly: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) (Devlin et al., 2019a). For MLM, a random subset of tokens in the input sequence are replaced with a [MASK] token. The MLM objective is a cross-entropy loss on predicting the masked tokens. NSP is a binary classification loss for predicting whether two segments appear after each other in the original text. At the time of its release, BERT yielded state-of-the-art results on 11 NLP tasks it was finetuned on including GLUE benchmarks (Wang et al., 2018), Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), and SQuAD-2.0 (Rajpurkar, Jia, and Liang, 2018). The large model<sup>2</sup> is trained on more than 3 billion words and consists of 24 blocks, 16 attentions heads and **340 million parameters**. Liu et al. (2019b) replicated the study from Devlin et al. (2019a) but trained the model for a longer time, with 10 times more data, larger batches, and longer sequences. The resulting model (ROBERTA) further pushed state-of-the-art and increased the parameter size by about 50 million.

ELECTRA is a model pretrained on a discriminative task called *replaced token detection*, in which the model learns to distinguish real input tokens from plausible but synthetically generated replacements (Clark et al., 2020). The large model is trained on **33 billion tokens** and contains **340 million parameters**. At the time of its release, ELECTRA achieved state-of-the-art results on several NLP tasks, beating the models previously mentioned.

More recently, Fedus, Zoph, and Shazeer (2021) pretrained an even larger architecture on more data, using the MLM objective and a sparse training technique. The resulting model is high performing and the largest yet, containing a **trillion parameters**.

All methods introduced in this section focus on pretraining a model on a generic language task on vast amounts of data

---

<sup>2</sup> Models of various sizes are typically trained.

to yield *universal language representations*, which can be easily adapted to *any* downstream task. Such universal language representations have shown to encode notions of syntax (Goldberg, 2019; Williams, Drozdov\*, and Bowman, 2018) and semantic relations (Liu et al., 2019a).

The trends show that models will continue to be scaled; we have not reached a plateau yet, and bigger models that learn from more data and contain more parameters will likely continue to push state-of-the-art in the next few years. Within the last 2 years, the list of Transformer-based pretrained models continues to grow, with about 40 easily available in the Huggingface Transformers package<sup>3</sup>, and new studies introducing them often.

### 2.1.1.3 *Adapting to various NLP tasks*

The focus on pretraining high-quality representations has reduced the need for both large amounts of data in the target task and the number of parameters in the adaptation phase. Adaptation is usually done in two main ways: (1) through feature extraction and (2) through finetuning (Ruder et al., 2019). Feature extraction can include, for instance, extracting pretrained language representations from the learned model and using them as features in a downstream model (Asghar et al., 2014; Lopez and Kalita, 2017; Mou et al., 2016).

Finetuning typically involves using the pretrained weights as initialization for parameters in the downstream model. The pretrained architecture is then further trained or *finetuned* on the downstream task (Chronopoulou, Baziotis, and Potamianos, 2019; Devlin et al., 2019b; Houlisby et al., 2019; Howard and Ruder, 2018). Such works investigating adaptation to new tasks focus on optimization schemes involving what weights to update, when to update them, and how.

## 2.2 TRANSFER LEARNING IN DIALOGUE SYSTEMS

While work on dialogue systems goes way back (Weizenbaum, 1966), recently, this has become a more active area due to advances in computation and data-driven statistical methods which have proved beneficial in plenty of NLP tasks.

Nowadays, humans rely on systems such as Siri, Google Assistant, and Amazon Alexa, which offer quick access to digital data via search and natural language interactions. The studies presented in the first part of this thesis investigate ways of improving dialogue systems by learning from previous user

<sup>3</sup> <https://github.com/huggingface/transformers>

interactions and feedback signals that align more with the real world and user expectations. Transfer learning techniques are incorporated for this purpose, including adaptation via both feature extraction and finetuning.

Dialogue research is typically divided into two main categories: 1) goal-oriented and (2) open-ended. The works presented in this thesis deal with the former but also integrate approaches from the latter. Both are briefly introduced below.

### 2.2.1 *Goal-oriented dialogue*

Goal-oriented dialogue systems are designed with the purpose of assisting users in achieving a goal in restricted domains (Chen et al., 2017b). In practice, such systems are modular, typically consisting of Natural Language Understanding (NLU), dialogue management and Natural Language Generation (NLG) modules which are independently trained (Budzianowski and Vulić, 2019; Chen et al., 2017b; Hosseini-Asl et al., 2020).

The NLU module maps the textual form of a user utterance into a semantic representation that is meaningful for the system (Chen et al., 2017b). The dialogue management module typically consists of DST and Policy learning. The DST sub module estimates the user's *belief state*, or updated goal throughout the conversation and typically involves slot filling. Usually, such slots and their possible values depend on a pre-existing domain ontology. The policy learning sub module uses the user's belief state to take an action. Methods for policy learning have typically involved supervised or Reinforcement Learning (RL) (Cuayahuitl, Keizer, and Lemon, 2015). The NLG module generates a natural language response from the dialogue action (Chen et al., 2017b). In commercial applications, the NLG module has typically consisted of template-based responses (Hosseini-Asl et al., 2020; Stent, Prasad, and Walker, 2004).

While traditional and commercial goal-oriented dialogue systems consist of independently trained modules, more recent methods have introduced approaches in which all modules are trained in an end-to-end fashion (Bordes, Boureau, and Weston, 2016; Wen et al., 2017).

### 2.2.2 *Open-ended chit chat*

Open-ended dialogue systems are typically concerned with maintaining a natural sounding and engaging conversation about any topic. Dialogue generation cast as a sequence-to-sequence problem has the advantage of being able to leverage large amounts of unannotated data such as conversations from

Twitter-style microblogs (Shang, Lu, and Li, 2015a) or movie scripts (Serban et al., 2015). The ability to incorporate information from previous turns in the conversation is important to keep conversations active. Sordoni et al. (2015b) introduced a method for encoding context with word embeddings. The response is then generated using an RNN language model. Serban et al. (2017) used a hierarchical neural model which encoded previous utterances and incorporated them into a context encoder.

Such methods, however, offer little control and have been shown to lead to fluent, yet meaningless and repetitive answers (Budzianowski and Vulić, 2019).

### 2.2.3 *Initial goals*

In this dissertation, the initial goal was to improve the interaction of users and goal-oriented dialogue systems. This section describes the more specific goals of the first studies in this thesis, some of the previous works motivating these goals, and some opportunities for using transfer learning.

**RESPONSE GENERATION** Open-ended dialogue models tend to be data-driven and often yield very fluent responses (Shang, Lu, and Li, 2015b; Wen et al., 2018; Zhang et al., 2018). However, these tend to offer flexibility but little control as to what the system will respond. In goal-oriented dialogue systems, where responses need to be precise and stay on topic and where there tends to be little annotated data; response generation is typically not data-driven.

As mentioned in Stent, Marge, and Singhai (2005), a good response generator should provide answers which fulfill the following criteria: adequacy, fluency, readability, and variation. An early goal of my PhD has been to improve dialogue generation of goal-oriented dialogue systems by incorporating neural sequence-to-sequence models for enhanced flexibility and variability of answers, but ensuring responses which are on topic, relevant and fluent. In this thesis, pretrained representations are used as a starting point and finetuned with an additional IR step (discussed in chapter 3).

**ADAPTATION TO NEW DOMAINS** Research in goal-oriented dialogue systems typically revolves around the DST and policy learning sub modules which have traditionally been trained on labeled data in a supervised fashion (Chen et al., 2017b). DST systems are usually based on the assumption that a domain ontology is available, simplifying the task to an intent classification problem (Henderson, Thomson, and Williams, 2014). In



past years, state-of-the-art approaches for state tracking have relied on deep learning architectures which represent dialogue state as a distribution over all possible slot values for each slot in the ontology (Henderson, Thomson, and Young, 2013, 2014; Mrkšić et al., 2016). Such systems tend to deal with one domain and a limited number of slots. In practice, multiple domains may be present throughout a conversation and obtaining an exhaustive ontology in advance is difficult. Additionally, such methods would not scale when the ontology and domain space gets significantly large.

Previous methods introduced for multidomain DST have typically involved: (1) sharing parameters across slots (Rastogi, Hakkani-Tür, and Heck, 2017; Williams et al., 2013), (2) sharing parameters across single domain systems (Williams et al., 2013), and (3) pretraining using disparate data sources and finetuning to a single domain (Mrkšić et al., 2015). These methods, however, transfer knowledge to unseen domains using *turn-level annotations* which may not be available or may be costly to obtain. An additional early goal in this dissertation was to improve domain adaptation of dialogue systems making use of feedback signals which may be obtained from real users in a more natural way (e.g., dialogue level signals) and making use of pretraining/finetuning scenarios (chapter 4).

With advances in Transformer-based pretraining and the promising results such methods have shown across many tasks, naturally these have begun to be adapted to dialogue systems. The next section describes key approaches which have been presented in recent years and which follow my early work.

#### 2.2.4 Recent advancements in dialogue using large pretrained models

Budzianowski and Vulić (2019) used GPT-2 pretraining and turned goal-oriented dialogue to a sequence-to-sequence problem. They evaluated their method on MultiWOZ (Budzianowski et al., 2018) and found that their simplified architecture performed about the same as the more complex pipelines.

Chao and Lane (2019) showed that using BERT as a dialogue context encoder coupled with parameter sharing across all slots led to significant improvements on several goal-oriented dialogue benchmarks (Budzianowski et al., 2018; Williams, Raux, and Henderson, 2016). Furthermore, Gulyaev et al. (2020) framed DST as a reading comprehension problem where the concatenation of slot and domain descriptions as well as the dialogue context serve as input, and the task is to return values for a dialogue state as an answer. They finetune BERT with several

classification and span-prediction heads for intent classification, categorical slot filling, free-form slot filling, among others.

Ham et al. (2020) present an end-to-end neural model for goal-oriented dialogue which is based on finetuning GPT-2 to perform the following steps in a single model: (1) DST, (2) Policy learning, (3) retrieval of appropriate records from a database and (4) NLG. This approach performed best in the DSTC-8 (Kim et al., 2019) and is also competitive with other state-of-the-art consisting of separate independently trained modules.

Hosseini-Asl et al. (2020) cast goal-oriented dialogue as a causal unidirectional LM task. This allows them to fully leverage transfer learning from causal language models such as GPT-2. They optimize for all submodules (NLU, dialogue management and NLG) jointly in an end-to-end manner and achieve state-of-the-art results on many automatic metrics. In recent years, dialogue-specific pretrained representations have also been developed (Mehri et al., 2019; Wu et al., 2020; Zhang et al., 2019).

As the data-driven approaches described earlier continue to push the state-of-the-art in many tasks and are deployed in real-world applications, several *ethical challenges* have been pointed out by researchers in the field of NLP as well as in other disciplines such as law and ethics. In dialogue systems for example, Henderson et al. (2018) analyze several datasets and models and find that: (1) discriminatory biases exist in most of the datasets evaluated and (2) the algorithms we use to develop word embeddings effectively encode and propagate such discriminatory patterns. They also point out several concerns in terms of model weaknesses, safety, and privacy.

The next section describes some of the ethical challenges emerging which have caused the focus of my research project to shift towards addressing the way in which humans are affected by and are interacting with NLP technologies.

### 2.3 RESEARCH SHIFT: ETHICAL CHALLENGES

Work on creating *universal* representations by pretraining on general language tasks, is an important and very active research area. Many of these methods have allowed us to create systems for a vast number of NLP tasks that exceed human performance on many benchmarks. However, there have been an increasing number of studies highlighting the challenges of such methods and urging researchers to consider the higher impact of the models they deploy (Hovy and Spruit, 2016). Below is a brief compilation of some of the ethical challenges which have been identified and outlined in recent years.



**IMPLICIT BIASES** In an extensive law review, Barocas and Selbst (2016) mentioned that “*discrimination may be an artifact of the data mining process*” and that statistical models learning from big data have the potential of placing marginalized groups “*at systematic relative disadvantage*”. This is an ethical concern that has been mentioned constantly in the last few years in the context of NLP models (Henderson et al., 2018; Hovy and Spruit, 2016) and researchers in NLP have started to address implicit biases in learned models and language representations (Caliskan, Bryson, and Narayanan, 2017; Manzini et al., 2019; Rudinger et al., 2018; Tan and Celis, 2019; Webster et al., 2019; Zhao et al., 2018).

**ADVERSARIAL EXAMPLES** Neural models which learn from raw historical data have shown vulnerabilities to adversarial examples (Goodfellow, Shlens, and Szegedy, 2014; Jia and Liang, 2017). A mainstream example in NLP is Tay, the Microsoft chat bot that learned from previous interactions without any control, and quickly learned inappropriate behavior. This was later attributed to *adversarial attacks*, in which users intentionally interacted with Tay using racial slurs and inappropriate language. Data-driven systems which have no human control are vulnerable to similar attacks.

**UNDEREXPOSURE NEGATIVELY IMPACTS EVALUATION** As has been an open discussion in the last few years in the NLP community, NLP tends to focus on Indo-European text sources as opposed to languages from Asia, Africa or the Americas (Hovy and Spruit, 2016). While there has been an increasing amount of work on multilingual and cross-lingual NLP, most commercial tools are geared towards English; and English-centric research has certainly shaped the way that NLP and evaluation has developed (Hovy and Spruit, 2016).

**PRIVACY CONCERNS** The vast amount of data used to train models can contain private information which can be recovered by using model inversion attacks (Song, Ristenpart, and Shmatikov, 2017; Yeom et al., 2018). In NLP, such attacks can be used to retrieve sensitive user information provided in conversational systems (Henderson et al., 2018), can pose risks for NLU systems for healthcare trained on private patients’ data (Huang et al., 2020) and demographic information could be retrieved from text (Li, Baldwin, and Cohn, 2018; Rosenthal and McKeown, 2011), among others.

**SAFETY CONCERNS** NLP systems are increasingly being used in critical domains such as health care and law (Dale, 2019; Fort and Couillault, 2016). In such domains, a wrong model decision which influences human decision making can have extremely negative consequences.

**DUAL USE PROBLEMS** Hovy and Spruit (2016) notes various instances of problematic *dual-use* of NLP applications. For example, the techniques used for the detection of fake reviews can also be used to generate them in the first place. While many technologies may be created with good intentions, unintended uses can have negative unforeseen effects on people’s lives.

The studies introduced in part iii of this thesis shift the focus from creating accurate and generalizable systems for human-AI interaction towards some of these issues. Chapter 5 focuses on the detection of negative social biases, particularly focusing on non-English models and leveraging linguistic phenomena not present in English. Additionally, several concerns about safety, diagnosing weaknesses to adversarial attacks, and detecting negative biases, can be addressed in part by increasing *transparency* in NLP systems. Chapters 6–8 investigate how *humans* interact with explainability methods to assess whether the current techniques introduced in NLP and ML are actually having the intended effects. The remaining section introduces some of the recent work within bias *diagnosis* in large pretrained models and *explainability* in NLP.

## 2.4 FAIRNESS AND TRANSPARENCY IN NLP SYSTEMS

Below, we give a brief introduction of recent work in bias detection and explainability. The discussion on bias is restricted to social biases (e.g., race, gender, etc.) rather than other types of biases such as inductive bias, cognitive bias, and media bias. We present some of the methods that are most relevant to understanding the motivation and the experiments in this thesis.

### 2.4.1 Detection of social bias in NLP

Distributed and deep contextualized representations (Devlin et al., 2019b; Mikolov et al., 2013; Peters et al., 2018), have led to huge improvements in a variety of NLP tasks. However, such representations are increasingly learned over large amounts of text data and taught to exploit statistical patterns in such corpora. These have been shown to encode social biases such as gender and racial biases (Bolukbasi et al., 2016; Caliskan, Bryson, and Narayanan, 2017; Manzini et al., 2019; Papakyriakopoulos et al.,

2020; Tan and Celis, 2019; Zhao et al., 2019). Some prior studies have measured bias in language representations using association tests from psychology (Caliskan, Bryson, and Narayanan, 2017; May et al., 2019b). Other works measure social biases in terms of representation bias (Pujari et al., 2019; Webster et al., 2019), meaning that based on disproportionate exposure to some groups, systems end up performing better or worse depending on the group. Recent work studied gender biases in word embeddings for German, which contains gendered pronouns, similar to the English case of *her*, *his* (Papakyriakopoulos et al., 2020). Similar to findings for English, they observed evidence for sexism, xenophobia and homophobic prejudice. Additionally, when using those word embeddings in a downstream task such as sentiment analysis, such biases were propagated.

Furthermore, Hutchinson et al. (2020) study biases in NLP models targeting people with disabilities. They devise a detection method using toxicity prediction and sentiment analysis. They find that three popular NLP models, which are readily deployed in many applications, *all* exhibit negative biases towards various types of disability groups.

The prior studies most related to the work explored in this thesis are presented next. Webster et al. (2019) created a *challenge dataset* for detecting gender bias in coreference models, using naturally occurring sentences from Wikipedia. Their dataset is gender balanced in terms of gender pronouns, and is useful for detecting when models have a preference for a particular gender. For example, they are correct more often for one gender rather than the other. Furthermore, Rudinger et al. (2018) and Zhao et al. (2018) both present challenge datasets for the task of coreference resolution using occupations as a probe. They evaluated state-of-the-art systems, correlated with occupation statistics from the U.S. and found that these systems encoded occupational stereotypes.

Benchmark datasets for detecting biases are a useful and fast way to check whether trained models have encoded negative and detrimental stereotypes. However, previous work is limited in terms of the tasks and the languages investigated. The previous diagnostic datasets mentioned have focused on English and only consider coreference resolution. As mentioned by Hovy and Spruit (2016), by focusing on English, the community also neglects a huge part of the world population who also interacts with NLP systems, and limits the type of evaluations and phenomena we study. In Chapter 5, work is presented tackling all these limitations. We introduce data for bias diagnosis in four languages and explore a linguistic phenomenon which does not

exist in English, to assess whether large multilingual models encode occupational stereotypes.

#### 2.4.2 Explainability methods

Work in explainability aims to make model decisions predictable or *transparent* to human end-users. Explainability is a very active area of research not only in NLP and other subfields of AI, but also in HCI. Desirable qualities of explanations are still being uncovered, however, explanations of model predictions should at least (1) improve human understanding, (2) improve confidence and trust in decision-making and (3) promote fair decisions (Das and Rad, 2020). Explanations can take many forms and can be presented in many ways. For example, an explanation can be *global*, meaning that it explains the overall behavior of the model, or *local*, meaning it explains the behavior for a specific decision. There are many kinds of explainability methods that can be used in NLP, however, this dissertation highlights two: *attribution-based* methods which are used to highlight important features (words) in the input and *natural language explanations* which provide a textual justification of the model behavior. The work presented here is limited to *local* explanations.

**ATTRIBUTION-BASED EXPLANATIONS** Local *model-agnostic* attribution methods work under the assumption that the predictions of a black box model around a neighborhood of the input can be approximated by an inherently interpretable model. Ribeiro, Singh, and Guestrin (2016a) introduced LIME, which generates new examples based on permutations of the input and trains an explainable model (such as decision trees or a lineal model) to see how the model’s predictions behave. In recent years, LIME has been improved and extended to several new use cases, for example, Bramhall et al. (2020a) presented Quadratic LIME (QLIME) which considers nonlinear relationships, and Sound-LIME (SLIME) (Mishra, Sturm, and Dixon, 2017), an extension aimed at music content analysis. Another popular *model-agnostic* explainability method is SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), a perturbation-based method which explains predictions on an input by computing individual feature contributions towards the output. Lundberg and Lee (2017) also explores several variations such as KernelSHAP and LinearSHAP.

Another direction involves *model-specific* methods. *Gradient-based feature attribution* is obtained by computing the gradient of the output class with respect to the input (Simonyan, Vedaldi, and Zisserman, 2013). These attributions are typically visualized

as saliency maps. Several variants of vanilla gradient attribution exists, for example, **IG** (Sundararajan, Taly, and Yan, 2017), **InputXGradient** (Adebayo et al., 2018), **Guided Backpropagation** (Shrikumar, Greenside, and Kundaje, 2017) and **Layer-wise Relevance Propagation (LRP)** (Bach et al., 2015).

This dissertation evaluates explanations derived from both model-agnostic and model-specific explainability methods.

**NATURAL LANGUAGE EXPLANATIONS** Another way of explaining model decisions is by providing natural language justifications of model behavior. Camburu et al. (2018) and Rajani et al. (2019) both introduced methods for training models using free-form natural language explanations collected from crowdsourced workers for the tasks of **NLI** and common sense reasoning. Recently, Lamm et al. (2020) introduce **QED** explanations for Open-domain Question Answering (**ODQA**), which are linguistically informed and consist of the sentence containing the answer, coreference and entailment information.

Atanasova et al. (2020b) introduce a method for generating explanations for fact verification using human veracity justifications. Lei, Barzilay, and Jaakkola (2016) introduced an approach for extracting *rationales* by selecting or *extracting* phrases from the input text which are sufficient to provide an output. Rationales are widespread in practice for applications such as Question Answering (**QA**), and have recently been introduced for a variety of other **NLP** tasks such as **NLI**, fact verification, and common sense reasoning (Chen et al., 2018a; DeYoung et al., 2020; Popat et al., 2017, 2018). This dissertation evaluates the effectiveness of extractive rationales. The work in this thesis also evaluates manually generated rationales in the form of *abstractive* summaries. With some notable exceptions (Kotonya and Toni, 2020), such abstractive rationales are not as frequently studied, however, they have the potential to provide benefits in scenarios where evidence spans multiple documents (Yang et al., 2018).

### 2.4.3 Evaluation of explainability

Natural language explanations such as rationales have been evaluated using discrete overlap metrics such as token **F1**, **BLEU**, and **Intersection Over Union (IOU)** to measure agreement with human rationales (DeYoung et al., 2020; Paranjape et al., 2020; Swanson, Yu, and Lei, 2020), or **Area Under the Precision-Recall Curve (AUPRC)** for continuous or soft token scoring (DeYoung et al., 2020).

Robnik-Šikonja and Bohanec (2018) evaluated attribution-based explanations based on consistency and stability properties. Consistency captures the difference between explanations of different models producing the same prediction, while stability measures the difference of explanations of similar instances within a model. Additionally, Atanasova et al. (2020a) extend on previous work and propose a list of diagnostic properties for evaluating explainability techniques such as agreement with human saliency rankings, confidence indication, faithfulness, rationale consistency and dataset consistency. Their protocol used metrics such as Mean Average Precision (MAP), Area Under the Threshold-Precision Curve (AUTPC), and spearman correlation. Other studies have used Area Over the Perturbation Curve (AOPC) to measure the local fidelity of explanations (Nguyen, 2018b). Automatic metrics may capture general aspects of explanations, which may be desirable. However, they do not capture the utility of explanations in the real world or give us notions of how explanations make model behavior transparent to human end-users. In this thesis, I argue that human evaluation is therefore, a more valuable way to investigate explanations.

Some work in human evaluation of explanations has been presented in NLP, but a larger quantity of human evaluations has been proposed by researchers in the HCI community and within other subfields of AI. Some recent methods are briefly mentioned next.

Lertvittayakumjorn and Toni (2019) proposed three human tasks for evaluating different desirable properties of explainability methods for text classification. The tasks consisted of humans assessing whether explanations (1) reveal model behavior, (2) justify model predictions, and (3) allow them to investigate uncertain predictions. Such studies, however, exhibited low inter-annotator agreement.

Other works used a simulatability task proposed by Doshi-Velez and Kim (2017), which consists of providing humans with explanations and having them decide what the model output would be (Hase and Bansal, 2020; Nguyen, 2018b). Other human evaluation has involved subjective measures (Ribeiro, Singh, and Guestrin, 2016a; Selvaraju et al., 2017; Weitz et al., 2019) and model ranking (Ribeiro, Singh, and Guestrin, 2016a). Fewer evaluations involved more realistic decision making tasks (Bansal et al., 2019a, 2020; Buçinca et al., 2020). However, recent work has found that current explanation strategies may not be any more effective in helping users in real decision making tasks than simply showing model confidence (Bansal et al., 2020).

The studies presented in this dissertation evaluate explainability methods through user studies exploring the following paradigms: simulatability, model ranking, and decision making.





## Part II

# LEARNING FROM USER INTERACTIONS



## RETRIEVAL-BASED GOAL-ORIENTED DIALOGUE GENERATION

---

### 3.1 ABSTRACT

Most research on dialogue has focused either on Natural Language Generation (NLG) for *open-ended dialogues* consisting of encoder-decoder neural architectures, or on *goal-oriented* dialogue focusing on Dialogue State Tracking (DST) and dialogue policy. In practice, generation of responses in goal-oriented dialogue systems (e.g. in domains such as customer service), relies on templates which provide more controlled and relevant but limited responses. In this work, we investigate a simple way of taking advantage of the flexibility provided by neural encoder-decoder models for the task of *goal-oriented dialogue generation*. More specifically, we incorporate retrieved past user interactions into a standard hierarchical dialogue generation model often used in open-ended dialogue systems. We show that adding this simple-yet-effective retrieval step leads to significant improvements in various automatic metrics and leads to responses that are rated more relevant and fluent by human evaluators in the customer support domain.

### 3.2 INTRODUCTION

Dialogue systems have become a very popular research topic in recent years with the expanding availability of personal assistants and the growing demand for online customer support. Research within dialogue has typically been split into two sub areas (Chen et al., 2017b): models presented for the generation of open-ended conversations (Li et al., 2017a; Ritter, Cherry, and Dolan, 2011; Serban et al., 2015; Shibata, Nishiguchi, and Tomiura, 2009; Sugiyama et al., 2013) and work on solving goal-oriented dialogue through dialogue management pipelines that include Dialogue State Tracking (DST) and dialogue policy (Bingel et al., 2019; Henderson, Thomson, and Young, 2013; Mrkšić et al., 2016; Ren et al., 2013; Ren et al., 2018; Sun et al., 2014; Yoshino et al., 2016; Zhao and Eskenazi, 2016).

DST typically consists of a Natural Language Understanding (NLU) step for detecting user intent and slot-value pairs and a step for updating the *belief state* of the user goals. Learning a

dialogue policy typically consists of determining what actions the system should take based on the updated belief state.

Work on open-ended conversation, in contrast, has largely been concerned with *dialogue generation*, and has relied on transduction architectures originally developed for Machine Translation (MT) (Shang, Lu, and Li, 2015b; Wen et al., 2018; Zhang et al., 2018). Such architectures encode an utterance into a fixed-sized vector representation and decode it into a variable length sequence that is linguistically very different from the input utterance. While MT-based approaches offer the flexibility of generating answers that are more varied, such methods often lack the ability to encode the context in which the current utterance occurs. As a result, such methods often lead to repetitive and meaningless responses (Li, Luong, and Jurafsky, 2015; Lowe et al., 2017a; Wen et al., 2018).

This observation has led researchers to extend simple encoder-decoder models to include context in order to deal with generation of larger structured texts such as paragraphs and documents (Li, Luong, and Jurafsky, 2015; Serban et al., 2016, 2017). Many of these models work by encoding information at multiple levels, e.g., encoding context consisting of multiple previous utterances and the most recent utterance. While popular in open-ended chit chat, it is not clear how such hierarchical methods can be used in practice for goal-oriented dialogue pipelines, where responses need not only be coherent and fluent, but also relevant.

In goal-oriented dialogue, there is often one (context-dependent) right answer to a question (e.g., *How many types of insurance do you offer?*); in chit-chat, there are many good answers to questions (e.g., *What do you want to talk about today?*). In addition, in narrower domain systems such as customer service domains, there may be more repetition in the types of questions users are asking. Therefore, we hypothesize that in goal-oriented dialogue, it may be beneficial to increase the inductive bias of the dialogue generation model by taking advantage of previous conversations to keep responses relevant. We do so by introducing a simple-yet-effective dialogue generation model that conditions decoding on retrieved user interactions from labeled past history.

Although retrieval approaches to dialogue generation have been introduced before, they have typically involved external sources to add more variety to the kind of answers the model can generate in open-ended conversations (Ritter, Cherry, and Dolan, 2011; Weston, Dinan, and Miller, 2018). Our model, in contrast, uses past conversations and is designed for improving NLG in *goal-oriented dialogue*.

**CONTRIBUTIONS** We present an effective **NLG** model aimed at improving the quality and relevancy of answers in goal-oriented dialogue systems. It is a hierarchical neural encoder-decoder model with an Information Retrieval (**IR**) component that *obtains the most informative turns from prior user interactions and conditions on those*. Our results show that this simple **IR** step leads to improvements over a traditional dialogue generation model intended for open-ended dialogue when evaluated on BiLINGUAL Evaluation Understudy (**BLEU**) and different embedding metrics previously used for evaluating dialogue generation models (Serban et al., 2016; Sharma et al., 2017). More importantly, the system responses of our proposed model are rated by *human evaluators* as more fluent and relevant than the responses generated by the strong baseline.

### 3.3 MODEL DESCRIPTION

We extend the Hierarchical Recurrent Encoder-Decoder (**HRED**) model presented by Sordoni et al. (2015a) for query suggestion, and subsequently adopted for dialogue by Serban et al. (2016), which has shown to be a strong baseline for the task of dialogue generation. In line with previous research, we consider a dialogue  $D$  between two speakers composed of  $M$  utterances so that  $D = [U_1, \dots, U_M]$  and each utterance  $U_n$  composed of  $N_m$  tokens so that  $U_m = [t_{m,1}, t_{m,2}, \dots, t_{m,N_m}]$ . Each token  $t_{m,N_m}$  represents a word from a set vocabulary.

#### 3.3.1 **HRED**

The **HRED** model for query suggestion (Sordoni et al., 2015a), predicts the next web query given previous queries already submitted by the user. The history of queries is encoded at two levels: a sequence of words for the last web query and a sequence of previous queries. The **HRED** model applied to dialogue (Serban et al., 2016), assumes that dialogues can be modeled in a similar way: by encoding previous user utterances at the word level and at the turn level.

**HRED** applied to dialogue, consists of an *utterance encoder* Recurrent Neural Network (**RNN**), a *context RNN* and a *decoder RNN*. The *utterance encoder* maps an utterance to a vector which is the hidden state obtained after the last token is produced. The *context RNN* summarizes the dialogue history (up to and including the current utterance) by keeping track of the previous hidden states. The *decoder RNN*, decodes the hidden state of the context **RNN** by producing a probability distribution over the tokens in the next utterance.

Just as in previous works (Serban et al., 2015; Serban et al., 2016, 2017; Shen et al., 2017), we use Gated Recurrent Unit (GRU) (Cho et al., 2014) for the utterance encoder, context encoder and decoder RNN. All modules are trained end-to-end.

### 3.3.2 Exemplar-HRED

In this study, we propose a simple enhancement to HRED by adding an efficient IR step. As already mentioned, similar approaches have been presented with the goal of incorporating factual information into open-ended conversations (Weston, Dinan, and Miller, 2018), to add variety and more topics to the conversation by retrieving facts from Wikipedia, however, our goal is to investigate how such methods can be used in goal-oriented dialogue, where relevancy of system responses is more important. To this end, we hypothesize that incorporating *exemplar* user conversations is beneficial.

**ARCHITECTURE** Our proposed model uses the same architecture as the HRED baseline, however, we include an additional RNN, which encodes the top example response (details later in this section). Just as in the baseline model, the user *utterance encoder* outputs a vector representation of the user utterance. Additionally, we encode the *exemplar* using the *example encoder*. The resulting representations of the *example response* and *user utterance* are concatenated and fed to the context RNN, which summarizes previous user interactions and examples of successful system responses to similar questions. This *global context* is then fed into the decoder. A graphical representation of our proposed model can be seen in Figure 3.1.

For all experiments, we use the MultiWOZ dataset for goal-oriented dialogue (Budzianowski et al., 2018), which is described in section 3.4.1. We initialize our model and the baseline model using GloVe embeddings. Our model uses the Adam optimizer (Kingma and Ba, 2014) for all encoders. All our encoders are one layer RNN's. We use a dropout rate of 0.3 and a learning rate of 0.001. We set a maximum of 50 training epochs, however, we use early stopping with a patience of 10. Most of our models converge by epoch 30. We use greedy search to generate the response during testing. More implementation details as well as our predicted utterances for each system can be found in the link provided.<sup>1</sup>

**RETRIEVAL STEP** The retrieval step happens offline (before training). For each user utterance, we extract a 300-dimensional

<sup>1</sup> [https://github.com/anavaleriagonzalez/exemplar\\_dialog](https://github.com/anavaleriagonzalez/exemplar_dialog)

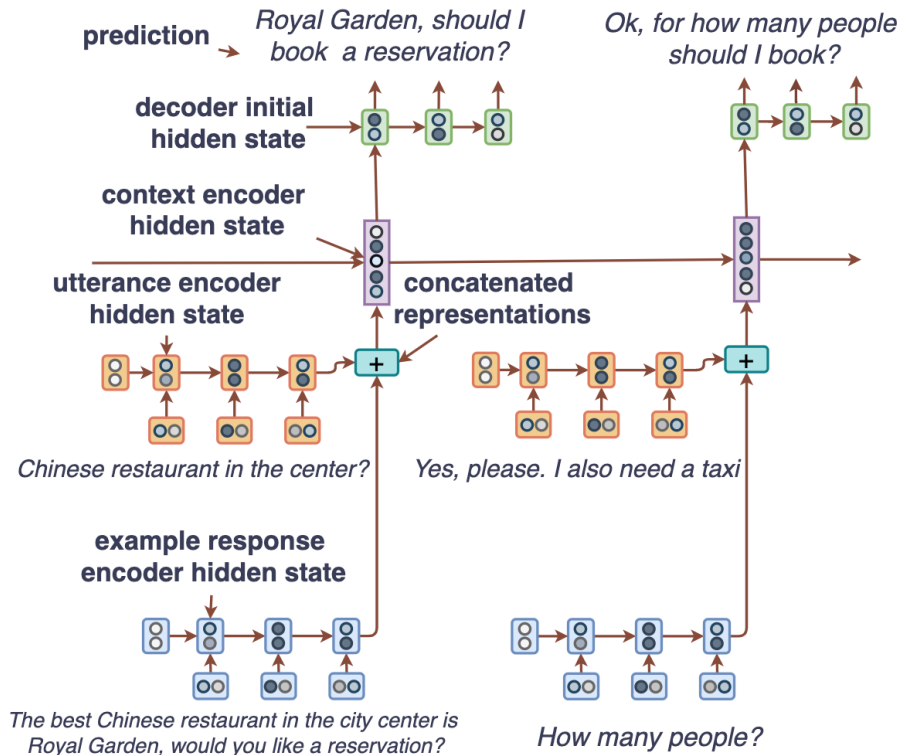


Figure 3.1: Our model is similar to HRED (Sordoni et al., 2015a), we include an utterance encoder, a context encoder and a decoder, however, unlike HRED, our model include a simple, yet effective retrieval step used to condition the decoder to generate responses that are more appropriate for a specific domain and context.

sentence representation by computing the average GloVe embeddings of sentence tokens (Pennington, Socher, and Manning, 2014). We obtain the ten most similar past user utterances from the training set using Approximate Nearest Neighbor Search (ANNS) (Indyk and Motwani, 1998). Nearest Neighbor Search (NNS) is defined as: given a set  $S$  of points in space  $M$ , and a query point  $q \in M$ , find the point  $p$  in  $S$  which is closest to  $q$ . Retrieval of the *exact* nearest neighbor requires exhaustive search which is not efficient in practice. ANNS allows us to increase the speed of retrieval by approximating a point  $p \in S$  which is close to  $q$ . While there are various methods for ANNS, we use the *annoy*<sup>2</sup> python package, which uses a random projection method. It builds a forest that can be indexed, where each tree is constructed by picking two points at random and splitting the space by the hyperplane equidistant from the two points. This is done  $k$  times ( $k$  is a hyperparameter, in our case  $k = 100$ ) in the subspaces until the points associated with a node are small enough. After the forest is constructed, it can be stored and easily indexed (in logarithmic time) by traversing

<sup>2</sup> <https://github.com/spotify/annoy>

the trees from the root. We further improve the ranking of the retrieved utterances using a feed-forward ranking model (Gonzalez, Augenstein, and Søgaard, 2018). In the end, we return the highest ranked user utterance and use *its response as the example* to be used in our model.

### 3.4 EXPERIMENTS

#### 3.4.1 Dataset and preprocessing

We use the MultiWOZ dialogue corpus which consists of 10,438 dialogues spanning several domains and annotated with dialogue states and acts (Budzianowski et al., 2018). We train on 8,438 dialogues and use 1000 dialogues for development and 1000 dialogues for testing. Although the data is primarily intended for DST and learning a dialogue policy, it is appropriate for generation tasks as it contains about 115k turns in total; making it considerably larger than many other goal-oriented dialogue corpora available. Dataset statistics can be seen in Table 4.1. The MultiWOZ dataset is also more difficult than the current benchmarks for goal-oriented dialogue, as it spans 7 different customer support domains and conversations are not limited to a single domain. We delexicalize the utterances to remove phone numbers, reference numbers, and train IDs, which would in practice usually be retrieved from a knowledge base. For delexicalizing, we use the ontology provided with the data and replace the value with the *slot names* e.g., replacing *The reference ID is A23N5* with *The reference ID is train-id* using regular expressions.

Statistic	MultiWOZ
# DIALOGUES	8438
TOTAL # TURNS	113,424
TOTAL # TOKENS	1,520,970
TOTAL UNIQUE TOKENS	24,071

Table 3.1: Statistics of the MultiWOZ training data

#### 3.4.2 Metrics

Evaluation of open-ended dialogue systems is an open problem (Pietquin and Hastie, 2013; Schatzmann, Georgila, and Young, 2005). Word overlap metrics such as the ones used for machine translation are often used to evaluate the quality of *dialogue gener-*



ation (Lowe et al., 2017b, 2016). We include these, as well as word embedding metrics previously used for measuring textual similarity (Wieting et al., 2015). While Liu et al. (2016) showed that these metrics tend to not correlate with human judgements in open-ended dialogue, Sharma et al. (2017) showed that they have stronger correlations with human evaluation in goal-oriented dialogues, where answers are in narrower domains and exhibit lower diversity. To calculate the results, we use the evaluation script from Serban et al. (2016)<sup>3</sup>.

In addition, we evaluate the fluency and relevancy of responses through *human evaluation*, which is more valuable. We collect human ratings from 7 evaluators, which gives us a better indication of how successful the model responses are. We briefly discuss all the metrics below:

**BLEU** BLEU (Papineni et al., 2002) is typically used for machine translation and has subsequently been used to evaluate the performance of many dialogue generation systems (Galley et al., 2015; Serban et al., 2016, 2017). BLEU analyzes cooccurrences of n-grams in a reference sequence and a hypothesis. It uses a modified precision to account for the differences in length between reference and generated output.

**AVERAGE WORD EMBEDDING SIMILARITY** We follow Wieting et al. (2015) and obtain sentence embeddings for the reference response by taking the average of its word embeddings. We do the same for the predicted output and obtain the final similarity score by computing cosine similarity of the two resulting vectors.

**VECTOR EXTREMA** Vector extrema is another way of obtaining sentence embeddings (Forgues et al., 2014). This consists of taking the most extreme value (minimum or maximum) of the embeddings of the words composing a sentence for each dimension. We do this for both reference and system responses and then compute the cosine similarity between them.

The goal of this metric as described by previous work (Liu et al., 2016; Sharma et al., 2017) is to consider informative words rather than common words, since the vectors for common words will tend to be pulled towards the zero vector.

**HUMAN EVALUATION** The previous metrics provide only a vague measure of similarity between the system output and a gold response. These metrics do not inform us of what is

<sup>3</sup> <https://github.com/julianser/hed-dlg-truncated/tree/master/Evaluation>

actually preferred by *humans*. Therefore, in addition to the previously mentioned standard metrics, we evaluate the performance of the baseline and the exemplar-HRED models using human evaluations.

We extract 100 baseline and exemplar-HRED system responses at random. 7 evaluators rated all 100 responses independently. They were presented with three-turn dialogues consisting of a system utterance (to provide additional context), user query, and then both baseline and exemplar-HRED system responses.

The evaluators were asked to do two things: choose the response that was more fluent and grammatical and choose the response that provides the *most relevant* answer given the context of the conversation. For each of those, they had 4 options to choose from: 1) the output of the baseline, 2) the output of the exemplar model, 3) both, or 4) none. The order of the options was shuffled.

### 3.5 RESULTS AND DISCUSSION

Overall, we found that in most cases, our model leads to significant improvements over all metrics; see Table 3.2 for the results. We observe that differences between our model and the baseline range widely across metrics. For example, while our proposed model leads to better average embedding similarity scores, this metric is high for both models which can mislead one into thinking both models are performing really well. However, we see that the differences observed in the human evaluation are considerably larger, with the baseline model performing very low and only being preferred 14-19% of the time.

The improvements in BLEU score suggest that our model is returning similar tokens to a reference response, more often than the baseline. The vector extrema similarity score suggests that our model is better than the baseline at matching the informative words of the reference responses. However, we turn our attention to our human evaluation results.

Overall, we found that when it came to *fluency*, evaluators perceived that 58% of the time, the exemplar-HRED response was better. The baseline beat our model's response for 19% of the evaluated dialogues. For the rest of the dialogues, either both or none were picked. For *relevancy* of system response, we see a similar pattern. Evaluators perceived the response produced by our model as most relevant given the context (59% of the evaluated dialogues). The baseline beat the proposed model only 14% of the time. These results can also be found on Table 3.2.

METRIC	HRED	EXEMPLAR-HRED
BLEU	23.6	<b>24.1</b>
Vector Extrema Sim	59.0	<b>65.0</b>
Avg Embedding Sim	93.0	<b>95.0</b>
Human Eval- Fluency	0.19	<b>0.58</b>
Human Eval- Relevancy	0.14	<b>0.59</b>

Table 3.2: The results of our dialogue generation experiments comparing HRED Serban et al. (2016) and Sordoni et al. (2015a) to our proposed exemplar-based model. We present results for standard metrics used in dialogue generation. For all the metrics we observe improvements over the strong baseline, with our best improvement of 6 percent in the vector extrema metric

**ANALYSIS OF SYSTEM OUTPUTS** As it is hard to interpret what the differences in the performance of each model really mean, we examined the output to spot the major differences in response generation of our proposed models versus the baseline. We provide a few example responses in Table 3.3, while other examples are mentioned throughout this section. We make the following observations:

**Exemplar-HRED chooses the right domain more often.** While inspecting the outputs, it seemed that our model generated responses containing slots for the correct domains more often than the baseline. This can be observed in the first example in Table 3.3. While both models are able to provide responses dealing with *booking*, the baseline model returns a response concerned with the *restaurant* domain, while the true domain is *hotel* (which our model generates correctly).

We quantify this by annotating 200 examples from the baseline and 200 examples from our proposed model (responses to the same dialogue contexts), to see if they match the domain of the reference. We stick to responses where the domain should be mentioned (as in the first example in Table A.11). We find that out of 200 responses, the baseline matches the domain only for 33 of them (16.5%). On the other hand, our model matches the domain 150 times (75% of the time).

This confirms the findings from the human evaluation, showing that the simple retrieval step we incorporate has a significant positive effect in the *relevancy* of responses, which can be further

Gold Response	Exemplar-HRED	HRED
your booking was successful and your reference number is hotel-reference . is there anything else i can help you with ?	your booking was successful . your reference number is hotel-reference . is there anything else i can help with ?	the booking was successful . the table will be reserved for 15 minutes . reference number is : restaurant-reference .
there is no option for me to book it . is there anything else i can help you with ?	certainly ! the reference number is restaurant-reference . can i help you with anything else ?	the train-trainid leaves at 09:01 and arrives in cambridge by 09:52 . would that work ?
where will you leave from ?	sure , when would you like to leave ?	have a nice stay .

Table 3.3: Examples of responses generated by both the baseline and our proposed model. By examining the outputs, it becomes noticeable that the baseline model tends to generate responses that are not precise about the current domain of the conversation (hotel, taxi booking, trains, restaurant, etc).

improved by incorporating other information such as the output of a [DST](#) model.

**Exemplar-HRED returns semantically compatible entities.** Additionally, we observe that baseline ([HRED](#)), despite the improvements over previous architectures, still provides responses which are grammatically correct but meaningless in content. For example, [HRED](#) outputs *the price is located at hotel-address. is there anything else i can help you with today ?* while our proposed model is able to provide entities that are semantically compatible e.g. *sure thing ! the postcode is post-code and the phone number is phone-number.*

Additionally, we observed that our model tends to add more follow up questions e.g. *anything else I can help with?* and filler words such as *sure, certainly!* , which may make the language sound more natural and in turn make these responses preferred by users.

### 3.6 RELATED WORK

Open domain dialogue systems aim to generate fluent and meaningful responses, however this has proven to be a challenging task. Most systems are able to generate coherent responses that are meaningless and at best entertaining (Lowe et al., 2017a; Serban et al., 2016; Wen et al., 2018). Much of the research on dialogue generation has tried to tackle this problem by predicting an utterance based on some dialogue history (Luan, Ji, and Ostendorf, 2016; Serban et al., 2016; Shang, Lu, and Li, 2015a; Vinyals and Le, 2015).

Most research on goal-oriented dialogue has focused almost exclusively on dialogue state tracking and dialogue policy learning (Bingel et al., 2019; Henderson, 2015; Henderson, Thomson, and Young, 2014; Li et al., 2017b; Mrkšić et al., 2016; Rastogi, Hakkani-Tür, and Heck, 2017; Sun et al., 2014, 2016; Yoshino et al., 2016). On the contrary, we focus on *dialogue generation* for goal-oriented dialogue, which has not been studied as often.

The idea of combining text generation with past experience has been explored before. White and Caldwell (1998) used a set of hand crafted examples to generate responses through templates. More recently, Song et al. (2016) also explored a hybrid system with an information retrieval component, but their system is very different: It uses a complex ranking system at a high computational cost, and they only evaluate their system in an open-ended chit-chat set-up, reporting only BLEU scores. In a similar paper, (Weston, Dinan, and Miller, 2018) tried to move away from short generic answers in order to make a chit-chat generation model more entertaining by using retrieval of relevant facts from Wikipedia. A similar method was recently shown to improve other generation tasks such as summarization. In Subramanian et al., 2019, the authors show that a simple extractive step introduces enough inductive bias for an abstractive summarization system to provide fluent and precise summaries. We extend this method to dialogue, with a retrieval step to include previous user conversations, which improves the *relevancy* and *fluency* of system responses.

### 3.7 CONCLUSION

We have introduced a simple-yet-effective way of conditioning a goal-oriented dialogue generation model. Generating fluent and precise responses is crucial for creating goal-oriented dialogue systems. We propose adding a simple retrieval step, where we obtain the past conversations that are most relevant to the current one and condition our responses on those. We find that this method not only improves over a strong baseline on word overlap metrics and other automatic metrics, but it also is preferred by human annotators. Finally, by inspecting the output of the baseline versus our proposed model, we identify different areas where our method leads to improvements.



## DOMAIN TRANSFER IN DIALOGUE SYSTEMS WITHOUT TURN-LEVEL SUPERVISION

---

### 4.1 ABSTRACT

Goal-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 and unrealistic 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 (RL) 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 that our method can improve models trained using turn-level supervision by subsequent finetuning optimization with dialog-level rewards.

### 4.2 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, 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 Challenge (DSTC) (Williams et al., 2013), this line of research has received considerable interest.

State-of-the-art models for DST 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 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 pretrains a model on a source domain for which turn-level annotations exist, then finetunes 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. In addition to this, we also report consistent gains using (modeled) 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 finetuning, we show that it is possible to transfer between domains in DST using RL, 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.

### 4.3 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 the slot representations, again derived from word vectors and intermediate layers. We update the hidden state of a Gated Recurrent Unit (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



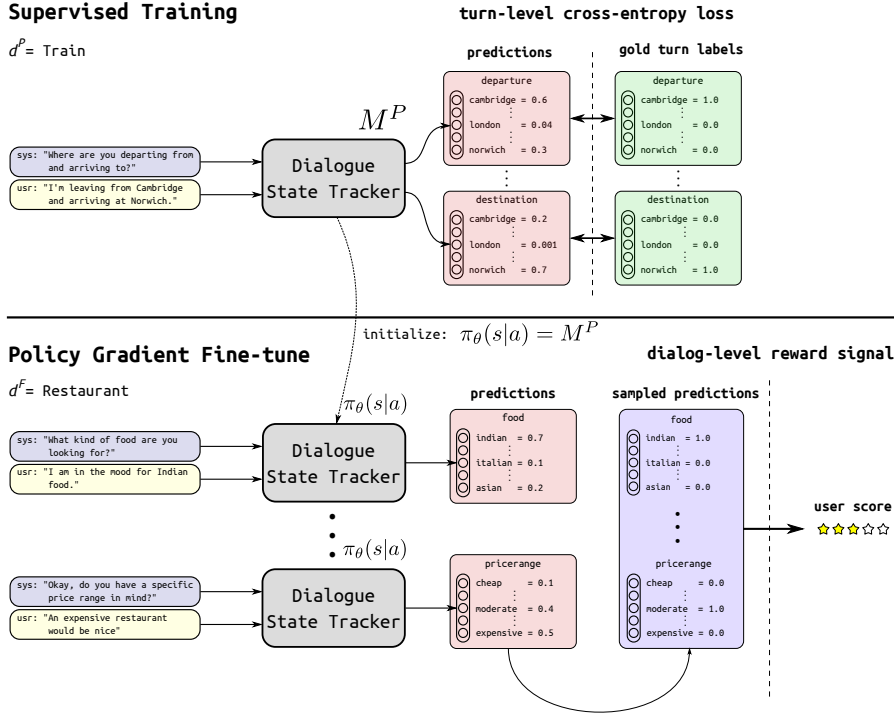


Figure 4.1: Illustration of our proposed domain transfer dialogue state tracker, using a model  $M^P$  trained with turn-level supervision on  $d^P$  as a starting point for the finetuning policy  $\pi_\theta(s|a)$  on domain  $d^F$ .

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 DSTC-2 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 pretraining 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 finetune  $M^P$  to the new domain, based on *simulated user feedback*, corresponding to how many goals we met at the end of the conversation. Figure 4.1 presents an overview of this process.

**PRETRAINING** In the pretraining 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 pretraining, we rely on turn level supervision, training models on a single domain and evaluating on a held out set from that same domain.

#### 4.4 DOMAIN TRANSFER USING REINFORCEMENT LEARNING

**DIALOGUE STATE TRACKING WITH RL** Given a pretrained model  $M^P$  trained on a domain  $d^P$ , we finetune 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 RL 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 interacts 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 4.1.

**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:

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

$$R_{goal} = \frac{|S_G \cap S_P|}{|S_G \cup S_P|} \quad (4.1)$$

**POLICY GRADIENT METHODS** We define the policy network  $\pi_\theta$  as the StateNet network, which is initialized with a pre-trained model  $M^P$ . The weights of the StateNet network are then finetuned using stochastic gradient ascent, i.e., in the direction of the gradient of the objective function  $\nabla J(\theta)$ . The update in the vanilla policy gradient algorithm is:

$$\nabla J(\theta) = \nabla_\theta \log \pi_\theta(a|s) R_{goal} \quad (4.2)$$

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

**VARIANCE REDUCTION METHODS** Policy gradient methods suffer from 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 pretrained 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. (4.2), we can rewrite the final gradient as:

$$\nabla J(\theta) = \nabla_\theta \log \pi_\theta(s|a) A_{goal} + \alpha \mathcal{H}(\pi_\theta(s|a)), \quad (4.3)$$

where  $\alpha$  is a term that controls the 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 that 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).

## 4.5 EXPERIMENTS

### 4.5.1 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 a 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), DSTC-2 (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 4.1 presents the statistics of domains used in the experiments with the distinction between the cases when the dialogue consists of only one or more domains.

**PREPROCESSING MULTIWOZ** The user utterances and system utterances used to train our model contain tokens that were randomly generated 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

Pretrain \ 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
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 4.2: Accuracy scores for our pretrained baseline (BL) and the policy gradient finetuning (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 finetuned results, and each pair of columns compare the 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.

this special token and add that to the ontology. 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 diverged from instructions. This led 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 in 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.<sup>1</sup>

#### 4.5.2 Implementation Details

Our pretrained StateNet model is implemented without parameter sharing and is not initialized with single-slot pretraining as in Ren et al. (2018). We use the Adam optimizer (Kingma and Ba, 2014) 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 pretrained embeddings

<sup>1</sup> <https://github.com/coastalcph/dialog-rl>

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 finetuning 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 finetuning 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.

#### 4.5.3 *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 contain the labels of that single domain. We hypothesize that this serves as an upper bound to the performance of the policy gradient finetuning. 2) Evaluating the pretrained models as a cross-domain zero-shot baseline. We take a model pretrained on  $d^P$  and measure its performance on  $d^F$  for all domains in  $D - d^P$ . This serves as the lower bound for the performance of the policy gradient finetuned models. We use this baseline and not a model finetuned 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) finetuning the pretrained model  $M^P$  to all other domains with policy gradient as described in Section 4.4. 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 finetuning a model using policy gradient on the same domain it was pretrained on ( $d^P$ ) after convergence 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.

#### 4.6 RESULTS

In Table 4.2 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 finetuning, compared to the zero-shot transfer baselines. In some instances we also see an increase in performance from further finetuning 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 finetuning. On average, we see relative improvements of the accuracy, ranging from 0.03 to 0.2, when applying our proposed method of finetuning for *DST* domain transfer.

#### 4.7 ANALYSIS

To illustrate the effectiveness of PG finetuning compared to zero-shot domain transfer, we plot in Figure 4.2 the results of training a model on the source domain *HOTEL* while evaluating 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 finetuning uses the pretrained model as a starting point to further improve the accuracy on the target domain using only the dialog-level feedback. Figure 4.2 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, the performance suddenly starts to deteriorate for the latter when the model overfits to the source domain.



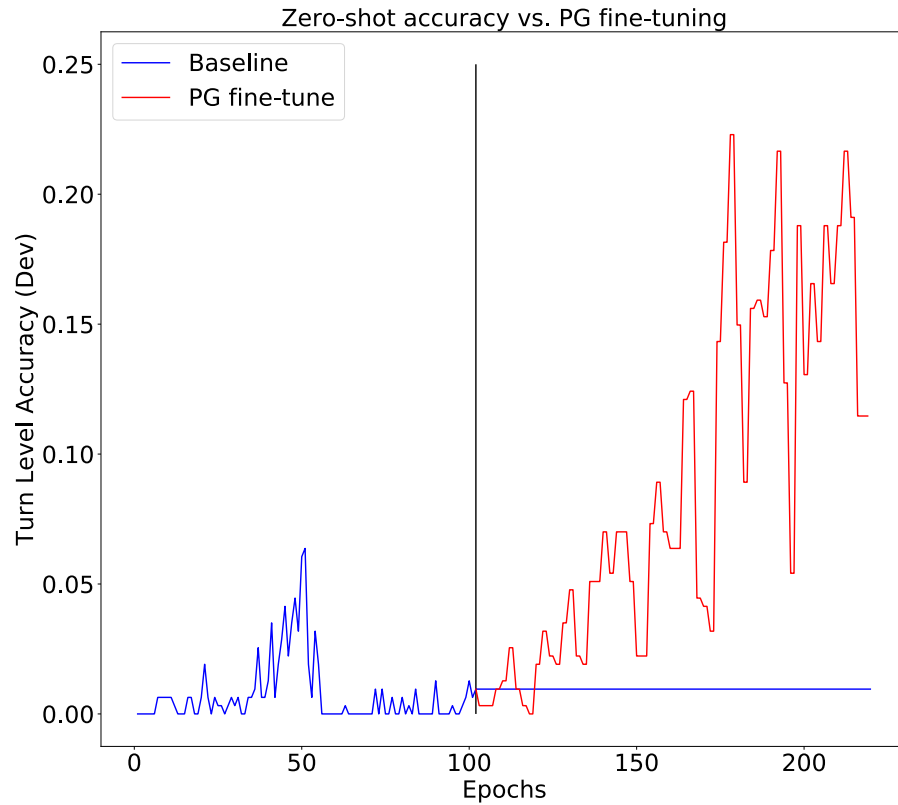


Figure 4.2: The performance of the supervised model trained on the HOTEL domain while evaluated on the development set of the TAXI domain after each epoch until convergence on HOTEL versus the improvements we get from the policy gradient finetuning using the supervised model as starting point.

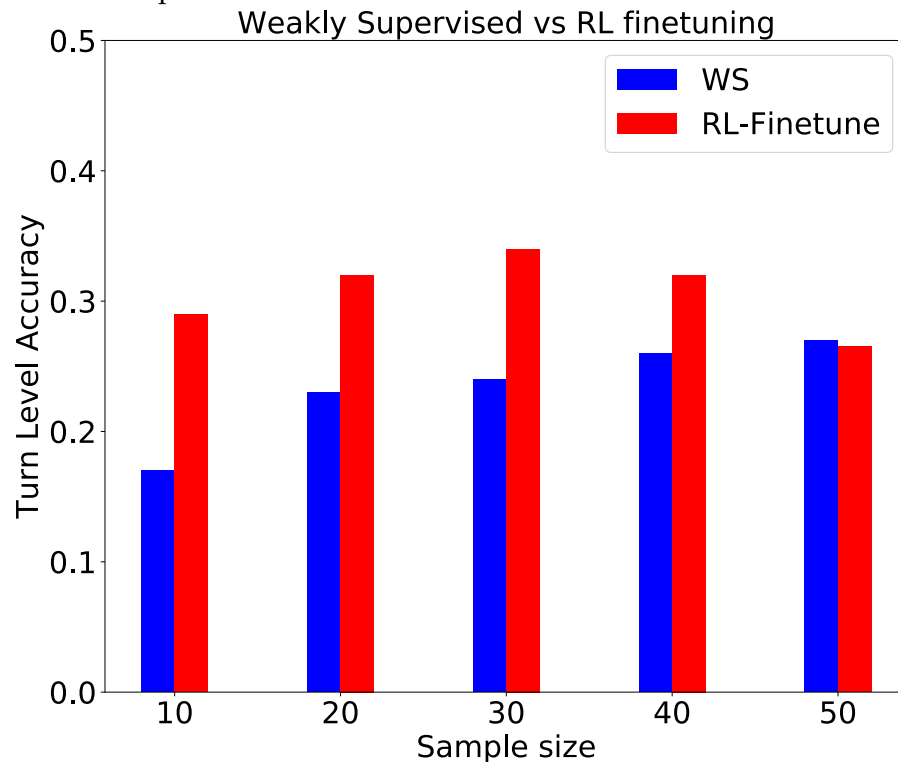


Figure 4.3: The turn level accuracy of our weakly supervised finetuning compared to finetuning using PG. Performance plateaus after about 50 samples for both methods.



## 4.7.1 Error Analysis

In general we observe lower scores for both the baseline models and in-domain finetuning 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.

System utterance	User utterance	Baseline prediction	PG finetune prediction
N/A	I'm looking for a cheap place to dine, preferably in the centre of town.	<code>inform(area=center)</code> <code>inform(pricerange=expensive)</code>	<code>inform(area=center)</code> <code>inform(pricerange=cheap)</code>
Yes, I have 4 results matching your request, is there a price range you're looking for?	I would like moderate price range please.	<code>inform(pricerange=expensive)</code>	<code>inform(pricerange=moderate)</code>
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.	<code>inform(pricerange=expensive)</code>	<code>inform(pricerange=cheap)</code>

Table 4.3: Comparison of example turn predictions from the MultiWOZ dataset between the baseline model trained on the `HOTEL` domains, and the policy gradient finetuned model. Green indicates a correct prediction whereas red indicates a wrong prediction.

In Table 4.3 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 finetuned 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 finetuned 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.

## 4.7.2 Comparisons to Weak Supervision

We also pose the question of how many annotated dialogues in the target domain are needed before policy gradient finetuning

with dialogue-level rewards is no longer beneficial, compared to finetuning a model trained with turn-level cross entropy. In order to further investigate this, we use our pretrained 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 finetuned each of the models on the RESTAURANT domain using the dialogue-level reward only. The results for these experiments are shown in Figure 4.3. Overall, we find that when we annotate just 10 complete dialogues and then finetune our model using reinforcement learning we still see an increase in performance. We observe that as we increase the sample size  $s$  for our weakly supervised models, finetuning 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.

#### 4.8 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 challenge (DSTC) (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) 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 pretrained 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 number 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 has 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 finetuned 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., 2018; 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) tackled the problem of domain adaptation using RL to learn generic policies and derive domain specific policies. In a similar study, Chen et al. (2018b) 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.

#### 4.9 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 finetuning using our dialogue level reward. Our results show consistent improvements over domain transfer baselines without finetuning, 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 finetuning can further improve supervised models in-domain.

## Part III

# INVESTIGATING FAIRNESS AND TRANSPARENCY IN NLP



## TYPE B REFLEXIVIZATION AS AN UNAMBIGUOUS TESTBED FOR MULTILINGUAL MULTI-TASK GENDER BIAS

---

### 5.1 ABSTRACT

The one-sided focus on English in previous studies of gender bias in NLP misses out on opportunities in other languages: English challenge datasets such as GAP and WinoGender highlight model preferences that are “hallucinatory”, e.g., disambiguating gender-ambiguous occurrences of ‘doctor’ as male doctors. We show that for languages with Type B reflexivization, e.g., Swedish and Russian, we can construct multitask challenge datasets for detecting gender bias that lead to unambiguously wrong model predictions: In these languages, ‘the doctor removed his mask’ is not ambiguous, since the coreferential reading requires a special, non-gendered pronoun, and the gendered, possessive pronouns are anti-reflexive. We present a multilingual, multi-task challenge dataset, which spans four languages and four NLP task and focuses only on this phenomenon. We find evidence for gender bias across all task-language combinations by state-of-the-art models and correlate model bias with national labor market statistics.

### 5.2 INTRODUCTION

A reflexive pronoun is an anaphor that requires a *c*-commanding antecedent within its binding domain (Chomsky, 1991).<sup>1</sup> In languages with *Type B* reflexivization (Heine, 2005), the referent of a reflexive possessive pronoun has to be the subject of the clause, while non-reflexive possessive pronouns (so-called *anti-reflexives*) trigger an interpretation where its referent is *not* the subject; see Table 5.1.

We focus on the subset of those languages in which *anti-reflexive possessive pronouns are gendered, but reflexives are not*. This includes Chinese, Russian, Danish, and Swedish, as well as other Scandinavian and Slavic languages (Battistella, 1989; Bílý, 1981;

---

<sup>1</sup> This means that the antecedent should be in the same sentence, be different from the pronoun and not command it, but any ancestor of the antecedent is an ancestor of the pronoun. This is why in *Lea*<sub>1</sub>'s *sister*<sub>2</sub> *taught herself*<sub>1\*/2/3\*</sub> the pronoun refers to *sister*, not to *Lea* or a discourse referent.

	Type A			Type B		
	1st	2nd	3rd	1st	2nd	3rd
Refl	✓	✓	✓	✓	✓	

Table 5.1: In Type B reflexivization (Heine, 2005), 3rd person pronouns cannot be used reflexively. We are interested in Type B languages with gendered pronouns, and where the non-gendered special (3rd person) reflexive marker has a possessive form.

Kiparsky, 2001).<sup>2</sup> Our motivation for highlighting this particular linguistic phenomenon is that the antecedents of reflexive and anti-reflexive pronouns are grammatically determined; if gender bias leads models (or humans) to predict alternative coreference chains, this violates hard grammatical rules and is thus a clear case of gender bias leading not only to ‘hallucinations’,<sup>3</sup> but to errors. To see this, consider the following examples:

- (1) The surgeon<sub>1</sub> put a book on PRON.POSS.REFL.3RD<sub>1</sub> table. → The book is on the surgeon’s<sub>1</sub> table.
- (2) The surgeon<sub>1</sub> put a book on PRON.POSS.3RD<sub>2</sub> table. ↗ The book is on the surgeon’s<sub>1</sub> table.

Examples (1) and (2) should not be thought of as examples of English, but placeholders for sentences in the languages above since this grammatical distinction is not possible in English: the possessive reflexive ( PRON.POSS.REFL.3RD) and the possessive anti-reflexive ( Pron.Poss.3rd) in these languages would be translated to the same pronoun in English. In Example (1), the reflexive possessive pronoun is co-referential with the grammatical subject (as indicated by subscripts), which leads to the conclusion that the book is now on a table that is associated with the subject, in other words, the *surgeon’s* table. In Example (2), in contrast, when an anti-reflexive possessive pronoun is used, this reading is no longer possible. Instead, Example (2) unambiguously means that the book is on someone else’s table. This distinction is not possible in English where the same pronoun (his/her) would be used in both Examples (1) and (2):

<sup>2</sup> This rules out languages such as German and French, where the reflexive (e.g., *sich* and *se*) does not have a possessive form (Steinbach, 1998). We focus on the reflexive and anti-reflexive *possessive* forms rather than pure reflexives, since they occur more freely, i.e., not only in the context of reflexive verbs, and they are thus more likely to interact with implicit gender assumptions.

<sup>3</sup> We use the word *hallucination* to refer to gender bias leading models to infer gender without evidence; see Tian et al. (2020) for a similar use of the term in abstractive summarization.



*The surgeon put a book on his table*, which is therefore ambiguous between a disjoint and a coreferent reading.

Here is why the interaction with implicit gender assumptions happens: Introducing a new referent in a discourse comes at a small cost if the referent is not already salient (Grosz, Joshi, and Weinstein, 1995). In other words, while Example (2) is grammatically unambiguous, language users may occasionally be willing to violate grammatical constraints to avoid the more costly non-coreferential reading, if the meaning of the grammatically correct disjoint reading does not align with their expectations about the world.<sup>4</sup> If a masculine possessive pronoun is used in this example, this aligns with a prevalent stereotype that surgeons are men; although in the US, in reality, only 62.1% are.<sup>5</sup> In other words, language users may be *more likely to prefer the ungrammatical reflexive reading if the gender of the anti-reflexive possessive pronoun matches their (possibly gender-stereotypical) expectations about the referent of the subject*, in this case, *the surgeon*. Such a reading is, however, clearly not intended, and this is an example of when gender bias prohibits effective communication. We will explore to what extent NLP models for languages with Type B reflexivization exhibit a similar bias, leading to wrong predictions, and correlate such predictions to labor market gender statistics for analysis (in section 5.6).

The challenge dataset that we present here consists of examples such as the one above and is intended as a diagnostic of implicit gender assumptions in NLP models. It is applicable across four languages (Danish, Russian, Swedish, and Chinese) and four NLP tasks: Natural Language Inference (NLI), Machine Translation (MT), coreference resolution, and Language Modeling (LM). We will, for example, be interested in whether models are more likely to produce errors when the anti-reflexive pronouns – PRON.POSS.3RD in Example (2) – exhibit the gender that is implicitly associated with the entity in the subject position, i.e., *surgeon*. As should be clear by now, the challenge dataset is fundamentally different from previously introduced challenge datasets in that it focuses on a single linguistic phenomenon that exists across many languages (Cohen, 1973; Honselaar, 1986; Lødrup, Butt, and King, 2011; Stoykova, 2012) and includes **four languages and four tasks**, and because it focuses on gender bias leading to prediction errors rather than ‘hallucinations’, i.e., unwarranted disambiguations. To the best of our knowledge, the dataset introduced below is in this way the first of its kind.

<sup>4</sup> Note that this is *not* a conflict between syntax and semantics, such as, for example, those studied in Kos et al. (2010), but a conflict between syntax, on the one hand, and belief bias and pragmatics.

<sup>5</sup> <http://www.bls.gov/cps/cpsaat11.htm>

**CONTRIBUTIONS** We present a multilingual, multi-task challenge dataset focusing on a specific linguistic phenomenon found in some Scandinavian, Slavic, and Sino-Tibetan languages, namely *gendered possessive anti-reflexive pronouns* in combination with non-gendered possessive reflexive pronouns. We show, by designing multilingual example generation templates by hand, how this phenomenon can interact with gender assumptions in interesting ways. This results in a unique challenge dataset, which we use to detect and quantify gender biases in state-of-the-art and off-the-shelf models across several tasks, including [MT](#), [NLI](#), coreference resolution, and [LM](#). Unlike all other previous challenge datasets focusing on gender bias, our examples quantify *to what extent gender bias in models leads to prediction errors*, rather than unwarranted disambiguation. Data and code is available at <https://github.com/anavaleriagonzalez/ABC-dataset>

### 5.3 THE ANTI-REFLEXIVE BIAS CHALLENGE

The Anti-reflexive Bias Challenge ([ABC](#)) dataset consists of challenging examples in four different languages for four different [NLP](#) tasks. The examples are designed to force humans and models to align with either widespread gender assumptions or hard grammatical rules. Note, again, that this is in sharp contrast to other gender bias challenge datasets, where gender biases lead to biases in semantic disambiguation, but do not interact with grammatical constraints. Our approach is similar to previous work in other respects.

Similarly to Rudinger et al. (2018) and other recent challenge datasets, [ABC](#) relies on hand-written templates, which are used to generate examples in conjunction with lists of occupations. We use the 60 occupations listed in Caliskan, Bryson, and Narayanan (2017) containing statistics about gender distribution across professions, taken from the U.S. Bureau of Labor Statistics. Specifically, we generate a base set of 4,560 sentences from 38 templates, two tenses (present and past), and 60 occupations. The 38 templates vary the position of the pronouns, e.g.:

- (3) The OCCUPATION lost PRON.POSS.3RD wallet at the house.
- (4) The OCCUPATION lost the wallet at PRON.POSS.3RD house.

where PRON.POSS.3RD, in this case, is a place holder for anti-reflexive and reflexive third-person pronouns. Our templates only include transitive verbs.

In our [LM](#) experiments, we predict the pronoun in question. For [NLI](#) and coreference, we introduce three variations of each datapoint (possessive masculine, possessive feminine

(anti-reflexive) pronouns and the non-gendered reflexive pronoun). This leads to a total of 13,680 examples for each language. For NLI, we use these as premises and add possible entailments to our templates. See Examples (1) and (2). For MT, we use the English versions of Examples (3) and (4) as source sentences, with feminine and masculine third-person pronouns. This leads to 9,120 translation problems. Native speakers manually verified and corrected all templates and sample examples for all tasks. Appendix A.1.1 shows examples of the four tasks in the four languages. We discuss each task in detail below.

**NLI** Examples (1) and (2) illustrate the entailment phenomenon that we are interested in. Reflexive possessive pronouns are coreferential with their subjects, which leads to the interpretation that the book is on the surgeon’s table. Anti-reflexive pronouns, on the other hand, prevent this reading and leads to an interpretation that a new discourse entity – another person – exists and that the book is located on that person’s table.

The general form of our inference examples is as follows:

- (5) OCCUPATION.DEF<sub>1</sub> [ VERB PHRASE] PRON.POSS.REFL.3RD<sub>1</sub> OBJECT PREP NOUN.DEF. → OCCUPATION.DEF.POSS<sub>1</sub> OBJECT [ VERB PHRASE.PASSIVE] PREP NOUN.DEF.
- (6) OCCUPATION.DEF<sub>1</sub> [ VERB PHRASE] PRON.POSS.3RD<sub>2</sub> OBJECT PREP NOUN.DEF. ↛ OCCUPATION.DEF.POSS<sub>1</sub> OBJECT [ VERB PHRASE.PASSIVE] PREP NOUN.DEF.

We will primarily be interested in the rate at which state-of-the-art NLI models (wrongly) predict examples of the form in Example (5) to be cases of entailment, and how this depends on whether the possessive pronoun Pron.Poss is masculine or feminine. To generate examples of this form, we translate one prototype example and then identify the variables in the output example. We also make sure to check that there are no morpho-syntactic dependencies, e.g., agreement, between these variables. We then generate all possible examples and have native speakers manually verify the correctness of samples of the generated examples.

**MACHINE TRANSLATION** For MT, we are interested in the way that gender assumptions play a role in the resolution of the gendered possessive pronoun in the source language. As an example, when translating the phrase *The doctor put the book on her table*, an English-Danish translation system would likely generate one of the following two options, a reflexive reading and an anti-reflexive one:

- (7) Lægen lagde bogen på *sit* bord

doctor. DEF put book. DEF on PRON.POSS.REFL.3RD table

(8) Lægen lagde bogen på *hendes* bord

doctor. DEF put book. DEF on PRON.POSS.3RD table

While [ABC](#) focuses on translating *from English*, it holds that similarly, if we translate the Danish sentence *mekanikeren har brug for sine.REFL værktøjer til at arbejde*, which uses a gender-neutral reflexive possessive pronoun *sine*, into English, the model will have to choose between two possible, correct translations:

(9) The mechanic needs *his* tools to work

(10) The mechanic needs *her* tools to work

The [MT](#) section of the [ABC](#) dataset consists of translations from English sentences with gendered possessive pronouns into one of the four target languages (Danish, Russian, Swedish, and Chinese). For a single occupation on the list, this would correspond to two English sentences (masculine and feminine possessive pronoun) per template. We quantify to what extent models translate English source sentences with possessive masculine or feminine pronouns into target sentences with reflexive pronouns.<sup>6</sup>

**COREFERENCE RESOLUTION** For coreference resolution, we generate variants of our templates in the four target languages with each of the gendered anti-reflexives and the reflexive pronoun. That is , for a sentence such as:

(11) The firefighter placed *her/his* shoes in the closet

we generate the following examples for Danish:

(12) Brandmanden placerede *hendes* sko i skabet ( Fem)

(13) Brandmanden placerede *hans* sko i skabet ( Masc)

(14) Brandmanden placerede *sine* sko i skabet ( Refl)

<sup>6</sup> In the context of examples such as Example (9) and (10), using an anti-reflexive pronoun in the target translation may seem more like a hallucination than violating grammatical constraints, and we acknowledge that in [MT](#), as well as in [LM](#), the difference concerning existing gender bias challenge datasets is less pronounced than with [NLI](#) and coreference resolution. Nevertheless, note that the model not only hallucinates a gender attribution, but also co-referentiality, making it relatively simple to construct semantically impossible examples, e.g., *The mechanic needs his tools, but not his own tools*. Furthermore, introducing a new referent without evidence also violates pragmatic economy principles (Gardent and Webber, 2001; Grosz, Joshi, and Weinstein, 1995). Google Translate incorrectly translates into a sentence with two reflexive pronouns (violating the semantic principle of bivalence).

In Examples (12) and (13), the use of anti-reflexive pronouns *hans* or the feminine anti-reflexive *hendes* means the shoes placed in the closet belong to someone other than the firefighter. In our coreference resolution experiments, we are thus interested in how often models wrongly link the anti-reflexive pronouns (*hans/hendes*) to the occupation. Such predictions violate grammatical constraints and are clear examples of gender assumptions overwriting morpho-syntactic evidence.

**LANGUAGE MODELING** For LM, we are interested in how likely the models are to predict a gendered anti-reflexive possessive pronoun when the original sentence contains a reflexive pronoun. In:

(15) Brandmanden placerede *sine* sko i skabet ( Refl)

we compute the sentence perplexity replacing the reflexive pronoun *sine* with a feminine anti-reflexive (*hendes*) or masculine (*hans*) anti-reflexive pronoun. A difference in perplexity reveals a gender bias, and if the model prefers an anti-reflexive reading, this possibly leads to a grammatically incorrect sentence.<sup>7</sup>

## 5.4 EXPERIMENTS

We are interested in the gender associations that *existing* models make. Because of this, we take off-the-shelf translation models and language models. As there were not any state-of-the-art models already pre-trained for coreference in the languages of interest, we train a state-of-the-art architecture for coreference resolution on languages where we could obtain data. To be able to evaluate NLI models on the target languages, we fine-tune a pretrained model for this task.

As previously found in (Rudinger et al., 2018), gender biases in models tended to correlate with labor statistics of the percentage of females in each occupation according to Bureau of Labor Statistics 2015<sup>8</sup> released with Caliskan, Bryson, and Narayanan (2017). We correlate our findings with these statistics as well as national statistics.

**NLI** NLI is originally a three-way classification task. Given two sentences; a premise and a hypothesis, the system classifies the relation between them as entailment, contradiction, or neutral. Since ABC is only intended for diagnosing gender bias in off-the-shelf models, and not for training models, we only

<sup>7</sup> See also the footnote above on whether our machine translation examples diagnose model ‘hallucinations’ or unambiguous prediction errors.

<sup>8</sup> <http://www.bls.gov/cps/cpsaat11.htm>

consider the entailment relation. If the premise contains a reflexive pronoun, the true class is entailment, and if the premise contains a masculine or feminine pronoun it is not entailment.

Crosslingual NLI (XNLI) (Conneau et al., 2018b) is a manual translation of the English NLI data into 15 languages. Chinese and Russian are among them and we benchmark the model on the XNLI test set. Singh et al. (2019) extend the XNLI train set to a wider set of languages, including Danish and Swedish but there is not test set for benchmarking. We use cross-lingual language model pre-training (XLM) (Conneau and Lample, 2019), i.e., we fine-tune on English NLI training data. For Chinese and Russian, we use a publicly available implementation<sup>9</sup> of the XLM-15 model (Conneau and Lample, 2019) and fine-tune it using a batch size of 4 and a learning rate of 0.000005 for 35 epochs, which led to the best performance on the XNLI development set. For Danish and Swedish, we use the XLM-100 model, which we fine-tune for 28 epochs.

**MACHINE TRANSLATION** For MT, we evaluate models for English → {Danish, Russian, Swedish, Chinese} to assess how often they predict the non-gendered reflective possessive pronouns when the source possessive pronoun is masculine versus feminine. For all languages, we report the performance of Google Translate. Additionally, for the languages where an off-the-shelf, near-state-of-the-art system was publicly available, we also report performance. For Chinese, we use the pre-trained models provided by Sennrich et al. (2017)<sup>10</sup> (E-WMT). For Russian, we use the winner system of WMT19 (Ng et al., 2019), which is provided as part of the Fairseq toolkit (F-WMT).<sup>11</sup>

**COREFERENCE RESOLUTION** We train coreference resolution models for Chinese and Russian using the model and code of Joshi et al. (2019). For Chinese, we use the Chinese version of Ontonotes as our training data, which is made up of about 1800 documents for training. For Russian, we use the RuCor corpus (Ju et al., 2014), which is small, containing only 181 documents total, but has been used to train coreference models for Russian before (Ju et al., 2014; Sysoev, Andrianov, and Khadzhiiskaia, 2017). The task consists of predicting the spans that make up a coreference cluster. We train the model using the hyperparameters specified in the source code<sup>12</sup>. We use a maximum

<sup>9</sup> <https://github.com/facebookresearch/XLM>

<sup>10</sup> <https://github.com/EdinburghNLP/nematus>

<sup>11</sup> <https://github.com/pytorch/fairseq/tree/master/examples/wmt19>

<sup>12</sup> <https://github.com/mandarjoshi90/coref/blob/master/experiments.conf>

















Task	Lang	System	Benchmark	ABC	Significance
NLI		XLM-100	–	0.380	✓
		XLM-15	0.736	0.370	✓
		XLM-100	–	0.362	✓
		XLM-15	0.742	0.330	✓
MT		Google Translate	0.204	0.395	✓
		Google Translate	0.260	0.406	✓
		F-WMT	0.268	0.421	✓
		Google Translate	0.211	0.422	✓
		Google Translate	0.460	0.594	†
E-WMT		0.360	0.194	✓	
Coref		e2eCoref-BERT	0.602	0.090	†
			0.630	0.600	✓
LM		BERT	2.4	11.4	✓
			3.9	13.4	✓
			1.2	11.2	✓
			6.7	22.1	✓

Table 5.2: Gender Bias Results. Performance on benchmarks and ABC. ✓: Pearson’s  $\rho$  of error  $\Delta$  on sentences with feminine pronouns and % of women in corresponding occupations significant ( $p < 0.01$ ); see S for a discussion of the statistics. †: Systems insensitive to variation in pronouns.

segment length of 128. See Appendix A.1.1 for statistics of the coreference resolution datasets used for training. While we do not have coreference resolution systems we can evaluate for Danish and Swedish, we include challenge examples for these languages that can be used to detect bias in future systems for these languages.

LANGUAGE MODELING For our LM experiments, we use the pretrained BERT masked LM architecture (Devlin et al., 2019a). We turn pronoun prediction into a Cloze task (Taylor, 1953). Specifically, we use Chinese BERT (for Chinese) and multilingual BERT for Russian, Danish, and Swedish.<sup>13</sup> The overall perplexities of these models on our challenge examples are low; again, this is because of the simple vocabulary and constructions used in the examples. We nevertheless see a strong gender bias in the language models, especially for Danish and Chinese.

<sup>13</sup> <https://github.com/google-research/bert/blob/master/multilingual.md>

## 5.5 RESULTS

Our evaluation results are found in Table 2, with results on Danish (🇩🇰), Russian (🇷🇺), Swedish (🇸🇪), and Chinese (🇨🇳), and for [MT](#), [NLI](#), coreference resolution (Coref), and [LM](#).

**NLI** For [NLI](#), the XLM models generally over-predict entailment for anti-reflexive pronouns. The models perform well on benchmark data, e.g., 0.742 on the Chinese [XNLI](#) test set, but much worse (0.330) on our challenge examples. For Chinese and Danish, the models perform slightly better on sentences with masculine anti-reflexive pronouns, whereas they perform slightly better on sentences with feminine anti-reflexives in Russian and Swedish. For all four languages, we see significant negative correlations between relative error increase on sentences with feminine pronouns and the ratio of women in the corresponding occupations; see section 5.6 for a discussion of the statistics. This suggests that the very poor performance numbers on sentences with anti-reflexive pronouns is, in part, the result of gender bias.

**MACHINE TRANSLATION** For [MT](#), we also observe strong negative correlations, suggesting gender bias. In the manual analysis of the output translations, we see a very clear pattern that English masculine possessive pronouns are more likely to translate into reflexive pronouns in the target languages than feminine possessive pronouns. For Danish, 93.7% of masculine pronouns were translated into reflexives, whereas only 72.9% of feminine pronouns were. For Russian, the two systems were consistent in this respect and both translated 69.3% of masculine pronouns and 18.1% of feminine pronouns into reflexive pronouns. For Swedish, the numbers were 90.0% and 73.1%, respectively. For Chinese, where the reflexive pronoun is used less frequently,<sup>14</sup> the [MT](#) models only produced a few translations with reflexive pronouns (for masculine source pronouns).

These differences are not reflected in BLEU scores, and in our correlations we correlate the increase in pronoun translation errors for source sentences with feminine pronouns and the ratio of women in the corresponding occupations. In general, our models achieve high BLEU scores on our challenge examples, which are all syntactically simple and use simple, in-vocabulary words.

<sup>14</sup> The systems are trained on a combination of traditional and simplified Chinese; the latter variant does not include the reflexive pronoun.



**COREFERENCE RESOLUTION** For coreference resolution, we observe clear performance differences between our Chinese and Russian models. This possibly reflects the fact that the Russian model was trained on a very small dataset and is less likely to generalize. For both models, we observe a clear bias towards clustering masculine anti-reflexive pronouns with their grammatical subjects, despite how this violates grammar. The Chinese model, which exhibits a strong gender bias, errs on 17% of sentences with masculine anti-reflexive pronouns, and on 14.6% of sentences with feminine anti-reflexives. For Russian, the differences are small, but note the model is trained on limited data, e.g., 140 documents. Out of around 13,000 examples, the model only predicts clusters for 475 pronouns, and 400 of those are in reflexive case. The remaining 75 are masculine (0 feminine). In other words, we see a similar tendency to Chinese, but since the overall performance is poor, and the model is in general rather insensitive to differences in pronouns, we do not include correlation results.

**LANGUAGE MODELING** Moreover, for LM, we observe a consistent bias when predicting a masculine pronoun in place of a reflexive for all languages. These differences are higher for Chinese and Russian. We are not interested in the model's ability to generate a particular pronoun, the more interesting observation is whether the perplexities for sentences containing masculine possessives are lower than for predicting feminine possessives when forcing the model to predict these in place of a reflexive. Our results show that perplexities are lower for masculine possessives in all languages with the biggest differences of 3.7 sentence perplexity for Russian.

## 5.6 ANALYSIS: BIASED STATISTICS?

We used occupations from Caliskan, Bryson, and Narayanan (2017) in creating our template data; this database also includes U.S. occupation statistics. In our results in Table 5.2, however, we rely on national statistics instead, but how much of a bias would it be to rely on the original American statistics? In this section, we explain how we collected the national statistics and show how they strongly correlate with the American statistics, but also that national statistics are slightly better at detecting gender bias:

Our Danish labor market statistics come from Larsen, Holt, and Larsen (2016), as well as Statistics Denmark<sup>15</sup> and Bevægelses-

---

<sup>15</sup> [www.dr.dk/nyheder/indland/](http://www.dr.dk/nyheder/indland/)

registeret,<sup>16</sup> which is a national database over authorised health staff. Some numbers (paramedic, scientist and receptionist) are based on graduation statistics. The Russian labor market statistics were mostly obtained from the Federal State Statistic Service.<sup>17</sup> For occupations not contained on this website we obtained the numbers from separate sources such as the Center of Fire Statistics (CFS) of International Association of Fire and Rescue Services (CTIF)<sup>18</sup> and the Organisation for Economic Cooperation and Development's statistics website<sup>19</sup>. We obtain most of our Swedish labor market statistics from Statistics Sweden (SCB).<sup>20</sup> We use the most recent statistic from 2017, which considers people aged 16-64 (Eriksson and Nguyen, 2019). For clerk and worker, we found labor market statistics in SCB (2018). For medical jobs, we used member statistics by Swedish Medical Association (SLF) from 2016.<sup>21</sup> Finally, we obtain statistics for China from National Bureau of Statistics (2004), which is based on census data from 2000.<sup>22</sup>

While labor statistics correlate strongly across countries (Table 5.1), U.S. statistics are not universal; e.g., almost all pathologist in the U.S. are women (97.5%), whereas the percentage for Denmark is 60%. In the U.S. and Sweden, the painter profession is very male-dominated, like mechanic and electrician (5.70% and 8% women, respectively), whereas in Russia, 57.0% of painters are women.

**CORRELATION RESULTS** To assess the potential bias of using U.S. labor market statistics in multilingual experiments, we correlate the gender bias of models for language  $l$  with labor statistics from the U.S. and the country in which  $l$  is a national language, i.e., we correlate performance differences on Swedish ABC examples with both U.S. and Swedish labor statistics, Danish ABC examples with U.S. and Danish labor statistics, etc. We do so for the subset of occupations, where national gender statistics are available:

**NLI.** Correlations were stronger with national rather than U.S. statistics for Danish and Swedish (-0.35 vs. -0.28; -0.36 vs. -0.34).

<sup>16</sup> [www.esundhed.dk/home/registre/](http://www.esundhed.dk/home/registre/)

<sup>17</sup> [eng.gks.ru/](http://eng.gks.ru/)

<sup>18</sup> [www.ctif.org/](http://www.ctif.org/)

<sup>19</sup> [stats.oecd.org/](http://stats.oecd.org/)

<sup>20</sup> [www.scb.se/](http://www.scb.se/)

<sup>21</sup> [slf.se/app/uploads/2018/04/](http://slf.se/app/uploads/2018/04/)

<sup>22</sup> We did not find reliable gender statistics for all occupations for all countries, but for 51 (Denmark), 50 (Sweden), 38 (Russia), and 10 (China) occupations. One reason was a mismatch between how gender statistics are reported in official reports, including how jobs are grouped. We release the numbers we were able to collect and will continually work on obtaining more statistics.

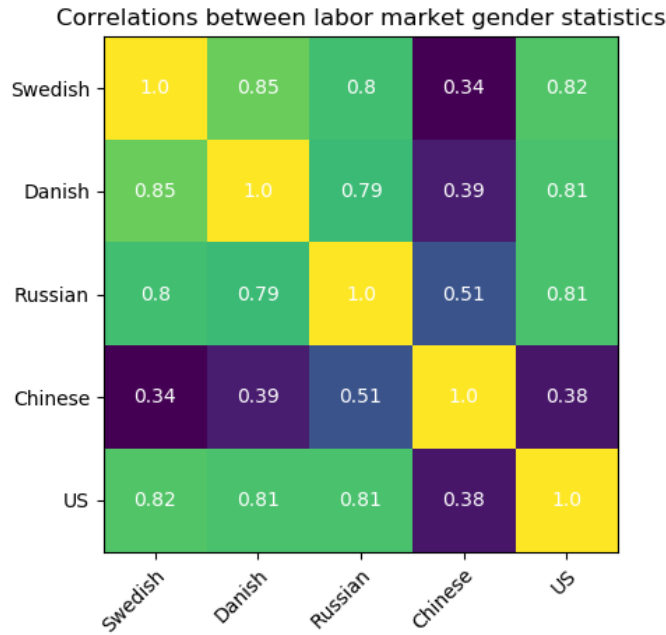


Figure 5.1: Correlations between collected labor statistics. Numbers  $> 0.7$  are significant ( $p < 0.01$ ).

**Machine Translation.** Correlations were stronger with national rather than U.S. statistics for Russian and Swedish (-0.31 vs. -0.20; -0.31 vs. -0.14).

**Coreference Resolution.** For coreference, we were able to correlate only the results for Chinese due to the fact that the coreference model for Russian only predicted clusters for sentences with male pronouns. The correlations with U.S. and Chinese labor market statistics were not significantly different because we only had statistics for 10 occupations.

**Language Modeling.** Correlations were stronger with national rather than U.S. statistics on average, but not significantly so.

## 5.7 RELATED WORK

The ABC dataset is not first to focus on pronouns and gender bias. The UD English-Pronouns<sup>23</sup> (Munro, 2020), a manually constructed, gender-balanced benchmark of English sentences with pronouns, was motivated by the observation that the genitive pronoun *hers* only occurs three times in the English Universal Dependencies (Nivre et al., 2017). The gendered, ambiguous pronoun (GAP) dataset (Webster et al., 2019) is a coreference resolution dataset of human-annotated ambiguous pronoun-name examples from English Wikipedia. Prates, Avelar, and

<sup>23</sup> [universaldependencies.org/](http://universaldependencies.org/)

Lamb (2018) constructed a translation challenge dataset of simple sentences in gender-neutral languages such as Hungarian and Yoruba and English target sentences such as *he/she is an engineer* to estimate gender biases in MT. Both these challenge datasets focus on gender hallucinations, not unambiguous errors induced by gender bias. Some of our examples share similarities with the English WinoGender schema (Rudinger et al., 2018). Consider the following minimal pair of Winograd schema taken from their paper:

- (16) The paramedic<sub>1</sub> performed CPR on the passenger<sub>2</sub> even though PRON<sub>1</sub> knew it was too late.
- (17) The paramedic<sub>1</sub> performed CPR on the passenger<sub>2</sub> even though PRON<sub>2</sub> was already dead.

In the Winograd schema, the context, i.e., the second clause, is supposed to disambiguate the pronoun on semantic grounds. In Example (16), the pronoun refers to the paramedic, because the patient is unlikely to know whether CPR is too late. In Example (17), the pronoun refers to the patient, because it is impossible to perform CPR if you are dead. Our examples, in contrast, do not disambiguate pronouns on semantic grounds, and this is why we are interested in reflexive possessive pronouns: they always refer to the subject, and their anti-reflexive counterparts never do, so there is no grammatical ambiguity. The disadvantage with semantic disambiguation, we argue, is that it ultimately becomes a subjective competition of belief biases. It is generally impossible to perform CPR if you are dead, but special cases exist:

- (18) Dr Jones<sub>1</sub> has turned into a zombie! He<sub>1</sub> performed CPR on the passenger even though he<sub>1</sub> was already dead.

The ABC dataset evaluates to what extent gender bias leads to unambiguous NLP errors not based on semantic grounds. Finally, Zhao et al. (2018) also include English examples with reflexive pronouns that can be resolved on syntactic grounds, such as:

- (19) The secretary called the physician and told *him* about a new patient.

This construction, however, is less interesting than the reflexive possessive pronominal construction, since in this case, pronouns are always co-referential with the object position, regardless of the pronoun. In sum, the ABC challenge dataset is, to the best of our knowledge, the first dataset to focus on cases where gender bias leads to unambiguous errors; it is also the first multilingual, multitask gender bias challenge dataset, and the first to focus on anti-reflexive pronouns.

## 5.8 CONCLUSION

In this work we have introduced the Anti-reflexive Bias Challenge (ABC) dataset for multilingual, multi-task gender bias detection, the first of its kind, including four languages and four tasks: MT, NLI, coreference resolution and LM. The ABC dataset focuses on a specific linguistic phenomenon that does not occur in English but is found in languages with *Type B reflexivization*: namely, anti-reflexive gendered pronouns. This phenomenon is shown to be useful for exposing unambiguous gender bias, because it quantifies to what extent gender bias leads to prediction errors, in contrast to just *unwarranted disambiguations* ('hallucinations'). The problem of anti-reflexive gendered pronouns has, to the best of our knowledge, not received attention before in the NLP literature, which tends to focus heavily on English (Bender and Friedman, 2018). Our evaluations of state-of-the-art models across the four tasks generally reveal significant gender biases leading to false predictions. Additionally, we find that for some tasks, these associations are more in line with national labor market gender statistics than with U.S. statistics, revealing another way that anglocentric biases can prohibit the detection of gender biases in our models.



## THE REVERSE TURING TEST FOR EVALUATING INTERPRETABILITY METHODS ON UNKNOWN TASKS

---

### 6.1 ABSTRACT

The Turing Test evaluates a computer program’s ability to mimic human behaviour. The Reverse Turing Test, reversely, evaluates a human’s ability to mimic machine behaviour in a forward prediction task. We propose to use the Reverse Turing Test to evaluate the quality of interpretability methods. The Reverse Turing Test improves on previous experimental protocols for human evaluation of interpretability methods by a) including a training phase, and b) masking the task, which, combined, enables us to evaluate models independently of their quality, in a way that is unbiased by the participants’ previous exposure to the task. We present a *pilot* human evaluation of [LIME](#) across five [NLP](#) tasks and analyze the effect of masking the task in forward prediction experiments. Additionally, we demonstrate a fundamental limitation of [LIME](#) and show how this limitation is detrimental for human forward prediction in some [NLP](#) tasks.

### 6.2 INTRODUCTION

Machine learning models have tremendous impact on our daily lives, from information storing and tracking (i.e. Google Search and Facebook News Feed), as well as on other scientific disciplines. Modern-day [NLP](#) models, for example, are complex neural networks with millions or billions of parameters trained with multiple objectives and often in multiple stages (Devlin et al., [2019b](#); Raffel et al., [2019](#)); they are often seen for that reason, as black boxes whose rationales cannot easily be queried. In other words, we are increasingly relying on models that we do not understand or cannot explain, in science, as well as in our daily lives. Model interpretability, however, is desired for several reasons: Humans often ask for the motivation behind advice, and in the same way, users are likely to trust model decisions more if they can ask for the rationale behind them. Model interpretability enables us to inspect whether models are fair and unbiased, and it enables engineers to detect when models rely on mere confounds. Combatting this type of overfitting will lead to more robust (or less error-prone) decision making with

better generalization to unseen data (and, hence, safer model employment).

Recent years has seen a surge in work on post-hoc interpretability methods for neural networks. See Murdoch et al. (2019) for a brief survey. Unfortunately, there is little consensus on how to compare interpretability methods. Some benchmarks have been introduced (DeYoung et al., 2020; Poerner, Schütze, and Roth, 2018; Rei and Søgaard, 2018), but some of these are flawed, and they are all only applicable to some of the proposed interpretability methods. See section 8.3 for discussion. In our view, a more promising approach to evaluating interpretability methods is by **human forward prediction** experiments. Nguyen (2018a) presented the first evaluations of **LIME** (Ribeiro, Singh, and Guestrin, 2016a) for sentiment analysis using human subjects through a series of Mechanical Turk experiments. Their study had two limitations: (a) They did not allow for a training phase for the human participants to learn model idiosyncracies, and participants instead had to rely on the assumption that the model was near-perfect. (b) Since the participants thus had to rely on their own sentiment predictions, their evaluations are biased by their beliefs about the sentiment of the input documents. (Hase and Bansal, 2020) recently presented evaluations of **LIME** with human participants that involved a training phase, enabling them to predict *poor* model behavior, and thereby addressing limitation (a), but they still only included known tasks for which forward prediction is biased by the participants' own beliefs. This paper aims to fill this gap.

**CONTRIBUTIONS** This work presents a simple-yet-insightful method for evaluating interpretability methods based on short experiments with human participants. Our experiments differ from previous work in a very important way: our evaluation of interpretability methods involves conditions where human subjects are less likely to rely on their belief biases. More specifically, we evaluate **LIME** (Ribeiro, Singh, and Guestrin, 2016a) across five **NLP** tasks in a Latin Square design (Shah and Sinha, 1989), including three tasks which were *kept secret* to our participants. We argue that *keeping the tasks secret to the participants makes the evaluation of interpretability methods more reliable* and investigate the impact of this difference in experimental design. Additionally, we also point out a weakness of **LIME** - namely, that its input/output dimensions are occasionally orthogonal to the relevant dimensions for interpretability. We include a task in which this happens and show how detrimental interpretability methods can be in such cases.



### 6.3 HUMAN BIAS IN FORWARD PREDICTION

One thing sets our experiments in this paper apart from previous evaluations of interpretability methods by A/B testing with human forward prediction (Hase and Bansal, 2020; Nguyen, 2018a): We will (in some cases) present participants with decisions by models trained on tasks that are *unknown* to the participants. In other words, humans are simply asked to predict  $y$  from  $x$ , with no prior knowledge of the relation that may exist between them, beyond an initial training phase. Three different cognitive biases are particularly important for motivating and analyzing our experimental design:

**BELIEF BIAS** An effect where someone’s evaluation of the logical strength of an argument is biased by the plausibility of the conclusion (Klauer, Musch, and Naumer, 2000). In human forward prediction of model behavior, this happens when the plausibility of the conclusion, e.g., this review is positive, biases the subject’s evaluation of her own conclusions, e.g., the model will predict this review is negative, because it includes this or that term. We argue that it is particularly important to evaluate interpretability methods with human forward prediction on *unknown* tasks to avoid belief bias.

**CONFIRMATION BIAS** This bias occurs when individuals seek information which supports their prior belief while disproportionately disregarding information that challenges this belief (Mynatt, Doherty, and Tweney, 1977). In our context, such a bias could, for example, lead subjects that already classified a document in one way to disregard LIME mark-up. In the extreme, confirmation bias could cancel out any effect of interpretability methods in human forward prediction, but our results show that in practice, LIME has a strong (positive or negative) effect on human forward prediction.

**CURSE OF KNOWLEDGE** This is the phenomenon when better-informed people find it extremely difficult to think about problems from the perspective of lesser-informed people (Ackerman, Pipek, and Wulf, 2003). In our case, the model plays the role of a lesser-informed agent. We believe the curse of knowledge amplifies belief bias and makes it very hard for participants to *unlearn* their prior knowledge of the underlying task relation.

Our experimental design is motivated by a desire to remove belief bias in our forward prediction experiments. Belief bias can interact with human forward prediction in a number of ways, e.g., making participants less confident about predictions that do

align with their beliefs, or leading them to ignore explanations that are inconsistent with their beliefs.

Additionally in comparison to previous work, we present experiments that highlight one of the weaknesses of LIME, by presenting participants with LIME explanations for a task where LIME is observed to provide poor explanations. In this scenario, explanations have a detrimental effect. We believe this is essentially an anchoring effect:

**ANCHORING** The tendency to rely too heavily, or *anchor*, on a few pieces of information when making decisions (usually the first pieces attended to) (Zhang et al., 2007). This explains why bad explanations can have detrimental effects on humans' ability to predict system outputs. Since explanations are the first information presented to subjects, this inhibits otherwise available counter-evidence.

#### 6.4 LIME – AND ITS LIMITATIONS

The Local Interpretable Model-agnostic Explanations (LIME) method (Ribeiro, Singh, and Guestrin, 2016a) has become one of the most widely used post-hoc model interpretability methods in NLP. LIME aims to interpret model predictions by locally approximating a model's decision boundary around an individual prediction. This is done by training a linear classifier on perturbations of this example.

Several weaknesses of LIME have been identified in the literature: LIME is linear (Bramhall et al., 2020b), unstable (Elshawi, Al-Mallah, and Sakr, 2019) and very sensitive to the width of the kernel used to assign weights to input example perturbations (Kopper, 2019; Vlassopoulos, 2019), an increasing number of features also increases weight instability (Gruber, 2019), and Vlassopoulos (2019) argues that with sparse data, sampling is insufficient. Laugel et al. (2018) argues the specific sampling technique is suboptimal.

We point to an additional, albeit perhaps obvious, weakness of LIME's: *It can only explain the decisions of a classifier in so far as the decision boundary of the classifier aligns with the feature dimensions of LIME.* In most applications of LIME to NLP problems, the feature dimensions are the input words. That is to say, LIME can only explain the decisions of a classifier if the decision boundary aligns with the dimensions along which the occurrences of words are encoded. LIME can, for example, not explain the decisions of a classifier " 1 if sentence length odd, else 0". In our experiments, we include a task in which a classifier is trained to predict the length of the input sentence (from a low-rank

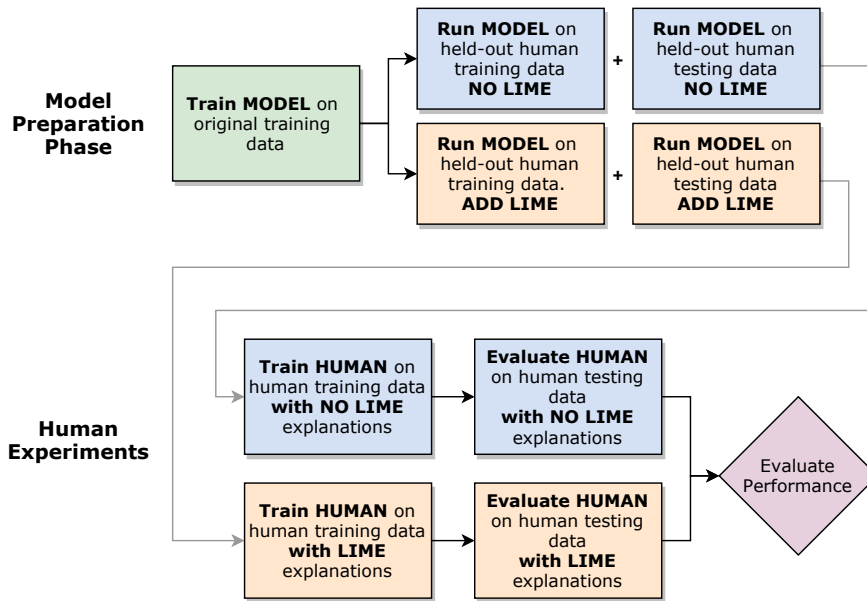


Figure 6.1: Our experimental protocol. For each task, we train our models using standard datasets and evaluate the model on held out training data and testing data to be used for the training and evaluation sessions involving humans. We also extract LIME explanations. In the human experiments phase, the humans train and evaluate in these 2 conditions (LIME explanation or no explanation). Finally, we compare the results.

representation), as a way of evaluating the effect of LIME on human forward prediction, on tasks that LIME is, for this reason, not able to explain.

Examples of real tasks where this limitation is a problem, include, for example, all tasks where sentence length is predictive, including readability assessment (Kincaid et al., 1975), authorship attribution (Stamatatos, 2009), or sentence alignment (Brown, Lai, and Mercer, 1991). We note this limitation is not unique to LIME, but shared among most post-hoc interpretability methods, e.g., *hot flip* (Ebrahimi et al., 2018), attention (Rei and Søgaard, 2018), and back-propagation (Rei and Søgaard, 2018). Other approaches to interpretability such as using influence functions (Koh and Liang, 2017) may have more explanatory power for such problems.

## 6.5 HUMAN FORWARD PREDICTION EXPERIMENTS

The experiments we describe below are examples of the Reverse Turing Test. The test resembles the Turing Test (Horn, 1995; Turing, 1950) in that it focuses on the differences between the behavior of humans and computer programs. In the Reverse Turing Test, we quantify humans' ability to simulate computer

programs, however; rather than computer programs' ability to simulate humans. Specifically, we quantify humans' ability to predict the output of machine learning models given previously unseen examples. The test is defined (for classification models) as follows: The Reverse Turing test is an experimental protocol according to which participants are presented with  $k$  examples of  $\langle \mathcal{I}(\mathbf{x}), \hat{y} \rangle$  pairs, with  $\hat{y} = f(\mathbf{x})$  the labeling of  $\mathbf{x}$  by some unknown machine learning model  $f(\cdot)$ , and  $\mathcal{I}$  is a possibly empty interpretation function, which, in the case of post-hoc interpretability methods, highlights parts of the input, e.g., input words. The training phase is timed. Subsequently, participants are presented with  $m$  unseen examples  $\mathbf{x}_1 \dots \mathbf{x}_m$  and asked to predict  $f(\mathbf{x}_1) \dots f(\mathbf{x}_m)$ . The evaluation phase is also timed. The result of the Reverse Turing test is the accuracy or  $F_1$  of the participants' predictions compared to  $\hat{y}_1, \dots, \hat{y}_m$ , as well as the training and inference times. The test is meant to evaluate the quality of different interpretations,  $\mathcal{I}(\cdot)$  and can be used for evaluation methods, like we do, or for evaluating models or interpretability methods during development (Lage et al., 2018). We believe our test is in some ways more critical than previous, as we are attempting to evaluate interpretability methods more reliably by reducing human belief bias.

### 6.5.1 Tasks and Data

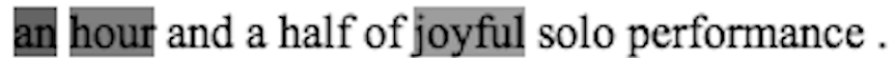
Based on the efforts of 30 annotators, we collected a total of 3000 example annotations in human forward prediction experiments, distributed across five different tasks (two known; three unknown) and two experimental conditions (with and without explanations). The overall experimental protocol is shown in Figure 7.1. All code for preprocessing data, training the models, and the experimental set ups are publicly available at [https://github.com/anavalerialgonzalez/reverse\\_turing\\_test](https://github.com/anavalerialgonzalez/reverse_turing_test).

**KNOWN TASKS** For our known tasks, we focus on two very common text classification tasks: sentiment analysis and hate/offensive speech detection. For sentiment analysis we use the Stanford Sentiment Treebank (SST) (Socher et al., 2013). The SST dataset consists of 6920 documents for training, 872 documents for development and 1820 documents for testing. For hate speech detection, we use the HatEval dataset from SemEval 2019 (Basile et al., 2019). The dataset consists of several binary tasks, however, we focus on the task of detecting presence of hate speech (disregarding which group is being targeted as this is considered a separate task). In total, there are 9000 tweets for training, 1000 for development and 3000 for testing.

**UNKNOWN TASKS** As our unknown tasks, we use 3 of the 10 probing tasks introduced in (Conneau et al., 2018a). The probing tasks were originally designed to evaluate the linguistic properties of sentence embedding models. In this study we are mostly interested in the differences in performance between humans and machines, and are not looking to evaluate linguistic properties of representations in depth, therefore chose only a few of these tasks. The first task is *sentence length prediction* in which the sentences are grouped in 6 bins indicating length in terms of number of words. This task was chosen in order to examine the effect on LIME in a task where LIME offers poor explanations. The second probing task is *tense prediction*, which involves predicting whether the verb in the main clause is present or past tense. The third task is *subject number prediction*, which focuses on predicting whether the subject in the main clause is plural or singular. These last two, are simple tasks where we expect LIME to offer good enough explanations. The training data for each of the probing tasks consists of 100k sentences, 10k sentences for validation and 10k sentences for testing. The sentences are taken from the Toronto Book Corpus (Zhu et al., 2015). More details on data extraction can be found on Conneau et al. (2018a).

### 6.5.2 Classification Model

For training sentiment and hate speech classifiers, we pass as our input pretrained BERT representations (Devlin et al., 2019b) through an LSTM layer (Hochreiter and Schmidhuber, 1997) ( $d = 100$ ) followed by a multi-layered perceptron with a single hidden layer ( $d = 100$ ). We use a learning rate of 0.001 and Adam optimizer. The hyper-parameters were *not* tuned for optimal performance. We use the same architecture for all tasks, except for sentence length prediction. For the sentence length prediction task, we use BERT token representations and pass them through a mean pooling layer followed by a multi-layered perceptron with a single hidden layer ( $d = 100$ ). Both models are trained for 20 epochs. Note also that we do *not* fine-tune the BERT representations. This, together with our hyper-parameters, gives us suboptimal performance, especially on the known tasks, but this was done *on purpose* to make our predictions different from the gold labels for the known tasks, in order to make it possible to quantify participants' belief bias: If results are too close to human performance, it would not be possible to distinguish human forward prediction performance with respect to model predictions from human performance with respect to predicting the true class. Our performance on the unknown probing tasks is comparable to the results in (Conneau et al., 2018a).



an hour and a half of joyful solo performance .

Figure 6.2: Example LIME explanation stripped of model decisions and class probabilities. We turn the images into gray scale to only highlight overall importance and avoid hinting the model’s final decision.

### 6.5.3 Stimulus Presentation

Each human forward prediction experiment consists of a training session where we present the participant with 25 training samples with model predictions, with or without explanations, followed by an evaluation session with 15 testing samples (without model predictions), also with or without LIME explanations. The participant is asked to predict the model’s labeling of these items. We use a Latin Square design (Shah and Sinha, 1989) to control for idiosyncratic differences between our participants. For each of the tasks  $t_i$ , we therefore randomly sample 120 examples, 75 of which we use for training our participants, and 45 of which we use for evaluation. We divide the 75 training samples into groups of 25:  $t_{t_1}, t_{t_2}, t_{t_3}$ . We have three different presentation conditions: no explanation, LIME explanation, or random explanation (for control). For the LIME explanations, we remove information about model decision and present participants with the original LIME output images, after turning them into grayscale in order to avoid revealing the class label. We rely on 500 perturbations of each data sample in order to obtain the top 3 most informative input tokens. See Figure 6.2 for an example of the visual stimuli under this condition. The training sessions are interactive, simulating the test interface, but providing the true answer whenever the participant has provided an initial guess. We shuffle the training sessions at random. The evaluation sets for each task  $t_e$  consist of 45 samples in total, split into chunks of 15:  $t_{e_1}, t_{e_2}, t_{e_3}$ . In the evaluation session, subjects are not provided with the true model responses, to avoid biases from additional training. We divide our participants in three groups, and for each task, the groups are assigned task subsamples in the following Latin Square design:

	$x$	LIME( $x$ )	Control( $x$ )
Subjects <sub>1</sub>	$t_{t_1}, t_{e_1}$	$t_{t_2}, t_{e_2}$	$t_{t_3}, t_{e_3}$
Subjects <sub>2</sub>	$t_{t_2}, t_{e_2}$	$t_{t_3}, t_{e_3}$	$t_{t_1}, t_{e_1}$
Subjects <sub>3</sub>	$t_{t_3}, t_{e_3}$	$t_{t_1}, t_{e_1}$	$t_{t_2}, t_{e_2}$

We include 3 unknown tasks, meaning that no information about the tasks was provided to the participants in advance of



the experiment. Instead, subjects had to try to infer patterns from the data sample, possibly augmented with LIME explanations. For the known tasks, we follow Nguyen (2018a) and Hase and Bansal (2020) and provide subjects with a brief explanation of the task, but emphasize the fact that the participants should predict *model decisions*, not the true labels; and hence, they should avoid being influenced by their own beliefs of whether a text is positive or an instance of hate speech. As in Hase and Bansal (2020), we make sure that true positives, false positives, true negatives, and false negatives are balanced across the training and test data. In total we have 30 participants, all with at least undergraduate education and some knowledge of computer science and machine learning. We collect 3000 human forward predictions: 1800 from training sessions and 1200 from the evaluation sessions. For each condition and item in the evaluation set, we have at least two human forward predictions. Some of the participants gave us optional feedback on strategies they used. This, as well as some examples of our interface can be found in the Appendix.

#### 6.5.4 Pre-Experiment: The Effect of Training on Forward Prediction

In addition to our main experiment with 30 participants, we also performed a human forward prediction pre-experiment with a single participant. In the pre-experiment we compare human forward prediction *with and without training*; we do so to motivate our experimental design, in which we follow Hase and Bansal (2020), but depart from Nguyen (2018a), in including a training phase in which humans can learn the idiosyncracies of the machine learning model. In the pre-experiment, we are only interested in seeing the effects of the training phase for the known tasks. We first ran the experiment without training; then ran the experiment with training. To clearly be able to quantify the effect of our interactive training phase, we only use examples with *false* model predictions in the pre-experiment. For each of the two tasks, sentiment analysis and hate speech detection, we use: (a) 20 distinct examples for evaluation for each of the two conditions; and (b) 25 distinct examples for training for the second experimental condition. Note that since we only use a single human participant in the pre-experiment, controlling for individual differences, we cannot control for the difficulty of data points and use different data points across the two experimental conditions.

The effect of training is positive. On the SST dataset, the accuracy with respect to model predictions ( $\hat{p}$ ) increases from **0.400**

to 0.550;<sup>1</sup> on the HatEval2019 dataset, performance increases from 0.3690 to 0.526. We see this as a very strong motivation for including a training phase. A training phase also makes it possible to perform human forward prediction experiments on tasks that are unknown to the participants, removing any belief bias that may otherwise affect results. We note that a training phase does not necessarily lead to faster inference times. On HatEval, average inference time was reduced from 08:56 to 07:21, but on STS, it increased from 06:24 to 08:53. This suggests that untrained annotators (after a few instances) learn superficial heuristics that enable them to draw fast, yet inaccurate, inferences.

### 6.5.5 Main Experiment: The Effect of LIME on Forward Prediction

We report the results of our main experiment in Table 6.1. Results show that LIME helps, both in terms of accuracy and time, on known and unknown target tasks, except when the decisions boundary does not align with LIME dimensions (Sent Len) (columns 1–4); and that while humans are biased by their beliefs and knowledge of the known tasks, they are *not* biased during unknown tasks, which can be seen by their *decrease* in accuracy with respect to human annotation. We make the following observations:

**THE EFFECT OF LIME ON KNOWN TASKS** This is the standard set-up considered also in previous work (Hase and Bansal, 2020; Nguyen, 2018a); see columns 1–2 and rows 1–2 in Table 6.1. We see that LIME leads to significantly better human forward prediction performance on both tasks. It also leads to (statistically) significantly faster inference times, approximately halving the time participants spend on classifying the test examples. This shows that LIME, in spite of its limitations (§3), is a very useful tool in some cases.

**THE EFFECT OF LIME ON UNKNOWN TASKS** The effect of LIME on human forward prediction accuracy on 2 of the *unknown* tasks is *not* significant. On the two tasks, where LIME provides meaningful explanations (subject number and tense prediction), LIME does lead to smaller reductions in inference time which are not statistically significant. The effect on the participants' accuracy is mixed and insignificant. In addition,

<sup>1</sup> Note that our human participant, without training had lower-than-random accuracy in both tasks. This is not surprising, since we have selected data points on which our model was wrong. Under the influence of belief bias, humans are likely to also classify these wrongly.



Task	Human Acc. ( $\hat{p}$ )		Human Time( $\hat{p}$ )		MODEL ACC. ( $p$ )	Human Acc. ( $p$ )	
	$x$	LIME( $x$ )	$x$	LIME( $x$ )		$x$	LIME( $x$ )
KNOWN TASKS							
SST	0.557	* <b>0.694</b>	03:00	* <b>01:50</b>	0.822	0.767	<b>0.794</b>
HatEval 2019	0.562	* <b>0.715</b>	02:18	* <b>01:10</b>	0.573	<b>0.706</b>	0.609
UNKNOWN TASKS							
Sent Len	* <b>0.470</b>	0.310	<b>05:32</b>	08:15	0.846	* <b>0.612</b>	0.360
Subj Num- ber	<b>0.500</b>	0.430	09:43	<b>08:50</b>	0.901	0.397	<b>0.491</b>
Tense	0.542	<b>0.581</b>	07:02	<b>04:51</b>	0.942	0.449	<b>0.500</b>

Table 6.1: RESULTS FROM MAIN EXPERIMENT. Columns 1–2: accuracy of human forward prediction results on plain input ( $x$ ) or augmented with LIME interpretations (LIME( $x$ )). \*: Significance of  $\alpha < .05$  computed with Mann-Whitney  $U$  test. Columns 3–4: average duration of evaluation sessions (human inference time). Column 5 lists the model accuracies with respect to human gold annotation; which we compare with human accuracies with respect to human gold annotation.

LIME is significantly detrimental on human forward prediction accuracy for the task of sentence length prediction; it also leads to longer inference times, although this difference was not statistically significant. This shows that while LIME is useful in some cases, this is not always the case. We speculate that since LIME explanations are partial, they are only effective when supplemented by (approximately correct) belief bias. If true, this suggests that LIME, even for the tasks that *can* be explained in terms of input words, is, nevertheless, only applicable to tasks that humans have experience with, and when the underlying models perform reasonably well.

**KNOWN AND UNKNOWN TASKS** In general, our participants are much slower at classifying examples when the task is unknown. This shows the efficiency of the belief biases our participants have in sentiment analysis and hate speech classification. The effectiveness of these biases is also demonstrated by the performance gaps between humans and models when comparing their predictions to ground truth labels. To see this, consider columns 5–7 in Table 6.1. Participants, while instructed to *predict model output* ( $\hat{p}$ ), actually significantly outperform our classifier in predicting the true labels (0.706 vs. 0.573)! In contrast, participants perform subject number and tense prediction at chance levels (0.491 and 0.500), while a simple classifier achieves accuracy greater than 0.9 on both tasks. This clearly demonstrates belief bias in human forward prediction experiments.

**HUMAN INFERENCE TIME** In addition to considering performance, we also recorded the time our participants spent on completing the forward prediction tasks. We present the average times of each condition in Table 6.1 with shorter times bolded. We used the Mann-Whitney  $U$  test to determine significance for these, which is also shown in the same table. We plot the total averages in Figure 6.3. All the results shown in the plot are significant with  $\alpha < 0.001$ .

## 6.6 RELATED WORKS

**INTERPRETABILITY METHODS** Interpretability methods come in different flavors: (a) post-hoc analysis methods that estimate input feature importance for decisions, including LIME, (b) post-hoc analysis methods that estimate the influence of training instances on decisions, e.g., influence functions (Koh and Liang, 2017) and (c) strategies for making complex models interpretable by learning to generate explanations (Narang et al., 2020b) or uptraining simpler models (Agarwal et al., 2020). In this paper

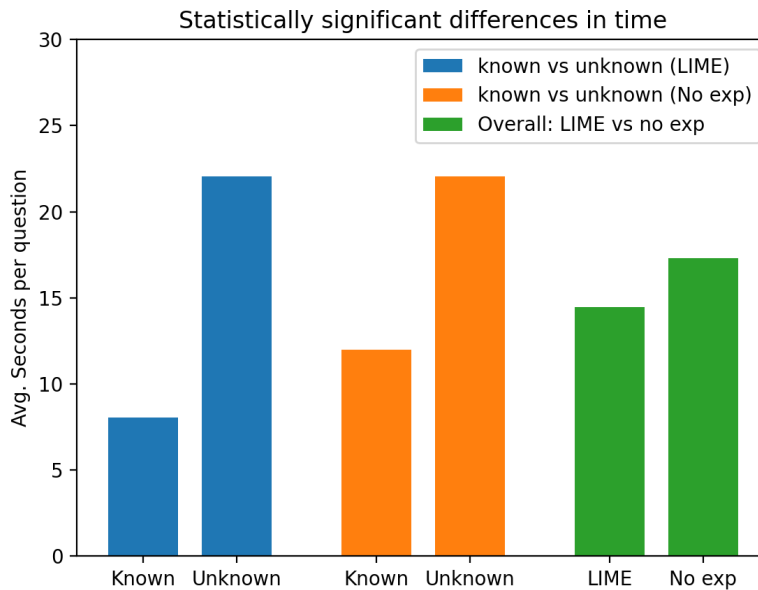


Figure 6.3: COMPARING KNOWN AND UNKNOWN TASKS. i) Left bars show mean inference time (secs) *with* LIME explanations; ii) middle bars show mean inference time *without*; and iii) right bars show mean inference time across *all* tasks, with and without LIME.

we have focused on post-hoc interpretability methods, but it is equally important, we argue, to evaluate other types of interpretability methods on unknown tasks, when running human forward prediction experiments, to avoid participants' cognitive biases.

#### INTRINSIC EVALUATION OF INTERPRETABILITY METHODS

One standard approach to evaluating explanations is to remove the parts of the input detected by the interpretability method and see whether classifier performance degrades (Samek et al., 2017). One drawback of this method is that the corrupted examples are now out-of-distribution, and classifiers will generally perform worse on such examples. Hooker et al. (2019) improve on this by evaluating classifiers retrained on the corrupted examples. This approach, however, now suffers from another drawback: If classifiers perform well on the corrupted examples, that does not mean the interpretability methods were wrong.<sup>2</sup> Jain and Wallace (2019) evaluate attention functions as

<sup>2</sup> To see this, consider a sparsity-promoting classifier relying on a single feature  $f$  in the context of feature swamping (Sutton, Sindelar, and McCallum, 2006), i.e., frequent features may lead to undertraining of covariate features in discriminative learning. If  $f$  is removed, but the classifier retains its original performance by now relying on covariate features, that does not mean the classifier did not solely rely on  $f$  when trained on the original data.

explanations and argue that they do not provide useful explanations, in part because they do not correlate with gradient-based approaches to determining feature importance; Wiegrefe and Pinter (2019), in return, show this test is not sufficient to show attention functions do not provide useful explanations.

#### EXTRINSIC BENCHMARKS FOR INTERPRETABILITY METHODS

Rei and Søgaard (2018) show how token-level annotated corpora can be converted to benchmarks for evaluating post-hoc interpretability methods. They train sentence classifiers to predict whether sentences contain labels or not, use interpretability methods to predict what input words were important, and use the  $F_1$  score of those predictions to evaluate the interpretability methods. The main drawback of their method is that it only works as an evaluation of interpretability methods under the assumption that the classifier is near-perfect (since otherwise the token-level annotations cannot be assumed to be explanations of model decisions); furthermore, it is only applicable to tasks for which we have token-level annotations. Poerner, Schütze, and Roth (2018) adopt a slightly different approach, augmenting real documents with random text passages to see whether interpretability methods focus on the *original* text passages. This method suffers from the same drawback, that it assumes near-perfect performance. It is also only designed to capture false positives; it cannot distinguish between true or false negatives. Finally, DeYoung et al. (2020) recently introduced ERASER,<sup>3</sup> a suite of NLP datasets augmented with rationales, including reading comprehension, natural language inference, and fact checking. ERASER also assumes near-perfect performance, and can be seen as extending the set of tasks for which the method proposed in Rei and Søgaard (2018), is applicable. In our current work, we present a method of evaluating post-hoc interpretability methods which is independent of model quality, which we argue is a great advantage of our proposed experimental design as opposed to many others, such as the studies mentioned above.

#### HUMAN EVALUATION OF EXPLANATIONS

The idea of evaluating explanations by testing human participants' ability to predict model decisions with and without explanations is not novel. Nguyen (2018a), Lage et al. (2018) and Hase and Bansal (2020), as already discussed, present such experiments. Schmidt and Biessmann (2019) is another example of human forward prediction experiments in a crowdsourcing platform. They perform experiments on the effect of LIME and COVAR on human

<sup>3</sup> <http://www.eraserbenchmark.com/>

forward prediction for a sentiment task that is known to be participants, in advance. Our criticism of Nguyen (2018a) also applies to their study.

Narayanan et al. (2018) also present evaluations of interpretability methods with humans. In contrast to our work, they design simple tasks in which humans verify whether an output is consistent with an input and an explanation. The human participants are provided with explanations of what the tasks they are working with are, and they only need to consider a handful of input features.

The Reverse Turing Test that we propose here is different from previous proposals to use human forward prediction to evaluate interpretability methods, in that it a) includes a training phase, and b) includes human forward prediction on *unknown* tasks, i.e., tasks about which they have no prior beliefs. We are, to the best of our knowledge, the first to propose such a protocol. In the above experiments, designed to motivate the design of the Reverse Turing Test, we see the limitations of a widely used interpretability method, LIME: On some tasks, i.e., tasks which cannot be explained by the occurrence of input words, the effect of LIME is detrimental; and on unknown tasks, for which LIME interpretations are not supported by participants' belief biases, its effect on human forward prediction is insignificant.

## 6.7 CONCLUSION

We presented an evaluation protocol for interpretability methods, which differs from previous work by including a training phase and by including unknown tasks. This makes our protocol work independently of model quality, and controls for belief bias. Using LIME as our test case, we find that on known tasks, LIME leads to statistically significant improvements in human forward prediction, both in accuracy and inference time. However, when tasks are unknown, differences are no longer significant. We see this as evidence of bias in the standard protocols, and argue that making tasks unknown, leads to more reliable evaluations. We also identify tasks, where model decisions cannot be explained in terms of input word occurrences, and for which the effect of LIME is detrimental for human forward prediction performance.



## ON THE INTERACTION OF BELIEF BIAS AND EXPLANATIONS

---

### 7.1 ABSTRACT

A myriad of explainability methods have been proposed in recent years, but there is little consensus on how to evaluate them. While automatic metrics allow for quick benchmarking, it is not clear how such metrics reflect human interaction with explanations. Human evaluation is of paramount importance, but previous protocols fail to account for humans' *belief biases* affecting performance, which may lead to misleading conclusions. We provide an overview of belief bias, its role in human evaluation, and ideas for NLP practitioners on how to account for it. Using two experimental paradigms, we present a case study of gradient-based explainability introducing simple ways to control for humans' previous beliefs: models of varying quality and adversarial examples. We show that *conclusions about the highest performing methods change when introducing such controls*, pointing to the importance of accounting for belief bias in evaluation.

### 7.2 INTRODUCTION

Machine learning has become an integrated part of our lives; from everyday use (e.g., search, translation, recommendations) to high-stake applications in healthcare, law, or transportation. However, its impact is controversial: neural models have been shown to make confident predictions relying on artifacts (McCoy, Pavlick, and Linzen, 2019; Wallace et al., 2019) and have shown to encode and amplify negative social biases (Caliskan, Bryson, and Narayanan, 2017; González et al., 2020; Manzini et al., 2019; May et al., 2019a; Rudinger et al., 2018; Tan and Celis, 2019).

*Explainability* is aimed at making model decisions transparent and predictable to humans; it serves as a tool for model diagnosis, detecting failure modes and biases, and more generally, to increase trust by providing transparency (Amershi et al., 2019b). While automatic metrics have been proposed to evaluate qualities of explanations such as: faithfulness, consistency and agreement with human explanations (Atanasova et al., 2020a; DeYoung et al., 2020; Robnik-Šikonja and Bohanec, 2018),

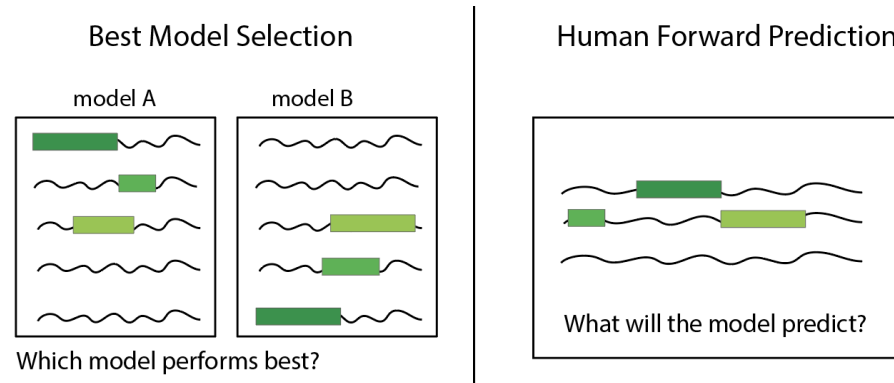


Figure 7.1: Evaluation protocols considered in this work

such metrics fail to inform us about human interaction with explanations.

Doshi-Velez and Kim (2017) suggested *human forward prediction*, a simulation task in which humans are given an input and an explanation, and their task is to provide the expected model output, regardless of the gold answer; recent studies include Hase and Bansal (2020), Lage et al. (2019), Nguyen (2018b), and Poursabzi-Sangdeh et al. (2018). Such protocols are widely used and can provide valuable insight into human understanding of explanations. However, prior work has not accounted for how humans’ previous beliefs (*belief biases*) interact with evaluation; simulating model decisions becomes an easier task when the model being evaluated makes predictions which align with human behavior. We argue that not considering belief bias may lead to misleading conclusions about which explainability methods perform best.

Other protocols have evaluated participants’ ability to select the best model based on explanations offered by different interpretability methods (e.g., decide which model would generalize ‘in the wild’) (Ribeiro, Singh, and Guestrin, 2016a). However, comparisons have been made between a model which is clearly in line with human beliefs, and another which exploits spurious correlations diverging from human expectations. When differences are less obvious, humans may not be able to leverage their belief biases, and conclusions may change.

This paper, which includes evaluations for both previously mentioned tasks, closes an important gap: to the best of our knowledge, no prior work in NLP addresses the interaction of belief bias with human evaluation of explainability.

**CONTRIBUTIONS.** We provide an overview of belief bias meant to highlight its role in human evaluation and provide some preliminary ideas for NLP practitioners on how to handle such cases. Using *human forward prediction* and *best model selection*



(Figure 7.1), we present a case-study where we compare two gradient-based explainability methods in the context of Reading Comprehension (RC), introducing conditions to take into account belief bias. We find that both explainability methods are helpful to participants in the standard settings (in line with most previous work), but the *conclusions about the best performing models change when incorporating additional control conditions*, reinforcing the importance of accounting for such biases.

### 7.3 BELIEF BIAS

**Belief bias** is a type of cognitive bias, defined in psychology as *the systematic (non-logical) tendency to evaluate a statement on the basis of prior belief rather than its logical strength* (Barston, 1986; Evans, Barston, and Pollard, 1983; Klauer, Musch, and Naumer, 2000). Cognitive biases are not necessarily bad; they help us filter and process a great deal of information Bierema et al., 2020, and have been widely studied in real human-decision making (Furnham and Boo, 2011; Kahneman, 2003; Tversky and Kahneman, 1974). However, in evaluations involving human participants, such biases may alter results and affect conclusions (Anderson and Hartzler, 2014; Wall et al., 2017).

Classic psychology studies of belief bias have assessed how prior beliefs affect syllogistic reasoning (Evans, Barston, and Pollard, 1983; “SJ and Harper, C. 2001”; Klauer, Musch, and Naumer, 2000; Markovits and Nantel, 1989; Newstead et al., 1992). Consider the following example by Anderson and Hartzler (2014):

- (a) *If all birds are animals, and if no animals can fly, then no birds can fly.*
- (b) *If all cats are animals, and if no animals can fly, then no cats can fly.*

In syllogistic reasoning, the task for humans is to assess the *logical* validity of such arguments while ignoring believability. While both arguments are logically valid, most work converges on the finding that humans will rate argument (a) as invalid more often than (b), biased by the fact that the premise in (a) is less believable.

In psychology, belief bias has been tied to the dual-processing theory, which assumes that reasoning is performed by two competing cognitive systems: (1) *system 1* which takes care of fast, heuristic processes and (2) *system 2* which handles slower, more analytical processes (Croskerry, 2009; Evans, 2003; Evans and Curtis-Holmes, 2005; Trippas and Handley, 2018). Generally, humans tend to have a cognitive preference for relying on fast, intuitive system 1 processes, rather than engaging in slow and more analytical system 2 processes. Belief bias is attributed to

system 1 (Evans, 2008; Evans and Curtis-Holmes, 2005; Evans and Frankish, 2009; Stanovich and West, 2008) due to several factors, reviewed in detail by Caravona et al. (2019) and Evans (2003).

For the purposes of NLP studies relying on crowd workers, one relevant finding is that **time pressures exacerbate reliance on previous beliefs** (Evans and Curtis-Holmes, 2005). Since crowd workers generally are incentivized to work as quickly as possible to maximize their hourly pay, reliance on belief bias is to be expected.

Another relevant finding for NLP is that threatening or negatively charged arguments (e.g., content violating political correctness, and social norms) leads to greater engagement of system 2, whereas **neutral content leads to increased reliance on belief bias** (Goel and Vartanian, 2011; Klaczynski, Gordon, and Fauth, 1997). Since NLP studies tend to be performed on neutral content such as passages from Wikipedia –content which may not sufficiently engage participants’ system 2 processes– belief bias is more likely to play a role in human performance.

This study aims to highlight the phenomenon of belief bias to encourage NLP practitioners to assess the role it plays in their evaluations, and introduce mechanisms to control for belief bias effects. We illustrate how belief bias effects can significantly affect the results of human evaluation of explainability for two paradigms: *human forward prediction* and *best model selection*.

#### 7.4 RELATED WORK

**HUMAN FORWARD PREDICTION** Human forward prediction experiments have been recently presented in the context of synthetic data (Lage et al., 2019; Poursabzi-Sangdeh et al., 2018; Slack et al., 2019) to evaluate explainability methods for their ability to make model decisions predictable to humans. In this paradigm, humans are presented with explanations and tasked with predicting the model’s decision regardless of the ground truth (Doshi-Velez and Kim, 2017).<sup>1</sup>

In NLP, Nguyen (2018b) introduced human forward prediction for LIME explanations (Ribeiro, Singh, and Guestrin, 2016b) of sentiment analysis of product reviews and correlated the results with automatic evaluations. Unlike with synthetic data, participants have prior beliefs on what the *true* outcome is. Since participants in Nguyen (2018b) had no training phase to learn how explanations correlate with predictions and the model be-

<sup>1</sup> Using synthetic data from fictitious domains effectively controls for belief bias (Lage et al., 2019; Slack et al., 2019). Slack et al. (2019), for example, evaluates explanations in the domain of recommending recipes and medicines to aliens.

ing evaluated sufficiently matched human behavior, humans likely relied *exclusively* on their prior knowledge, and beliefs to complete the task at hand.

Hase and Bansal (2020) improved on this protocol by adding a training phase. This is something we also do in our experiments (Section 7.6), but it is unlikely to solve the belief bias problem because even after training, humans will naturally opt for fast heuristic mechanisms (e.g., belief bias) to simplify tasks (Wang et al., 2019); this is particularly true if the model is high performing (aligns with human beliefs).

The protocol by Hase and Bansal (2020) had another key feature: they leave out the explanations for the test data points. This would seem like an advantage for evaluating explainability methods in the context of RC where explanations can, in theory, simply highlight the answer span, making it easy to guess the model output from the explanations. However, it is easy to control for the amount of explanation provided by the explanation methods you compare; in our experiments below, we highlight the top 10 tokens with the highest attribution scores. This part of their protocol is problematic for two reasons:

- It makes the human learning problem much harder, and we argue it is infeasible to expose participants to enough examples to make human forward prediction learnable (unless the task is made very easy on purpose; again by only evaluating high performing models). If it is not learnable, participants fall back on belief bias.
- It introduces a systematic bias between the training and test scenarios.

The protocol in Hase and Bansal (2020) also does not properly randomize the order in which participants are exposed to problems with or without explanations.

We improve on the above protocol by introducing a control condition for belief bias effect: evaluating explainability methods on low-quality models, the predictions of which substantially differ from human beliefs. This means that to succeed in the task, humans cannot simply rely on their previous beliefs. This condition would help us assess the ability of explanations in helping humans to *realign* their expectations with model behavior. The predictions of RC models can also be made different from human answers by introducing distractor sentences that fool machine reading models, but not humans (Jia and Liang, 2017). If in human forward prediction, participants predict the true answer rather than spans in the distractor sentences, this suggests that participants are relying on their belief biases.

**BEST MODEL SELECTION** Ribeiro, Singh, and Guestrin (2016b) presented an evaluation of explainability methods for text classification, where explanations for decisions of two different models on the same instance are presented side by side, and humans decide which model is likely to generalize better. With some exceptions (Lertvittayakumjorn and Toni, 2019), there has not been much follow up work on this task, but this scenario is important: it mimicks the decision about what model is *safer for deployment*. Ribeiro, Singh, and Guestrin (2016b) and Lertvittayakumjorn and Toni (2019) both make a single comparison between a model which clearly diverges from human intuition, and a model that generalizes and *aligns with humans' beliefs*. Accounting for the extent to which belief biases are leveraged (e.g. by introducing additional model comparisons where differences are not so obvious) is important in such paradigms, and can allow us to better evaluate where explanation methods may fail.

In the following sections, we show that introducing control conditions which take into account belief biases can alter conclusions for both *human forward prediction* and *best model selection*. We emphasize that many other potential strategies can be introduced and this is largely dependent on the goals of the evaluation protocol; we merely provide one example case with the following strategies:

- (1) Introducing low quality models which considerably diverge from humans' prior beliefs (*human forward prediction*)
- (2) Introducing evaluation problems with distractor sentences (*human forward prediction*)
- (3) Introducing model comparisons where relying on belief bias is not enough to obtain high performance (*best model selection*)

## 7.5 EXPERIMENTAL SETUP

This section introduces the general setup of the experiments, with details specific to each experimental paradigm described in [Section 7.6](#) and [Section 7.7](#).

### 7.5.1 Models

We evaluate gradient-based ([Section 7.5.3](#)) explanations produced by three BERT-based (Devlin et al., 2019c) models:

- (a) a high performing model (HIGH): BERT-base, finetuned on SQuAD 2.0. *This model is more aligned with human beliefs.*
- (b) a medium performing model (MEDIUM): tinyBERT, a 6-layer distilled version of BERT (Jiao et al., 2019), finetuned

on SQuAD 2.0. It performs about 20  $F_1$  points below HIGH. *This model somewhat aligns with human intuition, but performs significantly lower.*

- (c) a low performing model (Low): BERT-base, fine-tuned to always choose the first occurrence of the last word of the question. This system mimicks a rule-based system<sup>2</sup>; however, we evaluate gradient-based methods requiring a neural model. *This model diverges significantly from human beliefs.*

### 7.5.2 Data

We use SQuAD 2.0 (Rajpurkar, Jia, and Liang, 2018), a RC dataset consisting of 150k factoid question-answer pairs, with texts coming from Wikipedia articles. We opt for this data as it contains short passages that can be read by humans in a short time. In the human forward prediction experiments, we refer to experiments using this data as ORIG. As described in Section 7.3, Wikipedia texts could by themselves induce people to rely on their belief bias, but this particular dataset allows us to also introduce controls for the bias: the adversarial version of the data (Jia and Liang, 2017), has been shown to distract models but not humans. This means that in order to perform the task with success, humans need disregard their belief biases, and in some cases align with distractor sentences. We refer to this data in our simulation experiments as ADV.

### 7.5.3 Explainability Methods

We focus on gradient-based approaches, as they require no modifications to the original network, and are considerably faster than perturbation-based methods. We compare two explainability methods:

**GRADIENTS** Computing the gradient of the prediction output with regard to the features of the input is a common way to interpret deep neural networks (Simonyan, Vedaldi, and Zisserman, 2013) and capture relevant information regarding the underlying model.

**INTEGRATED GRADIENTS IG** (Sundararajan, Taly, and Yan, 2017) attributes an importance score to each input feature by approximating the integral of gradients of the model’s output with respect to the inputs along the path, from the references

<sup>2</sup> This model achieves about 0.90  $F_1$  for this task, but in the results we show its performance on the actual RC task

to the inputs. IG was introduced to address the sensitivity issues which are present in vanilla gradients and implementation invariance.

## 7.6 EXPERIMENT 1: HUMAN FORWARD PREDICTION

Human forward prediction for evaluating explainability was proposed by Doshi-Velez and Kim (2017). They argue that if a human is able to simulate the model’s behavior, they understand *why* the model predicts in that manner. For the reasons previously outlined, we suspect that belief biases may be considerably affecting performance in this task. We investigate this by asking the following: *Can humans predict model decisions, if model behavior considerably diverges from their own beliefs?*

**STIMULI PRESENTATION** We include: (i) HIGH, which is finetuned to solve SQuAD 2.0 and (ii) Low, which is finetuned to select the first appearance in the context of the last word in the question. We evaluate each of the two models twice: with or without adversarial data. We contrast using vanilla gradients and IG with a baseline condition, in which no explanations are shown (BASELINE).

We highlight the top-10 tokens<sup>3</sup> with the highest attribution scores wrt. the start and end positions of the predicted span, and zero out the rest.<sup>4</sup> The two sets of tokens often overlap.

Participants were provided with a question and a passage (with or without explanations) and were told to pick the *shortest* span of text which matched the model prediction. They saw the actual model answers before the next example (done for both baseline and explanation conditions), which was an important part of training to infer model behavior. Before seeing the model prediction, their answers were locked to prevent any further changes. An example of our interface can be found in Figure 7.2 and the instructions are shown in Section A.3.1.

We ran these experiments on Amazon, MTurk, recruiting participants with approval ratings greater than 95%<sup>5</sup> and ensuring different groups of participants per condition by specifying that participation is only allowed once, otherwise risking rejection.<sup>6</sup> We paid participants \$5.25 for about 20 minutes of work, (to ensure at least a \$15 hourly pay), and obtained at least three

<sup>3</sup> Explanations should be *selective* (Mittelstadt, Russell, and Wachter, 2019)

<sup>4</sup> Ribeiro, Singh, and Guestrin (2016a) use the top 6 attributes; we opt for 10 given that our texts are slightly longer.

<sup>5</sup> Previous research has shown that proper filtering and selection of participants on MTurk, can be enough to ensure high quality data (Peer, Vosgerau, and Acquisti, 2014).

<sup>6</sup> We also remove such (few) repetitions at analysis



**Question:**

when was the latin version of the word norman first recorded ?

**Context:**

the english name normans comes from the french words normans / normanz , plural of normant , modern french normand , which is itself borrowed from old low franconian nortmann northman or directly from old norse , latinized variously as nortmannus , normannus , or nordmannus ( recorded in medieval latin , 9th century ) to mean norseman , viking .

start of answer:

end of answer:

[Click here to show answer](#)

[Reset answers](#)

Figure 7.2: Interface for Experiment 1 for LOW condition. To select model predictions, participants clicked on tokens to select the start and end of the span. Then they would see the actual model prediction.

annotations per example. The data included 120 unique questions divided into small fixed batches (the same questions across conditions). About 75% of questions are accurate in the HIGH model, and around 15% are accurate for the Low model. In total, we obtained 4,300 data points across 123 participants (35 data points per participant).

**RESULTS** As humans often did not select the exact span that was provided as the ground truth, we *manually* labeled the spans as correct or incorrect. We also inspected the impact of training in human forward prediction, e.g., the learning effect of multiple exposures on annotator accuracy. Both with vanilla gradients and integrated gradients, we observe an increase in the participants' accuracy at around 15 examples. In contrast, in our baseline condition, performance either stays constant or drops slightly. To reduce the noise introduced due to the training period, we remove the first 15 examples of each participant. The results without this preprocessing (Section A.3.1) suggest that **the effect of training differed across explainability methods**, as will be discussed later in the section.

Using the average human accuracy per example, we run a one-way ANOVA to test for significant differences across the groups. As we obtained statistically significant results, we then

	MODEL	Human		
CONDITION	F1	$\hat{y}$	$y$	SEC
<b>Baseline</b>				
LOW-ORIG	0.17	0.16	0.48	33.9
LOW-ADV	0.15	0.12	0.34	63.3
HIGH-ORIG	0.79	0.45	0.46	34.6
HIGH-ADV	0.66	0.38	0.48	36.1
<b>Integrated (IG)</b>				
LOW-ORIG		*0.58	*0.22	*16.8
LOW-ADV		*0.63	*0.18	*22.3
HIGH-ORIG		* <b>0.84</b>	*0.88	36.1
HIGH-ADV		* <b>0.52</b>	*0.35	*18.9
<b>Gradients</b>				
LOW-ORIG		* <b>0.69</b>	*0.06	32.6
LOW-ADV		* <b>0.72</b>	*0.15	*25.6
HIGH-ORIG		*0.79	*0.81	47.4
HIGH-ADV		0.49	*0.60	48.4

Table 7.1: Human forward prediction results (  $HUMAN(\hat{y})$  ) for Low and HIGH models, compared to no explanations (  $BASELINE$  ). Each experiment is run on vanilla SQuAD 2.0 data (  $ORIG$  ) and adversarial SQuAD 2.0 data (  $ADV$  ).  $HUMAN(y)$  is the dataset ground truth and an indicator of belief bias. Statistically significant results are indicated with an asterisk. Time is the average time per question. The best  $\hat{y}$  results in each condition are bolded.

ran the Tukey Honest Significant Difference (  $HSD$  ) test (Tukey, 1949), comparing the means of every condition to the means of every other condition. The results are presented in Table 7.1.

As expected, in the absence of explanations (  $BASELINE$  ), **humans rely on belief bias and predict the gold standard answer more often than the model prediction** ( $y$  in Table 7.1). Even with training (seeing the true model prediction), humans fail to catch onto the simple rule used by the Low model, when no explanations are presented.

Overall, explanations derived from both of the gradient-based approaches lead to statistically significant improvements over the baseline. This indicates that the **explanations allow humans to realign their expectations of the model behavior**, better than with no explanations.



For HIGH-ORIG, the standard setting explored in previous evaluations, both IG and vanilla gradients perform well, with IG performing better. Given these results and the theoretical advantages of IG over vanilla gradients, one could arrive at the conclusion that IG are better for simulatability. However, **the differences between the two methods are reversed in the conditions where humans cannot rely on their previous beliefs (Low)**. The gap between gradients and IG as large as 0.11, and being statistically significant. This finding is *surprising*.

Finally, in the HIGH conditions, model behavior decreases about 13% F1 score with the presence of **adversarial examples**, meaning that the model we used does have weaknesses to adversarial inputs. We observe that human performance is considerably lower in HIGH-ADV as opposed to HIGH-ORIG. **With vanilla gradients, performance is more aligned with the ground truth labels than with model behavior**, showing that humans are also relying on their previous beliefs. **With IG, where performance is less aligned with previous beliefs (ground truth), the end performance increases**, but, in general it seems that this condition is considerably more difficult for humans.

**EFFECT OF TRAINING** In BASELINE, training does not have an effect on neither of the LOW or HIGH conditions (see [Table A.6](#) in [Section A.3.1](#) for the raw results). For the LOW model, multiple factors can be taking place (possibly at the same time): (1) the task is simply too far from humans' beliefs and there is no mechanism to help participants realign their expectations, (2) participants may simply not be incentivized to seriously engage and look for patterns, (3) participants opt for a mixed strategy, where for some questions they go with their prior beliefs and for others choice is random (as seen in their performance in  $y$ ).

For HIGH conditions in BASELINE, performance remains higher than LOW but this is likely due to belief bias and not training, given that performance remains constant after removing the training data points. We hypothesize that for HIGH, instances where the model does not align to human intuition might be more detrimental than in explanation conditions. More specifically, if humans are aware that the model aligns with their beliefs after some examples but encounter instances where it doesn't (model is not 100% accurate), they will likely develop an expectation that the model is bound to make some errors, without any indication of when.

In addition, our raw results suggest IG required longer training. While this does not mean IG is a worse method than vanilla gradients, explanations derived from IG may have confused the participants due to containing information which was irrelevant

to them. It may be that experts (e.g. system engineers knowledgeable about neural networks) can take advantage of such explanations; however, this is a direction for future work.

## 7.7 EXPERIMENT 2: BEST MODEL SELECTION

This section presents the setup and results of our model selection experiments; a task where humans select the model that is more likely to succeed in the wild. We present the participants with explanations from two models (HIGH vs LOW and HIGH vs MEDIUM), and ask them to decide which model is likely to perform better. As a follow-up, we also experimented with *soliciting explanations about what leads the worse model to fail*. Intuitively, comparative evaluation difficulty depends on how clear the difference is between the compared objects. Explanations should at least show the difference between a high-performing model and a low-performing one, enabling human participants to predict which is better (standard setting).

**STIMULI PRESENTATION** We presented participants with saliency information from both models (a high performing model + one of the lower performing models); their task was to determine which model performs best in the wild. We shuffled the order at random so that the best model would not remain in a fixed position. We obtain 120 samples (question-context pairs), and show the explanations next to each other as seen in [Figure 7.1](#). The participants are told that the highlighted attributes are words the model found important in making its final decision. A screenshot of the UI is shown in [Figure 7.3](#) and the instructions provided to the participants are shown in [paragraph A.3.2](#). These experiments were also ran on [MTurk](#) with the same general procedures and pay. The same subset of 120 examples is used in all conditions. We obtained at least three annotations per example and ended with a total of 1440 data points across 48 participants (30 examples each).

**RESULTS** For each example shown to annotators, we obtained the average accuracy scores and performed a standard T-test to compare the performance of the two methods. The results are shown in [Table 7.2](#). Using explanations from both methods, when shown the HIGH and LOW model, humans are clearly able to correctly select the better one. With [IG](#), humans achieve **0.95** accuracy on average, while with vanilla gradients they achieve **0.89**. The difference is *not* statistically significant. The fact that users are consistently able to discriminate between HIGH and

**Model A****Question:**

when was the **latin version** of the word **norman** first recorded ?

**Context:**

the english name **normans** comes from the french words **normans / normanz** , plural of **normant** , modern french **normand** , which is itself borrowed from old low franconian **normann northman** or directly from old **norse** , **latinized variously** as **nortmannus** , **normannus** , or **nordmannus** ( **recorded in medieval latin** , **9th century** ) to mean **norseman** , **viking** .

**Model B****Question:**

when was the latin version of the word **norman** first **recorded** ?

**Context:**

the english name **normans** comes from the french **words** **normans / normanz** , plural of **normant** , modern french **normand** , which is itself borrowed from old low franconian **normann northman** or directly from old **norse** , **latinized variously** as **nortmannus** , **normannus** , or **nordmannus** ( **recorded in medieval latin** , **9th century** ) to mean **norseman** , **viking** .

- Model A  
 Model B

Figure 7.3: Experiment 1 UI: Low(bottom) vs HIGH(top) condition.

Low models is expected, and serves as a *sanity check* that these explanations are meaningful for humans.

Condition	Gradients	IG
HIGH vs LOW	0.89	0.95
HIGH vs MEDIUM*	0.85	0.52

Table 7.2: Both methods do well in ( HIGH vs LOW). In HIGH vs MEDIUM, performance drops dramatically for IG. \* = statistical significant difference ( $\rho < 0.001$ )

When the same experiment was repeated in the HIGH vs MEDIUM condition, we found clear and statistically significant differences between the two explainability methods. Using IG, participants reach only **0.52** accuracy, while with vanilla gradients their performance is **0.85**. This is *surprising*, given that the difference in performance between the two models is still quite large (about 20% F1); the expectation is that both methods would capture this difference relatively well. It appears that

when both models *more or less* align with human beliefs, the task is much more difficult. To solve the task, humans now need to engage in more analytical thinking and cannot simply rely on belief biases to solve the task. We further investigate these differences through qualitative coding.

**QUALITATIVE ANALYSIS** After each instance, we asked participants to describe how the worse model will fail. We do not provide detailed guidelines in order to not further bias the participants by introducing specific criteria. The instructions given to the participants are shown in [Section A.3.2](#).

We collected 1440 responses, which were all inspected manually to uncover categories (codes). After multiple iterations, we tagged each response with one code (categories are mutually exclusive, no response can be placed in two). A description of the categories and their distribution are shown in [Figure 7.4](#), and examples of feedback per category are provided in the [Section A.3.2](#).

In the HIGH vs Low condition, feedback for both methods was generic (about 70-80% of the time), e.g., *model B is likely incorrect so it is worse*. This was expected: this task should be easy when model differences are large and humans can rely on their *system 1* processes to get through the task without thinking deeply about the explanations.

In the HIGH vs MEDIUM condition, the distribution of the feedback categories is very different. For IG, 50% of the time participants *felt* the highlighted tokens were irrelevant. This is not the case for gradients, where only about 15% of responses fell in that category. Additionally, for vanilla gradients, 50% of feedback is generic, signaling that in this condition, it may have been an easy task as well; explanations are making model behavior clear enough. It remains an open question whether IG explanations may in fact be more faithful to the model reasoning. In that case, *expert users* (e.g. a system engineer debugging a system) may not find IG attributions irrelevant and would be able take better advantage of the information provided. For this reason, we emphasize the importance of not overgeneralizing conclusions to other populations. Nevertheless, as evaluating on non-experts (crowdsourced workers for example) is common, this preliminary result is important: it shows that **conclusions can shift dramatically when introducing additional model comparisons**.

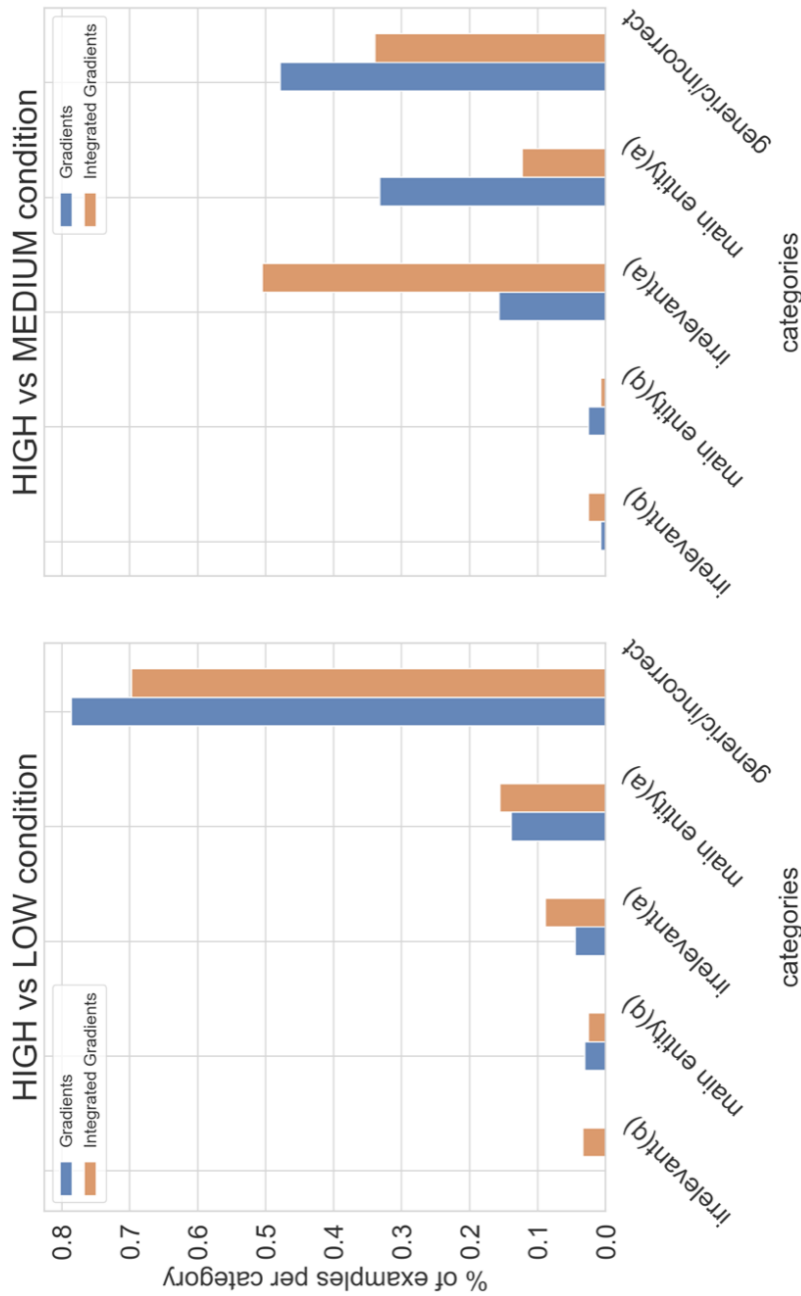


Figure 7.4: Feedback categories and their distribution. We observed that the HIGH vs MEDIUM condition results are considerably different from the HIGH vs LOW condition, with more participants giving generic answers for vanilla gradients, and emphasizing the irrelevant terms highlighted in the IG condition.

## 7.8 DISCUSSION: MITIGATING BELIEF BIAS

This study introduced *control conditions* in which the human participants could not rely on their belief biases to facilitate the task at hand. We presented a case study on evaluating RC models in model selection and human forward prediction paradigms, and we showed that this simple addition led to different conclusions in the evaluation and a better understanding of how humans interacted with explanations. Other tasks and paradigms might call for different setups, but generally control conditions with models of varying quality would be helpful both for the purposes of bias control and for the simulation of real-life use of explainability techniques to support decisions about which model is safer to deploy.

To conclude, we will briefly mention other directions for mitigating belief biases that can also be explored in future work.

**REDUCING AMBIGUITY** Ambiguity of task instructions leads humans to align interpretations with their own prior beliefs (Heath and Tversky, 1991); this may lead to misinterpretation and results which do not reflect the intended interaction with explanations. Ambiguity may also be present in other parts of the evaluation setup. For example, Lamm et al. (2020) evaluate the effectiveness of explanations in helping humans detect model errors for open-domain QA, but the data they use contains questions where multiple answers can be true. Users may deem an answer to be correct or incorrect based on their understanding of the question, which makes the effect of explanations blurry. Removing ambiguous instances from the data can be a way of reducing such confounds.

**REMOVING TIME CONSTRAINTS** Time constraints exacerbate the reliance of system 1 processes, which leads to humans relying on belief biases. In crowdsourced evaluations, it is common practice to provide workers with enough time to perform tasks, but workers may have intrinsic motivations for performing tasks quickly. A major challenge for evaluation research with crowd workers is creating better incentives for engaging in system 2 processes, e.g., pay schemes which encourage workers to be more analytical and accurate (Bansal et al., 2019b).

**INCLUDE FICTITIOUS DOMAINS** Using data from domains in which subjects have no prior beliefs, e.g., fictitious domains, may be an efficient way of controlling for belief bias in some

tasks<sup>7</sup>. This strategy has been used outside of NLP (Lage et al., 2019; Poursabzi-Sangdeh et al., 2018; Slack et al., 2019), where subjects are asked to imagine alternative worlds such as scenarios involving aliens. In QA for example, one could introduce context-question pairs that describe facts about fictitious scenarios that sufficiently differ from human reality.

## 7.9 CONCLUSION

The main contribution of this paper is bringing the discussion of belief bias from psychology into the context of evaluating explainability methods in NLP. We provide an overview of belief bias, making a connection between findings in psychology and the field of NLP, and present a case study of evaluating explanations for BERT-based RC models. We show that introducing models of various qualities and adversarial examples can help to control for belief bias, and that introducing such controls affects the conclusions about which explainability method works better.

---

<sup>7</sup> Again, we emphasize that some strategies are task dependent; fictitious domains may not be relevant in some tasks.





## DO EXPLANATIONS HELP USERS DETECT ERRORS IN OPEN-DOMAIN QA? AN EVALUATION OF SPOKEN VS. VISUAL EXPLANATIONS

---

### 8.1 ABSTRACT

While research on explaining the predictions of Open-domain Question Answering (ODQA) to users is gaining momentum, most works have failed to evaluate the extent to which explanations improve user trust. While few works evaluate explanations using user studies, they employ settings that may deviate from the end-user's usage in-the-wild: ODQA is most ubiquitous in *voice*-assistants, yet current research only evaluates explanations using a *visual* display, and may erroneously extrapolate conclusions about the most performant explanations to other modalities. To alleviate these issues, we conduct user studies that measure whether explanations help users correctly decide when to accept or reject an ODQA system's answer. Unlike prior work, we control for explanation *modality*, e.g., whether they are communicated to users through a spoken or visual interface, and contrast effectiveness across modalities. Our results show that explanations derived from retrieved evidence passages can outperform strong baselines (calibrated confidence) across modalities, but the best explanation strategy in fact changes with the modality. We show common failure cases of current explanations, emphasize end-to-end evaluation of explanations, and caution against evaluating them in proxy modalities that are different from deployment.

### 8.2 INTRODUCTION

Despite copious interest in developing explainable AI, there is increasing skepticism as to whether explanations (of system predictions) are useful to end-users in downstream applications. For instance, for assisting users with classifying sentiment or answering LSAT questions, Bansal et al. (2020) observed no improvements from giving explanations over simply presenting model confidence. Similarly, Chu, Roy, and Andreas (2020) observed that visual explanations fail to significantly improve user accuracy or trust. Such negative results present a caution-

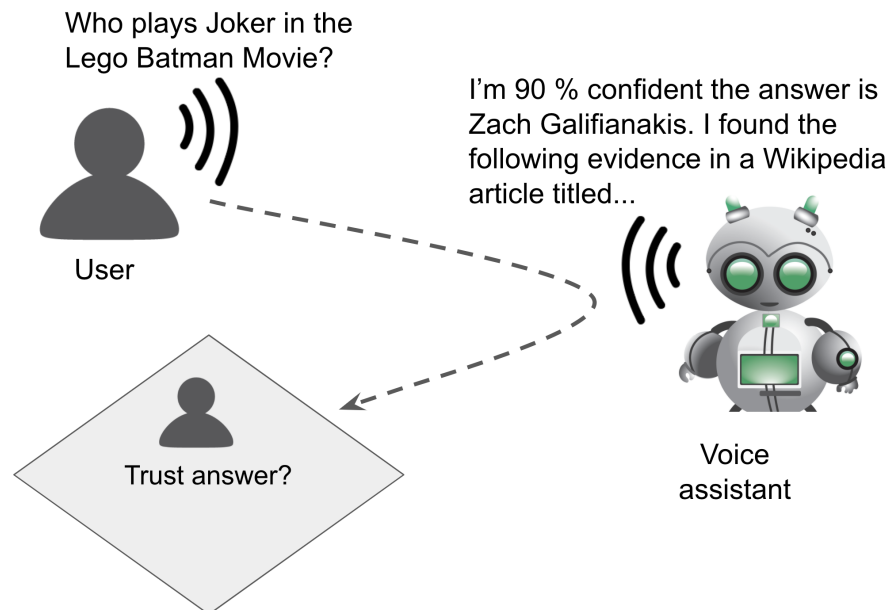


Figure 8.1: Using end-to-end user studies, we evaluate whether explanation strategies of open-domain QA assistants help users decide when to trust (or reject) predicted answers.

ary tale for explainability and emphasize the need to evaluate explanations using careful user studies.

We explore the effectiveness of explanations for *Open-Domain Question Answering* models, which involves answering users' factoid questions (e. g., "Who plays the Joker in the Lego Batman movie?") using a large corpus (e. g., Wikipedia). Such models are increasingly deployed not only in visual modalities (e. g., Web search) but also in spoken ones (voice assistants).<sup>1</sup> Spoken interfaces for *ODQA* are also important because they make systems more accessible for users with visual impairments. Despite improvements in accuracy, deployed *ODQA* models remain imperfect. This motivates the need to provide users with mechanisms (e. g., estimates of uncertainty or explanations) that can help improve *appropriate reliance* (Lee and See, 2004), e. g., by allowing users to detect erroneous answers. We henceforth refer to a user's ability to distinguish correct and incorrect answers as *error detectability*, and ask *Does explaining the system's reasoning, help improve error detectability?* (Figure 8.1)

Alongside recent negative results (Bansal et al., 2020), Lamm et al. (2020) showed that visually complex "QED" explanations that communicate coreference and entailment information along with evidence marginally improve error detectability. However, the study lacks the recommended baseline (Amershi et al., 2019b; PAIR, 2019) of communicating model confidence which has been

<sup>1</sup> <https://www.perflcient.com/insights/research-hub/voice-usage-trends>

shown to be effective on other domains (Bansal et al., 2020). Also, the transferability of complex visual explanations to the spoken modality remains unclear. Although Feng and Boyd-Graber (2019a) compare visual explanations with presenting model confidence on a different QA task, i.e., answering timed, multi-clue trivia questions, it was unclear whether explanations led to appropriate reliance (Bansal et al., 2020); thus the effectiveness of explanations for end users of QA systems still remains unclear. In this paper, we set out to evaluate the ability of Natural Language (NL) explanations in both visual and spoken modalities, to improve error detectability for the task of ODQA for non-expert users over strong baselines.

However, explaining ODQA systems in the spoken modality may pose unique challenges, e. g., because the same information content can impose higher cognitive demands when communicated by voice than visually (Leahy and Sweller, 2016; Sweller, 2011); potentially reducing effectiveness of longer, more complex explanations (e. g., QED) in the *spoken* modality. Thus we also ask, *Can the most useful explanation strategy change with presentation modality?* In summary:

1. We present user studies evaluating how well explanations for ODQA help users detect erroneous answers (error detectability). Unlike prior work, we evaluate explanations in both visual and spoken interfaces, and compare against calibrated confidence.
2. Our experiments with over 500 Mechanical Turk (MTurk) users confirm significant improvements in error detectability for ODQA over showing confidence. To the best of our knowledge, our work is the first to show statistically significant improvements in appropriate reliance through NL explanations for non-expert users. (Section 8.6.1)
3. We show that the best explanation approach can change with the modality: while longer explanations (evidence paragraphs) led to the highest error detectability in the visual modality, shorter explanations (evidence sentence) performed best in the spoken modality. We connect our observations with prior work on cognitive science and identify failure cases for ODQA explanations (Section 8.6.3).

### 8.3 RELATED WORK

**NATURAL LANGUAGE EXPLANATIONS** Recent work has introduced neural models that are trained to perform a task and output a NL explanation. Camburu et al. (2018) and Rajani et al. (2019), both introduce methods for training self-explaining models using free-form NL explanations collected from crowd-

sourced workers for natural language inference and common sense reasoning. Atanasova et al. (2020b) introduced a method for generating explanations for fact verification using human veracity justifications. Lei, Barzilay, and Jaakkola (2016) introduced an approach for extracting *rationales* by selecting phrases from the input text which are sufficient to provide an output. Rationales have since been introduced for various NLP tasks (Chen et al., 2018a; DeYoung et al., 2020; Yang et al., 2018).

Lamm et al. (2020) introduce QED explanations in ODQA consisting of the sentence containing the answer, coreference and entailment information. However, unlike free-form explanations or rationales, these explanations are too complex to adapt to the spoken modality. In question answering, many current models provide an answer and a rationale (or *extractive* evidence). We evaluate extractive evidences from a state-of-the-art ODQA model, along with human-written summaries.

**EVALUATING EXPLANATIONS** The quality of NL explanations has previously been evaluated using automatic metrics that measure the agreement of explanations with human annotations (Camburu et al., 2018; DeYoung et al., 2020; Paranjape et al., 2020; Rajani et al., 2019; Swanson, Yu, and Lei, 2020). It is not clear how these metrics reflect the usefulness of explanations in practice. As the goal of explainability is to make model decisions more predictable to *human* end users, a more useful way of evaluating explanations is through human evaluation.

Some human evaluations have used proxy tasks to evaluate explanations (Hase and Bansal, 2020; Nguyen, 2018b), however, Buçinca et al. (2020) showed that both subjective measures and proxy tasks tend to be misleading and do not reflect results in actual decision making tasks.

Within question answering, Feng and Boyd-Graber (2019b) evaluate how expert and novice trivia players engage with explanations. Lamm et al. (2020) evaluate how QED explanations help raters determine whether a model decision is correct or incorrect, and find marginal improvements on rater accuracy. Unlike these works, we simplify the presentation setup so that we can adapt explanations across different modalities. Bansal et al. (2020) observed that for sentiment analysis and answering LSAT questions, state-of-the-art explanation methods are not better than revealing model confidence scores and they increase the likelihood of users accepting wrong model predictions. We compare confidence to various explanation strategies for ODQA, but unlike previous work, we use *calibrated* model confidence.

OPEN-DOMAIN QA **ODQA** consists of answering questions from a corpus of unstructured documents<sup>2</sup>. Currently, **ODQA** models consist of two components: (1) a document *retriever* which finds the most relevant documents from a large collection and (2) a machine comprehension model or *reader* component, which selects the answer within the chosen documents (Chen et al., 2017a; Das et al., 2018; Karpukhin et al., 2020; Lee, Chang, and Toutanova, 2019a). Recent work focuses on identifying answers in Wikipedia (Karpukhin et al., 2020) as well as the web (Joshi et al., 2017), encompassing both short extractive answers (Rajpurkar et al., 2016) and long explanatory answers (Fan et al., 2019).

#### 8.4 VISUAL VS. SPOKEN MODALITIES

When presenting **NL** explanations to users, we must keep in mind that users typically process information differently across the spoken and visual modalities. In this section, we discuss work in learning and psychology research, which point to the differences motivating our evaluation.

1. **Real-time processing:** Flowerdew, Long, Richards, et al. (1994) observe that one of the main differences in how people process spoken versus written information is linearity. When listening, as opposed to reading, information progresses without you. Readers, on the other hand, are able to go back and dwell on specific points in the text, skip over and jump back and forth (Buck, 1991; Lund, 1991). Although in some scenarios it is possible to get spoken information repeated, it may not be as effective as re-reading (see below).
2. **Recall of information:** People tend to recall less after listening versus reading (Osada, 2004). Lund (1991) found that for some listeners, listening to information again was not as effective as re-reading. While advanced listeners benefited from listening multiple times, this was a controlled learning scenario simulating students learning classroom material; we would expect users in an **ODQA** setting to be slightly more passive.
3. **Effect on concentration:** The heavier cognitive load imposed by listening to information can make people lose concentration more easily. Thompson and Rubin (1996) found that optimal length for listening materials was around 30 seconds to 2 minutes. Beyond that, listeners

---

<sup>2</sup> <https://trec.nist.gov/data/qamain.html>

would lose full concentration. When people interact with voice assistants they may be on the go, or may be surrounded by additional distractions not present in a learning environment. This in turn may make the optimal length of material (explanations, in our case) much shorter.

We argue that these differences in the processing of spoken and written information can have tremendous consequences in the effectiveness of NL explanations in ODQA. Our experimental setup is the first to consider these differences.

## 8.5 EXPERIMENTAL SETUP

We design our user study to evaluate the explanation effectiveness for ODQA by varying two factors: *type* of explanation and *modality* of communication. We combine variations of each factor to obtain explanation conditions (Section 8.5.1) and obtain them using a state-of-the-art ODQA model (Section 8.5.3). We then deploy these conditions as HITs on Amazon MTurk to validate five hypothesis, each stating the relative effectiveness of conditions at improving error detectability (Section 8.5.2). Since MTurk studies can be prone to noise, to ensure quality control, we make and justify various design choices (Section 8.5.4).

### 8.5.1 Explanation Types and Conditions

ODQA models can justify their predictions by pointing to *evidence* text containing the predicted answer (Das et al., 2018; Karpukhin et al., 2020; Lee, Chang, and Toutanova, 2019a). We experiment with two types of *extractive* explanations:

- EXT-SENT: Extracts and communicates a sentence containing the predicted answer as evidence.
- EXT-LONG: Extracts and communicates a longer, multi-sentence paragraph containing the answer as evidence.

While extractive explanations are simpler to generate, we also evaluate a third explanation type that has potential to more succinctly communicate evidence spread across documents (Liu, Yin, and Wang, 2019).

- ABS: Generates and communicates new text to justify the predicted answer.

**FINAL EXPLANATION CONDITIONS** For the *voice modality*, we test five conditions, two baselines and three explanation types: (1) BASE: present only the top answer, (2) CONF, a second,



stronger baseline that presents the top answer along with the model’s uncertainty in prediction, (3) `ABS`, (4) `EXT-LONG`, and (5) `EXT-SENT`.

In the *visual modality*, we have 2 conditions corresponding to the `EXT-LONG` and `EXT-SENT` explanation types. Here, we were primarily interested in contrasting these with the voice modality. Examples of our explanations can be found in Appendix A.4.6.

### 8.5.2 Hypotheses

We investigated five (pre-registered) hypotheses about the relative performance of various explanation conditions at improving the accuracy of error detectability:

- H1** `CONF` will improve accuracy over `BASE`.
- H2** *Spoken* `EXT-SENT` will improve accuracy over `CONF`— the explanation would provide additional context to help validate predicted answers.
- H3** *Spoken* `EXT-SENT` will lead to higher accuracy than *Spoken* `EXT-LONG`. Since the spoken modality may impose higher cognitive limitations on people (Section 8.3), concise explanations may be more useful despite providing less context.
- H4** `ABS` will improve accuracy over *Spoken* `EXT-SENT`. `ABS` contains more relevant information than `EXT-SENT` (same length), which may help users make better accept/reject decisions.
- H5** *Visual* `EXT-LONG` will lead to higher accuracy than *Spoken* `EXT-LONG`.

### 8.5.3 Implementation Details for Conditions

**DATASET** For training our model and obtaining test questions for our user study, we used questions, answers, and documents from the Natural Questions (`NQ`) corpus (Kwiatkowski et al., 2019). `NQ` is composed of anonymized queries posed by real users on the Google search engine, and the answers are human-annotated spans in Wikipedia articles. The *naturally occurring* aspect of this data makes it a more realistic task for evaluating explanations. To simplify the study, we restrict our attention to the subset of questions with short answers (< 6 tokens) following Lee, Chang, and Toutanova (2019b). This subset contains 80k training examples, 8,757 examples for development, and 3,610 examples for testing.

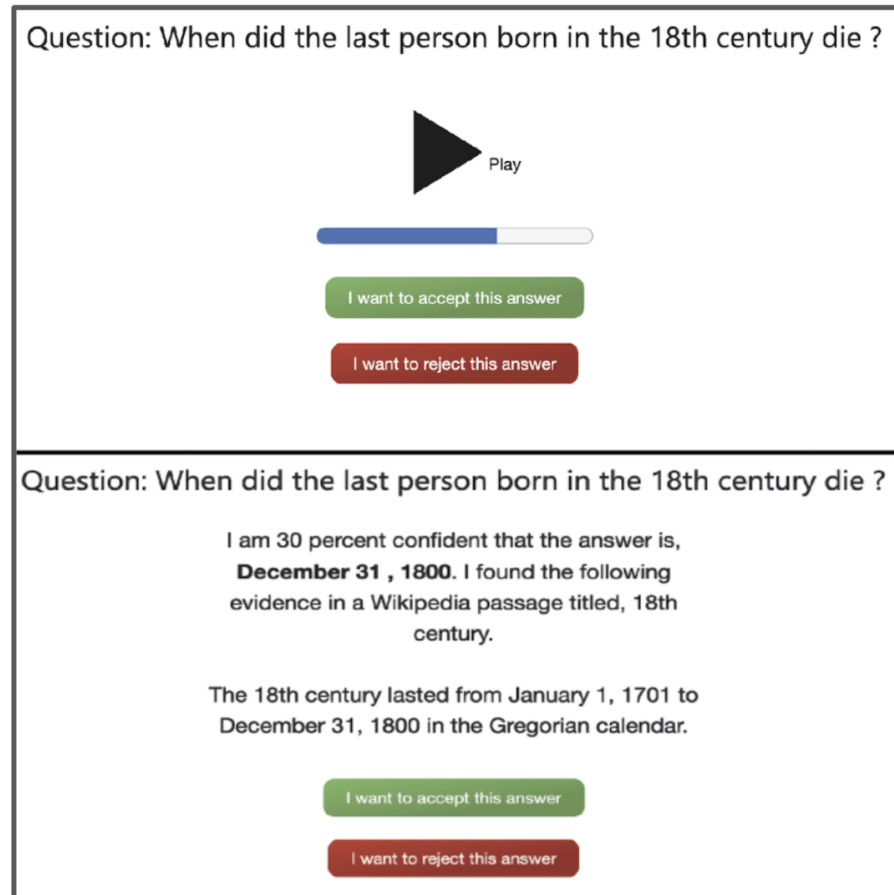


Figure 8.2: UI for visual (left) and spoken modalities (right) for EXT-SENT explanation type. Users either read or hear an explanation and decide whether to trust or discard the QA system’s prediction.

**MODEL** We train the current (extractive) state-of-the-art model on NQ: Dense Passage Retrieval (DPR) (Karpukhin et al., 2020). Similar to Karpukhin et al. (2020), we split documents (entire Wikipedia articles), into shorter passages of equal lengths (100 tokens). To answer an input question, DPR uses two separate *dense* encoders  $E_Q(\cdot)$  and  $E_P(\cdot)$  to encode the question and all passages in the corpus into vectors. It then retrieves  $k$  most similar passages, where *passage similarity* to a question is defined using a dot product:  $sim(q, p) = E_Q(q)^\top E_P(p)$ .

Given the top  $k$  passages, a neural reader (Section 8.3) assigns a passage selection score to each passage, and a *span score* to every answer span. The original model uses the best span from the passage with the highest passage selection score as the final answer. However, we re-score each answer using the product of passage and span score and use the highest-scored answer as the prediction. Our initial analysis showed that this rescoring improved the exact match scores of the predicted answers.



**GENERATING EXPLANATIONS** To create our extractive explanations, we use the passage associated with DPR’s answer—EXT-SENT is defined as the single sentence in the passage containing the answer and EXT-LONG is defined as the entire passage. Since DPR does not generate abstractive explanations, we simulate ABS by manually creating a single sentence that captures the main information of EXT-SENT and adds additional relevant information from EXT-LONG, whilst remaining the same length as EXT-SENT.

To improve transparency, in addition to presenting the evidence text in each explanation condition, we also inform users that the source of the text is Wikipedia and provide them with the *title* of the article containing the passage together with the model’s (calibrated) uncertainty in its prediction. Figure 8.2 shows an example of the final EXT-SENT explanation condition.

To convert text to speech, we use a state-of-the-art Text-to-Speech tool. For the questions we used in our study, when spoken, our final ABS and EXT-SENT conditions were on average 15 seconds long, EXT-LONG was between 30-40 seconds.

**CONFIDENCE CALIBRATION** Confidence scores generated by neural networks (e. g., by normalizing softmax scores) often suffer from poor calibration (Guo et al., 2017). To alleviate this issue and to follow the best practices (Amershi et al., 2019b) for creating strong baselines, we calibrate our ODQA model’s confidence using *temperature scaling* (Guo et al., 2017), which is a *post hoc* calibration algorithm suitable for multiclass problems. We calibrate the top 10 outputs of the model. We defer details on the improvement in calibration obtained through temperature scaling, and its implementation, to Appendix A.4.1.

#### 8.5.4 User study & Interface

We conduct our experiments using MTurk. Our task presents each worker with 40 questions one-by-one, while showing them the model’s answer (along with other condition-dependant information, such as confidence or explanation) and asks them to either *accept* the model’s prediction if they think it is correct or *reject* it otherwise. Figure 8.2 shows an example. For each of the 7 conditions, we hire 75 workers.

Additional details about our setting and the instructions can be found in Appendix A.4.3.

**QUESTION SELECTION** We deliberately sample a set of questions on which the model’s aggregate (exact-match) accuracy is 50%; thus any improvements in error detectability, beyond

random-performance, must be a result of users making optimal assessments about the model’s correctness. To improve the generalization of the results, we average results over three such mutually exclusive sets of 40 questions. Before sampling the questions, we also removed questions that were ambiguous or questions where the model was indeed correct, but the explanations failed to justify the answer. For brevity, we defer these details and justifications of these details to the Appendix [A.4.2](#).

**INCENTIVE SCHEME** To encourage [MTurk](#) workers to engage and pay attention to the task, we used a bonus-based strategy — When users accept a correct answer, we give them a 15 cent bonus but when they accept an incorrect answer they lose the same amount<sup>3</sup>. This choice aims to simulate real-world cost and utility from interacting successfully (or unsuccessfully) with [AI](#) assistants (Bansal et al., 2019a). Table 8.1 shows the final pay-off matrix that we used.

PREDICTION/DECISION	ACCEPT	REJECT
CORRECT	+\$0.15	\$0
INCORRECT	-\$0.15	\$0

Table 8.1: [MTurk](#) worker’s bonus as a function of the correctness of [ODQA](#) model’s prediction and the user’s decision to accept or reject the predicted answer.

**POST-TASK SURVEY** After the main task, we asked participants to (1) rate the length of responses, (2) rate their helpfulness and (3) give us general feedback on what worked and how explanations could be made better. For the *spoken modality*, we also asked participants to rate the clarity of the voice to understand if confusion originated from text-to-speech challenges. The survey as presented to the participants can be found in Appendix [A.4.4](#).

**QUANTITATIVE MEASURES OF ERROR DETECTABILITY** We quantify user performance at error detectability using the following three metrics:

- **Accuracy:** Percentage of times a user accepts correct and rejects incorrect answers. A high accuracy indicates high error detectability.

<sup>3</sup> If participants ended up with a negative bonus, no deductions were made from their base pay, instead their bonus was simply zero

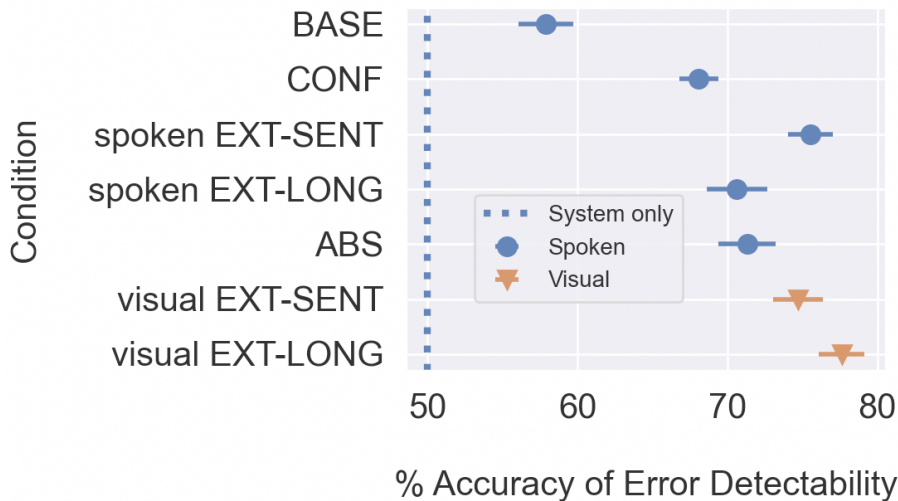


Figure 8.3: Accuracy of users at error detectability (75 workers per condition). In the *spoken modality*, *EXT-SENT* explanations yield the best results and is significantly better than *CONF*. In contrast, in the *visual modality*, *EXT-LONG* explanations perform best. We observe a statistically significant ( $p < 0.01$ ) difference between *EXT-LONG* in visual vs spoken, perhaps due to differences in user’s cognitive limitations across modalities.

- **% Accepts | correct:** Indicates the true positive rate, *i.e.*, percentage of times the user accepts *correct* answers.
- **% Accepts | incorrect:** Indicates the false positive rate, *i.e.*, percentage of times the user accepts *incorrect* answers. If a setting yields a high number, this would indicate that this setting misleads users more often.

We do not present true and false negative rates because the conclusions are similar. We additionally measure time spent on each question and the cumulative reward. These metrics are explained in Appendix A.4.5. When computing all metrics, we removed the first 4 questions for each worker to account for workers getting used to the interface. We pre-registered this procedure prior to our final studies.

## 8.6 RESULTS

To validate our hypothesis (Section 8.5.2) we compare explanation methods on the quantitative metrics (Section 8.6.1). To further understand participant behavior we analyze responses to the post-task survey (Section 8.6.2), and analyze common cases where explanations misled the users (Section 8.6.3). Results for reward and time metrics are included in Appendix A.4.5 and A.4.5.

### 8.6.1 Quantitative Results

Figure 8.3 shows average user accuracy at error detectability with 75 workers per condition. Similarly to Lamm et al. (2020), to validate hypotheses and compute statistical significance, we fit a generalized linear mixed effects model using the `lme4` library in R and the formula  $a \sim c + (1|w) + (1|q)$ , where  $a$  is the accuracy,  $c$  is the condition,  $w$  is the worker id and  $q$  is the question id. We run pairwise comparisons of these effects using Holm-Bonferroni to correct for multiple hypothesis testing. For both the spoken and visual modalities, all conditions lead to significantly higher accuracies than `BASE` ( $p < 0.01$ ).

**Model confidence improved accuracy of error detectability.** In Figure 8.3, `CONF` achieves higher accuracy than `BASE`—68.1% vs. 57.2%. This difference was statistically significant ( $p < 0.01$ ), thus **validating H1**. While previous guidelines recommend displaying confidence to users (Amershi et al., 2019b) and show its benefit for sentiment classification and LSAT (Bansal et al., 2020), our observations provide the first empirical evidence that confidence serves as a simple yet stronger baseline against which explanations for `ODQA` should be compared.

**Explaining via an evidence sentence further improved performance.** The more interesting comparisons are between explanation types and `CONF`. In both modalities, `EXT-SENT` performed better than `CONF`. For example, in the *spoken modality*, `EXT-SENT` improved accuracy over `CONF` from 68.1% to 75.6% ( $p < 0.01$ ); thus **validating H2**. Contrary to recent prior works that observed no benefit from explaining predictions, this result provides and confirms a concrete application of explanations where they help users in an end-to-end task.

**While longer explanations improved performance over more concise explanations in the visual modality, they worsened performance in the spoken modality.** Figure 8.3 shows that, for the visual modality `EXT-LONG` outperforms `EXT-SENT` explanations in the visual modality—77.6% vs. 74.7% ( $p < 0.4$ ). Conversely, for spoken, `EXT-SENT` is better than `EXT-LONG`—75.6% vs. 70.4% ( $p < 0.01$ ); thus **validating H3**. In fact, the decrease was severe enough that we no longer observed a statistically significant difference between long explanations and simply communicating confidence ( $p = 0.9$ ).

Although communicating the same content, *visual* `EXT-LONG` led to significantly better accuracy than their spoken version—77.6% vs. 70.4% ( $p < 0.01$ ); thus **validating H5**. These results indicate large differences, across modalities, in user ability to process and utilize explanations, and how these differences need

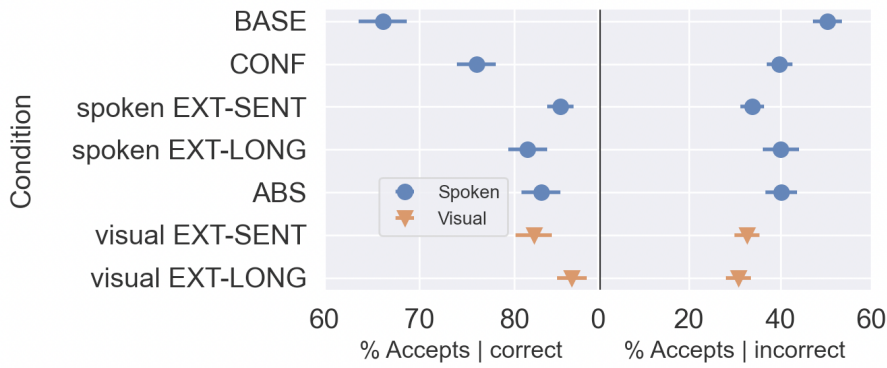


Figure 8.4: (Left) Explanations significantly increased participant ability to detect *correct* answers compared to simply displaying confidence. (Right) However, only `EXT-SENT` in the spoken modality and both explanations in the visual modality decreased the rate at which users are misled.

to be accounted for while evaluating and developing explanations.

**Despite improving conciseness, abstractive explanations did not help improve performance in the spoken modality.** Figure 8.3 shows that `ABS` performs significantly worse than `EXT-SENT` in the spoken modality—71.3% vs. 75.6% ( $p < 0.01$ ) and thus we could **not validate H4**. This result indicates that the length of the explanation (e. g., number of tokens) is not the only factor that affects user performance, instead, the density of information also increases cognitive load on users. This finding is in line with the Time Based Resource Sharing (TBRS) model (Barrouillet et al., 2007), a theory of working memory establishing that time as well as the complexity of what is being communicated, both play a role in cognitive demand. We also observe a similar effect in users’ subjective rating of length of explanation (Section 8.6.2).

**All explanations significantly increased participants’ ability to detect *correct* answers, but only some explanations improved their ability to detect *incorrect* answers.** Instead of aggregate accuracy, Figure 8.4 splits and visualizes how often users accept correct and incorrect answers. For accepting *correct* model predictions, all *visual* and *spoken* explanation conditions significantly helped compared to `CONF` (at least  $p < 0.05$ ).

In terms of accepting incorrect predictions, in the *spoken modality*, only `EXT-SENT` is significantly better (*i.e.*, lower) than `CONF`—34% vs. 40% ( $p < 0.05$ ). Whereas in the *visual modality*, both `EXT-LONG` and `EXT-SENT` lead to improvements over `CONF`—30% ( $p < 0.01$ ) and 32% ( $p < 0.05$ ), respectively. This shows that although explanations decrease the chance of being misled by the system, the least misleading explanations change with modality.

### 8.6.2 Qualitative results

We analyzed user responses to the post-task survey to understand their experience, what helped them and how the system could serve them better.

**VOICE CLARITY** To verify that the quality of the text-to-speech tool that we employed did not negatively affect our experiments, we asked users to rate the clarity of the assistant’s voice as *very poor*, *poor*, *fair*, *good*, or *excellent*. More than 90% of participants felt that the voice was good or excellent. These results can be observed in Figure 8.5

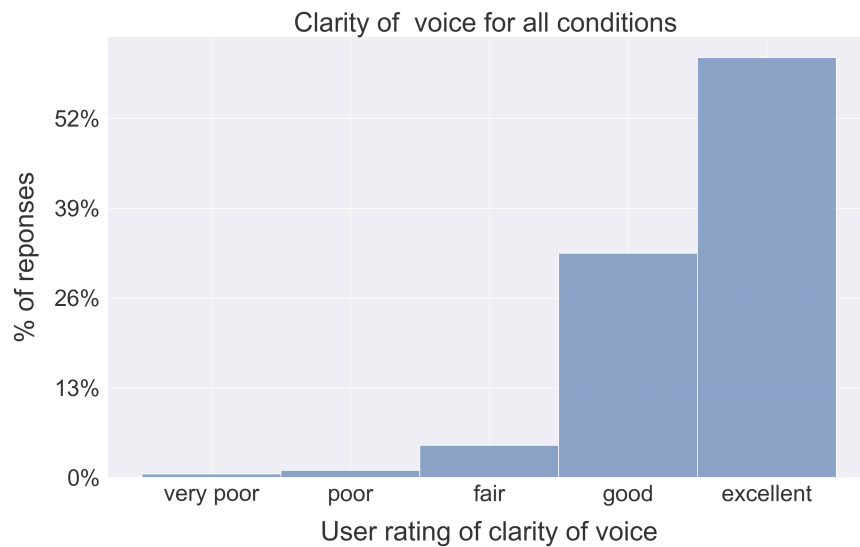


Figure 8.5: **Voice clarity:** Most participants found the voice of the assistant to be good or excellent.

**LENGTH PREFERENCE** We asked participants to rate the length of the explanation as *too short*, *short*, *right*, *long*, or *too-long*. Figure 8.6 shows the results. For **EXT-LONG**, over 85% of the workers perceived that in the *visual modality*, responses were the right length. On the other hand, in the *spoken modality*, only 30% of participants agreed the length was right. Thus, user’s subjective ratings for the same explanation type were dramatically different across modalities. Indicating, in addition to affecting error detectability, the modality also changes users’ subjective preferences.

Additionally, even though **ABS** and **EXT-SENT** were the same duration, in the post experimental survey, users indicated more often that they found **ABS** to be long, as opposed to **EXT-SENT**. As previously mentioned, this relates to the TBRS model of working memory (Barrouillet et al., 2007). We hypothesize that *our ABS* explanations, which integrate more information than



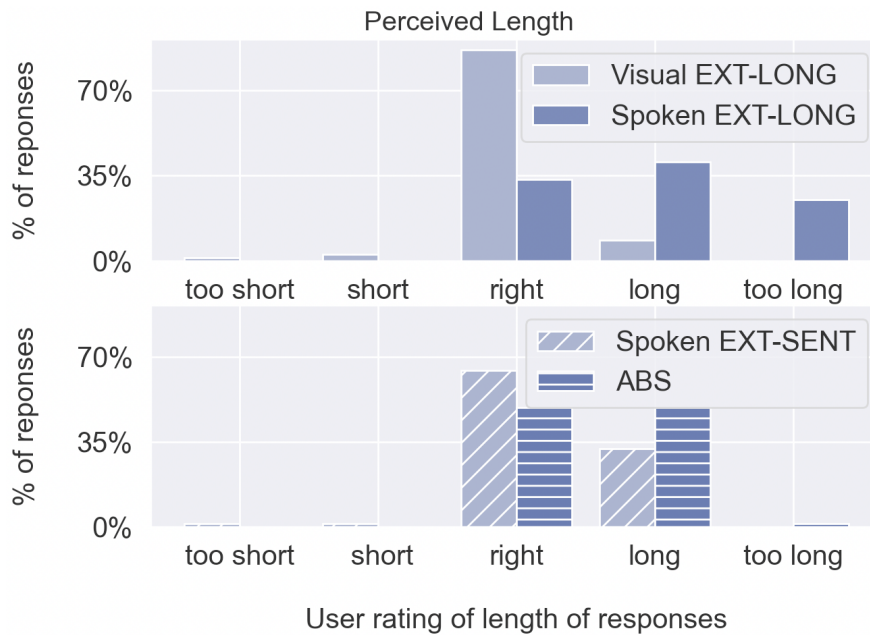


Figure 8.6: **Top:** Users perceive the same explanation to be longer in the spoken modality. **Bottom:** While EXT-SENT and ABS were the same length, participants rate the latter as longer more often perhaps because of they contain more content.

EXT-SENT in the same amount of time, might be more taxing in the working memory, making them less effective and in turn making users perceive them as being longer. These results are shown in Figure 8.6.

**PERCEIVED HELPFULNESS** Participants were asked whether the responses helped them in their decision making. Their responses showed that CONF and all explanation conditions were perceived as helpful by at least 80% of participants, with no real differences among them except for EXT-LONG in the visual modality (which is perceived helpful by close to 90% of users). Interestingly, 50% of participants indicated BASE to be helpful. In contrast, our results in Figure 8.3 show that different explanations actually differ in their eventual helpfulness. These results suggest that subjective measures can sometimes correlate with actual performance when the differences are large, but for the most-part and smaller differences, the result from subjective rating can be unreliable. These findings align with prior observations made (Bućinca et al., 2020) that showed that evaluating explanations on proxy metrics can lead to incorrect conclusions. These findings are shown in Figure A.4 in Appendix A.4.5.

**USER FEEDBACK** In order to get more specific details about how to improve the presentation of information, we asked participants at the end of the survey: *Do you have any additional*

*feedback on what the system can improve?* Two annotators read through about 400 responses across all conditions, and created codes to capture possible areas of improvement. The annotators then used these codes to classify responses. Many users gave feedback that was not insightful (e. g. "can't think of anything to improve"). After removing such responses, 175 responses remained for the final analysis. We computed the inter-annotator agreement using Cohen's  $k$  ( $k=0.74$ ). Here we briefly describe the most interesting findings, with more details about the codes we used and additional results in Appendix A.4.5.

In *BASE*, where the answer was provided with no additional information, about 50% of **participants mentioned that they would have liked it if the voice changed with system certainty**. In *CONF*, around 30% of participants give this feedback as well. Interestingly, for explanation conditions, this feedback is not seen as often.

For *EXT-SENT* in both modalities, *EXT-LONG* in the visual modality and *ABS*, 10-35 % of **participants would like the level of detail to adapt to the model certainty**. More specifically, users would like to have more details or additional answers *only* when the model is not confident in the prediction. This strategy seems similar to *adaptive explanations* proposed by (Bansal et al., 2020).

For the *EXT-LONG* condition in the *spoken modality* the feedback was mostly about the length of the responses. 78% of participants mentioned that responses should be shorter, which aligns with the higher perceived length of the explanations in Figure 8.6. For the *visual modality*, 40 % of participants mention that highlighting some key items would have made it even easier and faster. In fact, introducing highlights would improve the visual interface and would likely increase the differences in modalities that we already observe.

Finally, for all explanation conditions, 20-45 % of **participants would like to see more than one source containing an answer**. This means that the system would need to find multiple sources that converge to the provided answer. To provide users with this additional information without overloading cognitive capacity, an interactive strategy can be adopted. For example, evidence and additional sources can be presented through an explanatory dialogue (Miller, 2019), where users are initially provided with limited information, and more can be provided upon request.

### 8.6.3 *What misleads users?*

To better understand how explanations mislead users and how they can be further improved, we analyzed cases leading to user



error. We compiled a set of unique questions alongside their frequency of errors across users in all explanation conditions.

A single annotator followed a similar coding procedure as previously described, where questions were analyzed in order to detect emergent error categories. Following this initial analysis, questions were categorized into error types. We found that users tend to be misled on the same questions, with most of the errors happening on around 50 questions per condition, and about 40 of these questions overlapping across conditions.

Below we describe the three main cases:

**PLAUSIBLE EXPLANATIONS.** A concept is plausible if it is conceptually consistent to what is expected or appropriate in some context (Connell and Keane, 2006). Work has consistently identified that people often fail to evaluate the accuracy of information (Fan et al., 2020; Fazio and Marsh, 2008; Marsh and Umanath, 2013), particularly when no prior knowledge exists and information seems plausible (Hinze et al., 2014). We find many cases where a model response and explanation do not answer the question, yet the plausibility misleads users into accepting incorrect responses. For example:

**Question:** *Who is the patron saint of adoptive parents?*

**Response:** I am 37 percent confident that the answer is, **Saint Anthony of Padua**. I found the following evidence in a wikipedia passage titled, Anthony of Padua: Saint Anthony of Padua, born Fernando Martins de Bulhoes, also known as Anthony of Lisbon, was a portuguese Catholic priest and friar of the Franciscan order.

In the example above, the model is incorrect (true answer is Saint William of Perth), but users were often misled to accept this answer because the evidence makes the prediction sound plausible. Such errors make up 60 to 65% of the errors for all explanation conditions.

**LEXICAL OVERLAP.** In our error analysis, the second most common mistake (from 30 to 35% of errors) that both *the model* and *the users* make is related to the lexical overlap (McCoy, Pavlick, and Linzen, 2019) between the question and the evidence. For example:

**Question:** *How many teams are in the MLB national League?*

**Response:** I am 60 percent confident that the answer is, **30**. I found the following evidence in a wikipedia passage titled, Major

League Baseball: *A total of 30 teams play in the National League( NL) and American League (AL) , with 15 teams in each league .*

The evidence contains the correct answer (15 teams) but many users are misled by the phrase “*A total of 30 teams play in the National League*”.

**BELIEF BIAS.** Humans tend to rely on prior belief when performing reasoning tasks (Klauer, Musch, and Naumer, 2000). In model evaluation, this has consequences affecting validity. For example, if instructions are not specific, participants are left to use their beliefs to infer what is required of them, leading to varied interpretations. People often rely more heavily on belief bias when processing information under pressure (Evans and Curtis-Holmes, 2005), therefore in time limited evaluations this phenomenon might be more prominent. We reduced these confounds by carefully designing instructions, a straightforward interface, allowing workers plenty of time and removing ambiguous questions. However, some interesting cases of belief bias do occur — take, for instance, the example below:

**Question:** *Where is the longest bone in the body found?*

**Response:** I am 17 percent confident that the answer is, **femur**. I found the following evidence in a wikipedia passage titled, Femur: The Femur or thigh bone, is the proximal bone of the hindlimb in tetrapod vertebrates .

Such errors in our evaluation make about 3-5% of total errors in each explanation condition.

## 8.7 DISCUSSION

### 8.7.1 *Why Explanations Worked for ODQA?*

Unlike previous studies (Bansal et al., 2020; Chu, Roy, and Andreas, 2020; Hase and Bansal, 2020), we observed significant improvements from explanations over only communicating confidence. One reason for our positive results could be owing to the nature of ODQA i.e. unlike tasks such as sentiment classification, where humans may be able to solve the task without relying on explanations, ODQA requires satisfying a user’s information need, which may take considerably longer without explanations; users *require* additional help to navigate through vast amounts of information.

Another potential reason is, in ODQA, presenting a single good explanation can allow users to verify whether the prediction

is correct. In contrast, in sentiment analysis, even if the explanation points to evidence for a positive sentiment (“the smell was delicious”), there is always a chance that another phrase (“but the taste made me puke”) renders the net correct label as negative. It is worth noting that like previous works, not all of our explanation methods provide significant value (Figure 8.3); thus the success from showing explanations still cannot be taken for granted but should instead be measured using well-designed user studies.

### 8.7.2 *Implications and Recommendations*

Another interesting question is how can our findings inform future research in explainable NLP.

**DEVELOP MODALITY-SPECIFIC EXPLANATIONS** Our results showed that the best explanation varied across modalities, indicating that evaluating explanations on one modality (e. g., visual UI) and deploying them on another (e. g., voice assistant) can lead to sub-optimal deployment decisions. As a result, explanations should be optimized for and evaluated in the task and settings in which they will be deployed in-the-wild.

**FURTHER STUDY ABSTRACTIVE EXPLANATIONS** Longer explanations helped in the visual case, showing that communicating more evidence has potential to help users. However, they hurt in the spoken case, perhaps because longer explanations increase the cognitive load on users. This may indicate a trade-off between *information content* of explanation and its *cognitive load* for ODQA. We had hoped abstractive explanations would achieve a more optimal balance between fidelity and comprehensibility for spoken. However, Figure 8.3 shows that they did not improve the end performance. One reason is that even though abstractive explanations were concise, they had high information density and thus did not sufficiently decrease cognitive load.

That said, while abstractive explanations did not significantly improve accuracy compared to longer explanations, they did improve user speed at the task by 2.2 sec (Table A.8) and were satisfactory rated in terms of their perceived length compared to longer explanations (Figure 8.6). The utility of such generated explanations, over longer explanations, may further increase when explaining multiple sources (e. g., in Hotpot QA (Yang et al., 2018)) or candidate answers, where communicating multiple entire passages seems infeasible.

**ENABLE INTERACTIVE EXPLANATIONS** All explanation conditions we tested were static— they assumed a single trade-off between detail and conciseness. For example, `EXT-SENT` always conveyed a single sentence to the user as an explanation, which was concise but may not always convey all the context required to validate answers. A different strategy may be to use interactive explanations, in which the system first gives a concise explanation and then lets users request additional information. Such explanations may be especially used to accommodate user suggestions such as, including and explaining multiple candidate answers or multiple answer sources. Another possible strategy is to use *adaptive explanations*, where the model switches explanation strategies based on its confidence (Bansal et al., 2020).

**LIMITATIONS** While our user study addresses issues of many similar previous evaluations of explanations, it still has limitations. First, although the interaction of users with the QA system was kept as realistic as possible, in reality users may have the option to double-check the model’s answer using external tools, such as Web search. Accommodating that case would require encoding the additional cost of the reject action (e. g., due to time spent and effort) into the payoff. In addition, unlike an interaction in-the-wild questions were not posed by participants themselves, which may lead to different kinds of biases in the interpretation of the questions, as discussed before. Second, we conducted studies with MTurk workers who may not have the same motivation for performing the task as real users. To address this, we incentivized them by rewarding high-performance through bonuses and penalties specified by a payoff matrix. In practice, the values of the payoff matrix can vary depending on the stake of the domain and may vary with users. Finally, we only registered hypotheses that compared the performance on one metric— accuracy of error detectability. However, there may be other metrics that may be of interest e. g., improvements in speed and user satisfaction.

## 8.8 CONCLUSION

Contrary to recent user-studies for other tasks (such as classification), ours suggest that for ODQA, explanations of model’s predictions significantly help end-users decide when to trust the model’s answers over strong baselines such as displaying calibrated confidence. We observed this for two scenarios where users interact with ODQA model using spoken or visual modalities. However, the best explanation may change with the modal-

ity, e. g., due to differences in users' cognitive abilities across modalities. For example, for the spoken modality, concise explanations that highlight the sentence containing the answer worked well. In contrast, for the visual modality, performance improved upon showing longer explanations. Thus, developers and researchers of explainable ODQA systems should evaluate explanations on the task and modalities where these models will be eventually deployed, and tune these explanations while accounting for user needs and limitations.

Despite the success of explanations on our domain, explanations sometimes still mislead users into trusting an incorrect prediction, and sometimes as often as displaying the baseline. These results indicate the need to develop better explanations or other mechanisms to improve appropriate user reliance on ODQA agents, e. g., by enabling abstractive explanations that balance conciseness and detail while taking into account the user's cognitive limitations, interactive explanations that can explain multiple answer sources and candidates and adaptive explanations where the model strategy changes based on its confidence.



## Part IV

# DISCUSSION AND CONCLUSION





## DISCUSSION OF THE CONTRIBUTIONS

---

The preceding chapters have introduced work that falls within the framework of *human-centered NLP*, a general perspective to building language technology which must consider humans, society and the impact that it has on both. My research has gradually acknowledged that within a human-centered framework, a more interdisciplinary understanding of humans, their needs, and society, is required.

The work during my studies has specifically been concerned with improving human-AI interaction by (1) improving the predictive performance of dialogue systems utilizing user interactions and feedback signals, and (2) introducing methods for gender bias detection in NLP models and better evaluations of model transparency. This section revisits the main research questions along these dimensions, which were introduced in chapter 1, and reviews how the chapters in this thesis contribute to answering those questions.

The first two studies presented in this dissertation are concerned with improving Human-AI interaction via goal-oriented dialogue systems. Systems which can learn from users in an appropriate and safe manner will not only increase user trust on the system but will likely improve the quality of human-AI interaction over time (Amershi et al., 2019b). The studies in chapter 3 and 4 answer the following questions along this dimension:

*How can systems leverage previous user interactions to improve relevancy and fluency of answers?*

Traditionally, goal-oriented dialogue systems have relied on templated NLG modules due to the difficulty of obtaining the large amounts of data needed for sequence-to-sequence models. In addition, as opposed to chit-chat models, the responses from goal-oriented dialogue systems must be focused and relevant. Chapter 3 is one of the early works showing that it is possible to use more flexible sequence-to-sequence neural models to generate responses which are deemed more relevant and fluent by humans in multi-domain goal-oriented dialogue systems. This was done by incorporating similar previous user interactions as a simple-yet-powerful prior for the NLG module. With the advancement of pretrained models in the last couple of years, and models specifically pretrained for dialogue, the starting

performance for systems is much higher. However, this simple technique may still prove to help systems remain on topic in domain-restricted scenarios.

*How can dialogue-level user feedback help dialogue systems further improve and adapt to new domains?*

Chapter 4 investigated how to integrate a user feedback signal which is more natural to obtain during a human-machine interaction. Throughout the PhD, the disconnect between the systems being developed and the humans who interact with them has become more clear; often the systems do not match the usability expectations of the user. One of the early realizations was that RL methods for dialogue rely on reward signals which deviate strongly from real world use. As mentioned in chapter 1, work in HCI has pointed out the importance of feedback strategies matching the human expectations (Cakmak, Chao, and Thomaz, 2010). They find that asking for too much feedback in the course of a dialogue is perceived by users as imbalanced and annoying, hurting the effectiveness of systems. Chapter 4 employed a feedback signal which is collected at the end of the dialogue which is more in line with real-world interactions. The results showed that it is possible to leverage signals which match user expectations to be able to better generalize to new domains and further improve performance in-domain.

As was mentioned in part i, the second part of the thesis focused on investigating topics such as fairness and transparency, as these are crucial to ensuring ethical and responsible deployment of technologies. This part of the dissertation introduced more interdisciplinary approaches and knowledge from cognitive science, linguistics, psychology, and HCI to investigate how humans and NLP systems interact. The contributions in this later work are mostly in terms of evaluation of NLP models. The next four questions are answered throughout chapters 5–8.

*How can we diagnose negative social biases in multilingual systems, which currently make the adoption of systems by end-users difficult?*

Systems which are built within the human-centered framework must be created with awareness of the impact on society and humans. This means that NLP systems must not only display the predictive power that comes from AI methods, but should also be responsibly deployed and match socially-safe behaviors, should not exacerbate existing inequities, and should promote inclusivity. Chapter 5 is the first study to introduce a diagnostic

testbed for gender bias in languages other than English, exploiting a linguistic phenomenon which exists in some (non-English) languages. Furthermore, while previous works have focused on coreference resolution, the study presented in this thesis introduced three different NLP tasks in addition to coreference resolution including Language Modeling (LM), Machine Translation (MT), and Natural Language Inference (NLI). We showed that popular and frequently deployed models do exhibit behaviors which align with gender stereotypes that must be mitigated. We also demonstrated the importance of looking beyond English for linguistic clues on how to best diagnose negative social biases in multilingual NLP systems.

*To what extent do human cognitive biases and previous world knowledge affect the explainability of NLP systems, and does this change the conclusions we make about the best performing methods?*

Model transparency and explainability in ML models is typically evaluated using automatic metrics which may not reflect how methods are used in the wild or how they interact with humans in the real world. To inform ourselves on those dimensions, human evaluations are vital. However, as mentioned throughout chapters 6 and 7, previous evaluations do not properly account for the cognitive biases playing a large role in the decision making process of users. In chapter 6, we presented a small pilot experiment showing that for classification tasks, several cognitive biases can interact with evaluation of explainability. When such cognitive biases are reduced, the positive effect of a popular explainability method becomes much smaller for the task of predicting model decisions. This work was extended in 7; here, we are the first to bring the discussion of *belief bias* from the field of psychology to NLP. We observed how such bias plays a role in human evaluation of explainability for two previously used evaluation protocols. We found that compared to standard setups from previous works, introducing conditions which account for humans' belief biases altered the main conclusions about which models work best. This is an essential finding, as not accounting for such biases can potentially mislead the community to develop more methods which may not be optimal or robust during real interactions with humans. We presented several recommendations and insights to help NLP practitioners develop better evaluation paradigms which take more knowledge from psychology and properly account for humans' cognitive biases.

*Which natural language explanations help users in a real world down-*

*stream decision making task such as deciding when to trust a model prediction and does the effectiveness of explanations change with presentation modality (eg. voice vs. visual)?*

In chapter 8, we further investigated explainability and presented a human evaluation where we assessed the effectiveness of natural language explanations in helping users decide when to trust (reject or accept) model decisions for the task of ODQA. We were interested in simulating a real life scenario, therefore we were the first to evaluate the effectiveness of explanations across two presentation modalities; in the wild users will interact with systems via a voice assistant or a visual display. Unlike previous studies evaluating tasks like sentiment analysis, our findings showed that *some* explanation strategies are more effective than showing model confidence, but similarly to prior work, the majority still significantly mislead humans into trusting wrong model predictions. This result showed that there is vast room for improvement and that NLP researchers should not take the effectiveness of explanations for granted. In addition, the effectiveness of explanation strategies varied with presentation modality due to the differing cognitive limitations imposed on users in visual versus spoken interfaces. This finding points out that it may be important to build NLP systems or explanation strategies which take presentation modality into account; this requires more interdisciplinary collaboration between NLP researchers and researchers in other fields such as HCI and cognitive science.

*What explanation strategies would users prefer to be presented with?*

In chapter 8, we also collected extensive qualitative data to accompany our quantitative studies, to obtain a better understanding of what users would prefer and to drive our future work. This also allowed us to better understand their experience with our interface and make better conclusions. Overall, we found that users tend to prefer adaptive explanation scenarios where the model changes its strategy based on model confidence. In addition, users mentioned that they would like the model to be more interactive and change their explanation strategy based on whether the user asked for more information or not. These findings are extremely valuable to both NLP and HCI: they indicate that there are various directions for explainability work which are founded on *real* user feedback and experience.

## FUTURE DIRECTIONS

---

The work presented throughout this dissertation has *gradually* and *increasingly* emphasized a need to consider the impact that technology has on users and society. While the earlier studies emphasized the development of new models which aimed to be more user-centric, the later work focused on model evaluation and incorporates knowledge about humans and society which spans various fields. In this closing chapter, I elaborate on some reflections and conclusions based on my work throughout the PhD. More specifically, (1) I briefly revisit dialogue systems and discuss some opportunities for new work, extending even further within the human-centered framework, and (2) conclude with a short discussion of how the insights from my studies will inspire my future research.

**REVISITING DIALOGUE SYSTEMS** Despite my earlier research showing that predictive performance can be improved by incorporating *human* signals, work in dialogue systems can be extended in a more human-centered direction. For example, recent work suggests that in order to create **AI** systems which collaborate with humans, the systems must be trained to optimize for human-machine team performance (Bansal et al., 2019b, 2020). Dialogue systems may be useful for improving human-machine teamwork, therefore, devising dialogue systems which optimize using team-based metrics might be an interesting avenue for future work.

Within explainable **AI**, Miller (2019) mentions that *explanatory dialogues* might be better suited for humans' needs for explanations, arguing that explanations will often be needed in collaborative environments. This suggestion is supported by the findings in this thesis, which show that humans prefer explanation strategies which are collaborative, adaptive, and interactive. This is another interesting direction for future work.

**FUTURE RESEARCH** Chapter 5 showed that looking into linguistic phenomena which do not occur in English, and looking at occupational statistics from different countries, could inform us of how multilingual systems encode gender stereotypes. This study left plenty of room for further work which incorporates knowledge from linguistics, psychology and social sciences to devise new ways of detection of social biases across many different languages.

In terms of explainability, user studies across different domains may inform us of what type of explanations are needed by humans in more specific use cases. In chapter 8, we showed that, unlike previous works, *some* explanations actually helped users detect model errors better against simply showing model confidence. One interesting area of work is to uncover in which domains explanations are actually useful. This requires more controlled user studies which simulate real decision-making scenarios. This is not only an HCI concern, I argue this is vital for the field of NLP; the development of explainability methods should prioritize the domains which are more critical and which require explanation the most.

Part V

APPENDIX





## APPENDIX

## A.1 CHAPTER 5

## A.1.1 Example Data

We present example data for each task.

**NLI** For each occupation, template and language in our NLI benchmark, we have three premise-hypothesis pairs. The hypotheses are identical. For the premises, we have three versions, one for each pronoun (the non-gendered reflexive possessive pronoun and the two gendered, anti-reflexive possessive pronouns). Examples with reflexives are labeled entailment; examples with anti-reflexives are labeled neutral. An example datapoint for this task, with the label entailment, is shown in Table A.1.

Natural language inference	
A) PREMISE AND B) HYPOTHESIS	
da	a) Ingeniøren tager <i>sine</i> briller fra skrivebordet b) Ingeniørens briller tages fra skrivebordet
sv	a) Ingenjören tar <i>sina</i> glasögon från skrivbordet b) Ingenjörens glasögon tas bort från skrivbordet.
ru	a) Inzhener snimayet <i>svoi</i> ochki so stola. b) Ochki inzhener snyaty so stola.
zh	a) Gōngchéngshī cóng zhuōzi shàng zhāi xià <i>zìjǐ</i> de yǎnjìng b) Gōngchéngshī de yǎnjìng yǐ bèi cóng zhuōzi shàng yí kāi.
ENGLISH TRANSLATION	
en	a) The engineer takes <i>his/her</i> glasses from the desk b) The engineer's glasses are taken from the desk

Table A.1: Example data for NLI. For NLI, we only generate entailments and neutral statements. The English translation is shown for reference only.

**MACHINE TRANSLATION** For machine translation, we have 4560 pairs of source sentences with masculine (*his*) and feminine possessive pronouns (*her*), respectively. We translate these into the target languages using off-the-shelf models and assess the tendency of these models to predict reflexive possessive pronouns in the target languages, instead of anti-reflexive possessive pronouns. An example of the data can be found in Table A.2.

MACHINE TRANSLATION	
SOURCE SENTENCE	
en	The engineer takes <i>his/her</i> glasses from the desk
TRANSLATIONS	
da	Ingeniøren tager <i>sine</i> briller fra skrivebordet
sv	Ingenjören tar <i>sina</i> glasögon från skrivbordet
ru	Inzhener snimayet <i>svoi</i> ochki so stola.
zh	Gōngchéngshī cóng zhuōzi shàng zhāi xià <i>zìjǐ</i> de yǎnjìng

Table A.2: Example data for machine translation.

**COREFERENCE RESOLUTION** For coreference resolution, we are interested in whether the model is more likely to cluster a masculine possessive pronoun with the subject of the sentence than a feminine pronoun, even when this reading violates grammatical constraints. In Table A.3, we list examples of how the task data would look. In brackets, we have mentions of an entity that can be clustered together by the system as belonging to the same coreference chain.

COREFERENCE RESOLUTION	
da	[Ingeniøren] tager [ <i>sine/hans/hendes</i> ] briller fra skrivebordet
sv	[Ingenjören] tar [ <i>sina/hans/hennes</i> ] glasögon från skrivbordet
ru	[Inzhener] snimayet [ <i>svoi/yego/yeye</i> ] ochki so stola.
zh	[Gōngchéngshī] cóng zhuōzi shàng zhāi xià [ <i>zìjǐ/tā/tā</i> ] de yǎnjìng
ENGLISH TRANSLATION	
en	[The engineer] takes [ <i>his/her</i> ] glasses from the desk

Table A.3: Example data for coreference resolution. In brackets, we have the mentions that the system could cluster as coreferent. We include an English translation only for reference.

**LANGUAGE MODELING** For language modeling, we take a sentence containing a reflexive pronoun and swap the reflexive for the possessive masculine and feminine anti-reflexives; we then compute the perplexities of the original and perturbed

sentences. Example of how this is framed can be found in Table A.4.

LANGUAGE MODELING	
da	<b>Truth:</b> Ingeniøren tager <i>sine</i> briller fra skrivebordet <b>Prediction(Fem):</b> Ingeniøren tager <i>hendes</i> briller fra skrivebordet <b>Prediction(Masc):</b> Ingeniøren tager <i>hans</i> briller fra skrivebordet
sv	<b>Truth:</b> ingenjören tar <i>sina</i> glasögon från skrivbordet <b>Prediction(Fem):</b> ingenjören tar <i>hennes</i> glasögon från skrivbordet <b>Prediction(Masc):</b> Ingenjören tar <i>hans</i> glasögon från skrivbordet
ru	<b>Truth:</b> Inzhener snimayet <i>svoi</i> ochki so stola. <b>Prediction(Fem):</b> Inzhener snimayet <i>yeye</i> ochki so stola. <b>Prediction(Masc):</b> Inzhener snimayet <i>yego</i> ochki so stola.
zh	<b>Truth:</b> Gōngchéngshī cóng zhuōzi shàng zhāi xià <i>zìjǐ</i> de yǎnjìng <b>Prediction(Fem):</b> Gōngchéngshī cóng zhuōzi shàng zhāi xià <i>tā</i> de yǎnjìng <b>Prediction(Masc):</b> Gōngchéngshī cóng zhuōzi shàng zhāi xià <i>tā</i> de yǎnjìng

Table A.4: Example data for the language modeling task

### A.1.2 Coreference Dataset Statistics

In table A.5 we show the number of documents used to train each system. For Chinese, the data is available with predefined train, development and test sets. For Russian, however, this is not specified, therefore we split the data 80-20-20.

Lang	Training	Dev	Test
zh	1810	252	218
ru	144	18	18

Table A.5: Statistics for the coreference data used for training.

## A.2 CHAPTER 6

**PRESENTATION OF STIMULI** We created a web application using Flask<sup>1</sup> in order to collect participant data. Participants would randomly get assigned a known or unknown task and LIME explanations or no explanations. For all tasks we provide the same general instructions. In addition, we had task specific instructions. For known tasks we provided short descriptions of the task, while emphasizing the fact that subjects should imitate

<sup>1</sup> <https://flask.palletsprojects.com/en/1.1.x/>

the model rather than follow their own opinions about the true labels.

The training and evaluation sessions were almost the same, with the only difference being that during training, subjects could check the model’s answer after making an initial guess. See Figure A.1 for an example of what the items looked like. The example here is for the task of sentence length prediction using LIME explanations.

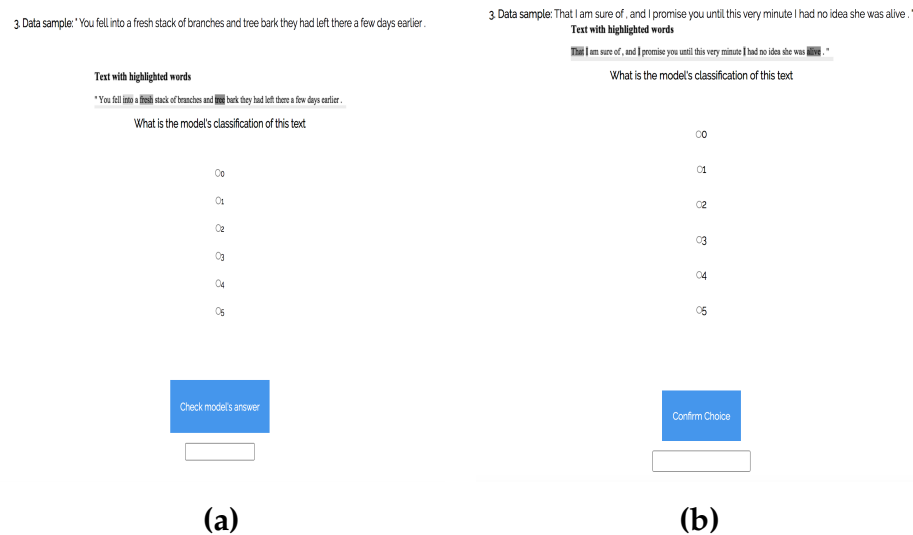


Figure A.1: **(a)** Example of item in the training session for sentence length prediction. Note that the participants are able to check the model answer **(b)** Example of item in the evaluation session for sentence length prediction. Here the participants are no longer able to check the model answer

### A.3 CHAPTER 7

#### A.3.1 *Experiment 1: Human Forward Prediction*

Below we show the instructions provided to the participants, as well as an example of the saliency maps presented to participants for adversarial examples.

**INSTRUCTIONS.** Question-answering systems are a particular form of artificial intelligence. The task here is for you to learn to predict how the system answers questions. In other words, when in a bit, you are presented with questions, the task is not to provide the right answer, but to guess the answer the system provided. For each question, you will also see a context paragraph. The answer is a span of text in this paragraph. Instead of writing out the answer, you can simply mark the relevant span.

If you want to select a new answer, please click *reset answer*, if you are ready to see the model answer, please click *show answer*. Note that your answer will lock at that time.

**RAW RESULTS.** In our evaluation, we use the first 15 points as training, therefore, we discard them from the main evaluation but show them in this section. Overall, we see that training, for the most part has a positive effect, or not so much of an effect. These scores can be seen in Table A.6.

	MODEL	Human		
		<b>Baseline</b>		
LOW-ORIG	0.17	0.14	0.52	52.27
LOW-ADV	0.15	0.10	0.36	54.36
HIGH-ORIG	0.79	0.53	0.58	37.12
HIGH-ADV	0.66	0.35	0.48	47.64
		<b>Integrated (IG)</b>		
LOW-ORIG		*0.34	0.35	41.68
LOW-ADV		*0.36	0.28	44.38
HIGH-ORIG		*0.71	0.76	46.87
HIGH-ADV		0.46	0.47	42.99
		<b>Gradients</b>		
LOW-ORIG		* <b>0.64</b>	*0.09	*32.16
LOW-ADV		* <b>0.63</b>	0.23	*30.05
HIGH-ORIG		* <b>0.82</b>	*0.84	44.65
HIGH-ADV		* <b>0.57</b>	*0.62	*52.30

Table A.6: Raw scores, before removing data points on training session

### A.3.2 Experiment 2: Best Model Selection

Below we show the instructions given to the participants, and more details about the qualitative analysis of the feedback we obtained.

**INSTRUCTIONS.** Question-answering (QA) systems are a particular form of artificial intelligence. We have trained two QA systems and have extracted the most important words the model uses to make its final decision. Based on these highlighted words, your task is to select the model that you think is more likely to

QUALITATIVE CODES	EXAMPLES
Irrelevant (q)	<ol style="list-style-type: none"> <li>1. Model A only extracted some important words but also some punctuations in the question which is insufficient to derive to a good answer. Model B extracted a number of key important words that would lead to the correct answer.</li> <li>2. Option b chose quantitative statements, while option A seems confused about what it's looking for since it highlights all sorts of things in the question.</li> </ol>
main entity (q)	<ol style="list-style-type: none"> <li>1. The words "year" and "norman" in the question were not extracted by Model A. The Model will not be able get the correct answer without knowing what to look for.</li> <li>2. The question was asking about the year lavoisier's work was published but neither of the key words in this question were highlighted. Model A had no idea where to locate the answer without considering those key words.</li> </ol>
main entity (a)	<ol style="list-style-type: none"> <li>1. The answer requires a year; it hasn't highlighted any years as part of the answer.</li> <li>2. Answer needed to be a name and option A chose nothing that could be a name.</li> </ol>
Irrelevant (a)	<ol style="list-style-type: none"> <li>1. Model B has highlighted many extra words in the answer</li> <li>2. Both models selected the correct terms, but model A selected more irrelevant terms in the answer too, so it's less likely to choose the correct one from those numerous options.</li> <li>3. B highlighted the answer but also too much unneeded info.</li> </ol>
Generic	<ol style="list-style-type: none"> <li>1. Model A does not highlight the right answer</li> <li>2. Model B is wrong and model A is correct</li> </ol>

Table A.7: Examples of some of the feedback categorized into these classes

perform best. Additionally, please write how the low-performing model fails and/or how it could be better (try to be detailed)

QUALITATIVE ANALYSIS OF FEEDBACK. In Table A.7, we include a few examples of the sentence that were categorized using the qualitative codes. Unsurprisingly, once participants found a strategy for giving feedback, they mostly stuck to it.

After categorizing all the feedback into each category, we visualize the distribution per condition. This can be found in Figure 7.4. We find that for the HIGH vs Low conditions, the distribution is very similar between gradients and integrated gradients. Many participants gave very generic feedback, for example by simply saying that "model A is better because it is correct, and model B is wrong". This was not surprising, as here the differences were supposed to be clear and it is likely most participants did not have to think too hard before making a decision. However, the distribution is very different for the HIGH vs MEDIUM conditions. Here, for standard gradients, the feedback followed a similar pattern as in the previous condition, but about 30% less examples received generic feedback than before. For integrated gradients, most examples received feedback regarding the irrelevant terms being highlighted, showing that

even when the difference in performance between models is large (20 F1 points), this method makes the distinction difficult for the best model selection task.

## A.4 CHAPTER 8

### A.4.1 *Temperature Scaling*

Temperature scaling (Guo et al., 2017), a multiclass extension of Platt Scaling (Platt et al., 1999), is a post-processing method applied on the logits of a neural network, before the softmax layer. It consists of learning a scalar parameter  $t$ , which decreases or increases confidence.  $t$  is used to rescale the logit vector  $z$ , which is input to softmax  $\sigma$ , so that the predicted probabilities are obtained by  $\sigma(z/t)$ , instead of  $\sigma(z)$ .

In our experiments, the model is set to pick from the top 100 solutions, however, in many cases the correct answer occurs within the top 10 items. For our purposes we calibrate the confidence scores of the top 10 outputs. We use the publicly available scripts provided by Guo et al. (2017).<sup>2</sup>

The model confidence before and after calibration can be seen in Figure A.2.

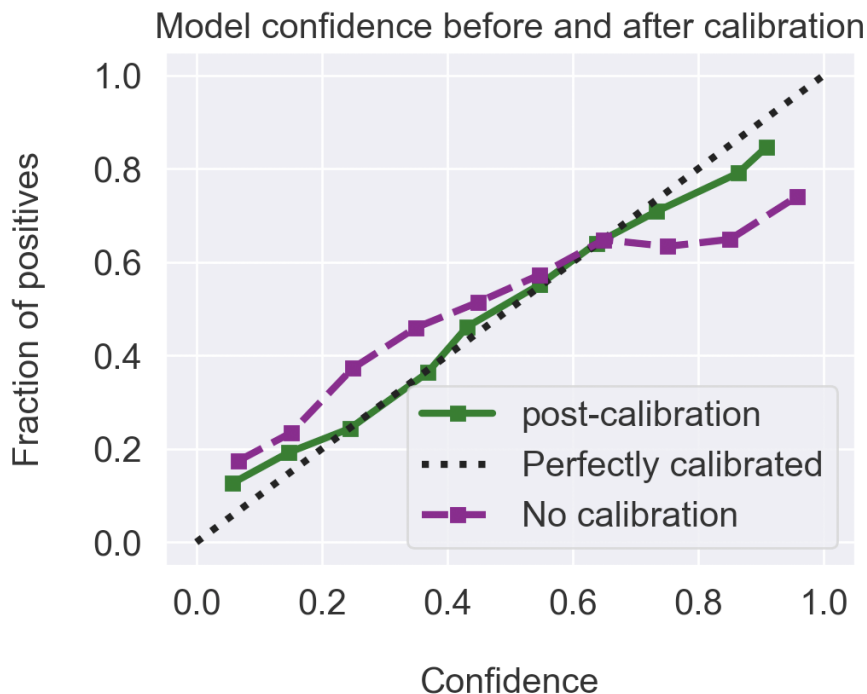


Figure A.2: Confidence before and after calibration.

<sup>2</sup> [https://github.com/gpleiss/temperature\\_scaling](https://github.com/gpleiss/temperature_scaling)

#### A.4.2 Additional Preprocessing

Additional preprocessing to ascertain the quality of stimuli in each modality was required; more details can be found in Appendix . Before sampling questions for the task, to ensure a high-quality and non-ambiguous experience for MTurk workers, we manually filter out several “problematic” questions:

- **Ambiguity in the question:** For various questions in NQ, multiple answers can exist. For example, the question: *when was King Kong released?*, does not specify which of the many King Kong movies or video games it refers to. These cases have been known to appear often in NQ (Min et al., 2020). We remove such questions from our subset.
- **The gold answer was incorrect:** Many examples in NQ are incorrectly annotated. As it is too expensive to re-annotate these cases, we remove them.
- **Answer marked incorrect is actually correct :** We present both correct and incorrect questions to users. There are cases where the predicted answer is marked incorrect (not exact match) but is actually correct (a paraphrase). We manually verify that correct answers are paired with contexts which support the answer.
- **Correct answer but incorrect evidence:** The model sometimes, though not as often, chooses the correct answer but in the incorrect context. We discarded examples where the explanation was irrelevant to the question e.g. *who plays Oscar in the office? Oscar Nuñez, is a Cuban-American actor and comedian..* In order to be able to make more general conclusions about whether explanations help in , we restrict our questions to ones containing correct answers in the correct context.
- **Question and prediction do not match type.** We removed cases where the question asked for a certain type e.g. a date, and the prediction type did not match e.g. a location.

In the visual modality, to ensure readability, we fixed capitalizations. For the spoken modality, to ensure fluency and clarity, we manually (1) inserted punctuation to ensure more natural sounding pauses, and (2) changed abbreviations and symbols to a written out form e.g. *\$ 3.5 billion* to *3.5 billion dollars*.



### A.4.3 *Task Setup: Additional details*

**PLATFORM AND PARTICIPANT DETAILS** We conduct our experiments using Amazon Mechanical Turk<sup>3</sup>. We recruited 525 participants in total, with approval ratings greater than 95 % and had a maximum of 8 days for approval of responses in order to minimize the amount of spamming.

We use a random sample of 120 questions from our dataset which remains the same across all conditions. In order to keep each session per participant at a reasonable time and ensure the quality of the data wouldn't be affected by workers becoming exhausted, we opted for three fixed batches of 40 questions, all split as 50 % correct and 50 % incorrect. Workers could only participate once (only one batch in one condition). Participants took around from 35-45 minutes to complete the HITs, but were given up to 70 minutes to complete.

We monitored if their screen went out of focus, to ensure that participants did not cheat. We ensured that we had 25 user annotations per question. When analyzing the data, we remove the first 4 questions of each batch, as it may take participants a few tries before getting used to the interface. In the end, we collect about 21,000 test instances.

**TASK INSTRUCTIONS** Imagine asking Norby a question and Norby responds with an answer. Norby's answer can be correct or wrong. If you believe Norby's answer is correct, you can accept the answer. If you believe it is wrong, you can reject it. If the answer is actually correct and you accept it, you will earn a bonus of \$0.15. But, if the answer is wrong, and you accept it, you will lose \$0.15 from your bonus. If you reject the answer, your bonus is not affected. (Don't worry, the bonus is extra! Even if it shows negative during the experiment, in the end the minimum bonus is 0). In total you will see 40 questions in this HIT (you will only be allowed to participate once) and the task will take about 40 to 45 minutes. You can be compensated a maximum of \$13.50 for about 40-45 minutes of work. Some things to note:

1. You must listen to the audio before the options become available.
2. If you make it to the end there is a submit button there, however, in case of an emergency you can hit the quit early button above and you will get rewarded for the answers you provided.

<sup>3</sup> <https://www.mturk.com/>

3. You can play the audio as many times as you need but as soon as you click a choice you will be directed to the next item.
4. **IMPORTANT!!** Please do not look up questions in any search engine. We will monitor when the screen goes out of focus, so please keep the screen on focus or you might risk being rejected.
5. Finally, please do not discuss answers in forums; that will invalidate our results.

#### A.4.4 *Post-task survey*

1. I found the CLARITY of Norby’s voice to be:  
(a) Excellent (b) Good (c) Fair (d) Poor (e) Very Poor
2. I found Norby’s responses to be HELPFUL when deciding to Accept or Reject:  
(a) Strongly Agree (b) Agree (c) Undecided (d) Disagree (e) Strongly Disagree  
Can you give a few more details about your answer?
3. I found the LENGTH of Norby’s responses to be:  
(a) Too Long (b) Long (c) Just right (d) Short (e) Too short
4. No AI is perfect and Norby is no exception. We are interested in helping Norby provide responses that can help users to determine whether to trust it or not (to accept or reject, just as you have done in this experiment). From your interaction with Norby, **do you have any additional feedback on what it can improve?**

#### A.4.5 *Results*

**REWARD** We compute the differences in overall reward for each condition. We observe the same trends as we discussed for accuracy. More specifically, all explanation conditions improve the final user reward, with `EXTRACTIVE-SENT` performing best in the spoken modality and `EXTRACTIVE-LONG` performing best overall. These differences are shown in Figure [A.3](#).

**TIME DIFFERENCES** We measured the time (in seconds) that it took participants to complete each question. In Table [A.8](#) we present the median times averaged over all workers per condition. We also include an adjusted time, subtracting the length of the audio, in order to measure decision time.

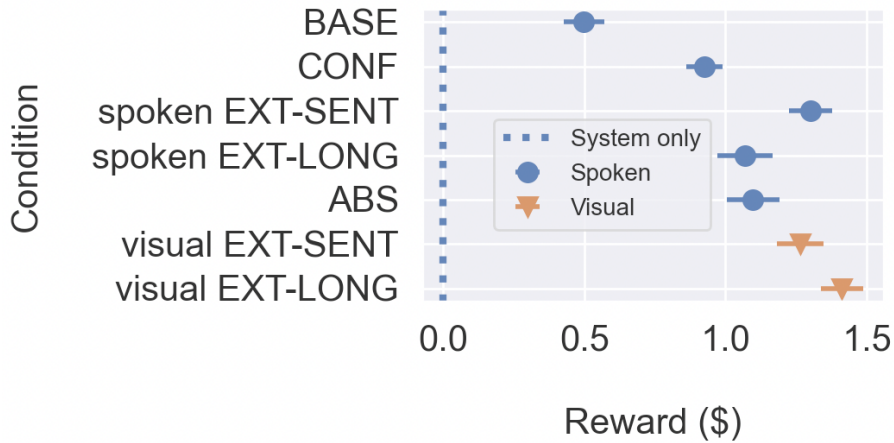


Figure A.3: **Reward**: The scores presented here are out of \$ 2.70. Although all explanations are better than `CONFIDENCE`, the explanations leading to the highest rewards change across modalities.

CONDITION	SEC/QUESTION	ADJUSTED
Spoken Modality		
Baseline	$10.2 \pm 1.6$	$8.3 \pm 1.6$
Confidence	$9.4 \pm 1.5$	$6.0 \pm 1.5$
Abstractive	$24.4 \pm 1.5$	$7.0 \pm 1.4$
Extractive-long	$44.9 \pm 1.6$	$9.2 \pm 1.6$
Extractive-sent	$24.3 \pm 1.7$	$7.6 \pm 1.7$
Visual Modality		
Extractive-long	$16.1 \pm 1.7$	-
Extractive-sent	$10.4 \pm 1.1$	-

Table A.8: Time differences across modalities. Time differences in the right column have been adjusted by removing the duration of the audio files. We observe that with additional information, users can make faster decisions than the `BASILINE` condition.

**HELPFULNESS** Differences in perceived helpfulness are shown in Figure A.4.

**USER FEEDBACK** Users provided free-form written feedback on possible ways to improve the system. The prompt they saw was: *do you have any additional feedback on what the system can improve?* After converging on a final set of codes, two annotators coded up about 400 responses across all conditions. The codes and their descriptions can be found in Table A.9. The codes are *not* mutually exclusive.

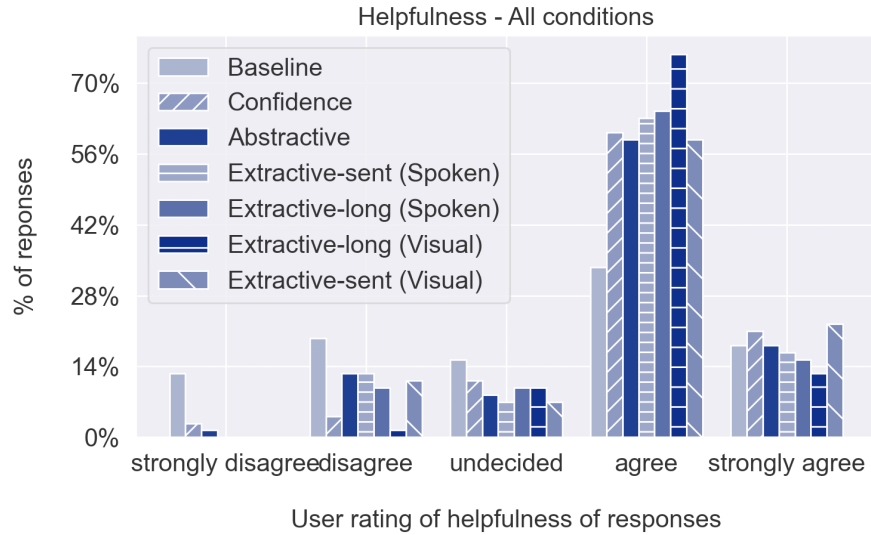


Figure A.4: **Helpfulness:** Participants indicated how helpful responses were. These results reflect the large differences we see in performance (BASELINE vs the rest of the settings), but are not able to capture the more subtle differences among explanation strategies and CONFIDENCE.

CODE	DESCRIPTION	CATEGORY
len-conciseness	users wish explanation was shorter	improvement on length
len-expand	users wish explanation was shorter	
adapt-detail	users wish details adapted with confidence	adaptability feature
adapt-voice	users wish voice adapted to confidence	
pres-change-confidence	users wish confidence would be communicated differently e.g. the answer is probably...	improve presentation
pres-highlighting	users wish important facts would be highlighted	
need-more-sources	users wish more source were provided	need additional info
need-confidence	users wish confidence was provided	
need-source	users wished a source was provided	
need-explanation	users wish an explanation would be provided	
need-link	users wish a link was provided	
need-multiple-answers	users wish more than 1 answer was provided	

Table A.9: The codes used to uncover areas of improvement from the post-experimental user feedback.

We found that many users across most conditions, would like **adaptability features** added. Additionally, we found that participants would like to be provided with multiple sources which converge on the answer. We also observe that for spoken

CONDITION	CODE	% PARTICIPANTS
baseline	adapt-voice	50
	need-confidence	36
	need-explanation	25
	need-source	17
confidence	need-explanation	38
	adapt-voice	29
	pres-change-confidence	14
	adapt-detail	10
	need-multiple-answers	10
	need-link	5
extractive-sent (spoken)	need-more-sources	44
	adapt-detail	28
	len-conciseness	22
	need-multiple-answers	17
	need-link	11
	len-expand	11
	pres-change-confidence	6
extractive-long (spoken)	len-conciseness	78
	need-more-sources	15
	pres-change-confidence	4
abstractive	len-conciseness	52
	need-more-sources	22
	adapt-detail	22
	pres-change-confidence	13
	need-multiple-answers	4
extractive-sent (visual)	need-more-sources	33
	adapt-detail	33
	len-expand	27
	need-multiple-answers	7
extractive-long (visual)	pres-highlighting	40
	need-more-sources	33
	adapt-detail	10
	need-link	10
	pres-change-confidence	7

Table A.10: Distribution of codes across all conditions. Codes are **not** mutually exclusive.

conditions, **improvements on length** are mentioned more often. The full distribution of codes across conditions is shown in Table [A.10](#).

EXPLANATION TYPE	RESPONSE+EXPLANATION	MODALITY
Baseline	The answer is, <b>two</b> .	Spoken
Confidence	I am 41 percent confident that the answer is, <b>two</b> .	Spoken
Abstractive	I am 41 percent confident that the answer is, <b>two</b> . I summarized evidence from a wikipedia passage titled, Marco Polo (TV series). Netflix cancelled the show after two seasons, as it had resulted in a 200 million dollar loss.	Spoken
Extractive-sent	I am 41 percent confident that the answer is, <b>two</b> . I found the following evidence in a wikipedia passage titled, Marco Polo (TV series). On December 12, 2016, Netflix announced they had canceled "Marco Polo" after two seasons.	Spoken/visual.
Extractive-long	I am 41 percent confident that the answer is, <b>two</b> . I found the following evidence in a wikipedia passage titled, Marco Polo (TV series). On December 12, 2016, Netflix announced they had canceled "Marco Polo " after two seasons. Sources told "The Hollywood Reporter" that the series' two seasons resulted in a 200 million dollar loss for Netflix , and the decision to cancel the series was jointly taken by Netflix and the Weinstein Company. Luthi portrays Ling Ling in season 1, Chew in season 2. The series was originally developed at starz, which had picked up the series in January 2012.	Spoken/visual

Table A.11: **Explanation examples:** Example of how system responses looked for each explanation type and baseline, for the question *How many seasons of Marco Polo are there?*

#### A.4.6 Explanation Examples

In Table A.11, we show an example of how the responses and explanations looked for each of the conditions. We also indicate in which modalities each explanation is shown in our experiments.

## BIBLIOGRAPHY

---

- Ackerman, Mark, Volkmar Pipek, and Volker Wulf, eds. (2003). *Sharing expertise beyond knowledge management*. MIT Press.
- Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2018). "Sanity checks for saliency maps." In: *Advances in Neural Information Processing Systems*, pp. 9505–9515.
- Agarwal, Rishabh, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, and Geoffrey E. Hinton (2020). "Neural Additive Models: Interpretable Machine Learning with Neural Nets." In: *CoRR* abs/2004.13912. arXiv: 2004.13912. URL: <https://arxiv.org/abs/2004.13912>.
- Amershi, Saleema, Maya Cakmak, William Bradley Knox, and Todd Kulesza (2014). "Power to the people: The role of humans in interactive machine learning." In: *Ai Magazine* 35.4, pp. 105–120.
- Amershi, Saleema, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. (2019a). "Guidelines for human-AI interaction." In: *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13.
- Amershi, Saleema, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. (2019b). "Guidelines for human-AI interaction." In: *CHI*.
- Anderson, Richard B and Beth M Hartzler (2014). "Belief bias in the perception of sample size adequacy." In: *Thinking & Reasoning* 20.3, pp. 297–314.
- Asghar, Muhammad Zubair, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi (2014). "A review of feature extraction in sentiment analysis." In: *Journal of Basic and Applied Scientific Research* 4.3, pp. 181–186.
- Atanasova, Pepa, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein (2020a). "A Diagnostic Study of Explainability Techniques for Text Classification." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3256–3274.

- Atanasova, Pepa, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein (2020b). "Generating Fact Checking Explanations." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7352–7364.
- Auernhammer, Jan (2020). "Human-centered AI: The role of Human-centered Design Research in the development of AI." In:
- Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." In: *PloS one* 10.7, e0130140.
- Bannon, Liam (2011). "Reimagining HCI: toward a more human-centered perspective." In: *interactions* 18.4, pp. 50–57.
- Bansal, Gagan, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz (2019a). "Beyond accuracy: The role of mental models in human-AI team performance." In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 1, pp. 2–11.
- Bansal, Gagan, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz (2019b). "Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 2429–2437.
- Bansal, Gagan, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld (2020). "Does the whole exceed its parts? the effect of ai explanations on complementary team performance." In: *arXiv preprint arXiv:2006.14779*.
- Barocas, Solon and Andrew D Selbst (2016). "Big data's disparate impact." In: *Calif. L. Rev.* 104, p. 671.
- Barrouillet, Pierre, Sophie Bernardin, Sophie Portrat, Evie Vergauwe, and Valérie Camos (2007). "Time and cognitive load in working memory." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.3, p. 570.
- Barston, Julie Linda (1986). "An investigation into belief biases in reasoning." In:
- Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti (June 2019). "SemEval-2019 Task 5:



- Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter." In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63. DOI: [10.18653/v1/S19-2007](https://doi.org/10.18653/v1/S19-2007). URL: <https://www.aclweb.org/anthology/S19-2007>.
- Battistella, Edwin (1989). "Chinese reflexivization: a movement to INFL approach." In: *Linguistics* 27, pp. 987–1012.
- Bender, Emily and Batya Friedman (2018). "Data statements for natural language processing: Toward mitigating system bias and enabling better science." In: *TACL* 6, pp. 587–604.
- Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (2003). "A neural probabilistic language model." In: *Journal of machine learning research* 3.Feb, pp. 1137–1155.
- Bierema, Andrea, Anne-Marie Hoskinson, Rosa Moscarella, Alex Lyford, Kevin Haudek, John Merrill, and Mark Urban-Lurain (2020). "Quantifying cognitive bias in educational researchers." In: *International Journal of Research & Method in Education*, pp. 1–19.
- Bílý, Milan (1981). *Intrasentential pronominalization and functional sentence perspective*. Lund Slavonic Monographs.
- Bingel, Joachim, Victor Petrén Bach Hansen, Ana Valeria Gonzalez, Paweł Budzianowski, Isabelle Augenstein, and Anders Søgaard (2019). "Domain Transfer in Dialogue Systems without Turn-Level Supervision." In: *3rd Conversational AI Workshop at NeurIPS 2019*.
- Blitzer, John, Ryan McDonald, and Fernando Pereira (2006). "Domain adaptation with structural correspondence learning." In: *Proceedings of EMNLP*.
- 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*, pp. 4356–4364.
- Bordes, Antoine, Y-Lan Boureau, and Jason Weston (2016). "Learning end-to-end goal-oriented dialog." In: *arXiv preprint arXiv:1605.07683*.
- Bowling, Michael and Manuela Veloso (2001). "Rational and Convergent Learning in Stochastic Games." In: *IJCAI*.

- Bramhall, Steven, Hayley Horn, Michael Tieu, and Nibhrat Lohia (2020a). “QLIME-A Quadratic Local Interpretable Model-Agnostic Explanation Approach.” In: *SMU Data Science Review* 3.1, p. 4.
- Bramhall, Steven, Hayley Horn, Michael Tieu, and Nibhrat Lohia (2020b). “QLIME-A Quadratic Local Interpretable Model-Agnostic Explanation Approach.” In: *SMU Data Science Review* 3.
- Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer (June 1991). “Aligning Sentences in Parallel Corpora.” In: *29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, California, USA: Association for Computational Linguistics, pp. 169–176. DOI: [10.3115/981344.981366](https://doi.org/10.3115/981344.981366). URL: <https://www.aclweb.org/anthology/P91-1022>.
- Buçinca, Zana, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman (2020). “Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems.” In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 454–464.
- Buck, Gary (1991). “The testing of listening comprehension: an introspective study<sup>1</sup>.” In: *Language testing* 8.1, pp. 67–91.
- Budzianowski, Paweł and Ivan Vulić (2019). “Hello, It’s GPT-2—How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems.” In: *arXiv preprint arXiv:1907.05774*.
- 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.
- Cakmak, Maya, Crystal Chao, and Andrea L Thomaz (2010). “Designing interactions for robot active learners.” In: *IEEE Transactions on Autonomous Mental Development* 2.2, pp. 108–118.
- 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.
- Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom (2018). “e-snli: Natural language inference

- with natural language explanations." In: *Advances in Neural Information Processing Systems*, pp. 9539–9549.
- Caravona, Laura, Laura Macchi, Francesco Poli, Michela Vezzoli, Miriam AG Franchella, and Maria Bagassi (2019). "How to get rid of the belief bias: boosting analytical thinking via pragmatics." In: *Europe's Journal of Psychology* 15.3, pp. 595–613.
- Card, Stuart K (1983). *The psychology of human-computer interaction*. Crc Press.
- Carreras, Tomás and Joaquín Carreras (1939). *Historia de la filosofía española: filosofía cristiana de los siglos XIII al XV*. Vol. 2. Real academia de ciencias exactas, físicas y naturales.
- Chao, Guan-Lin and Ian Lane (2019). "BERT-DST: Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representations from Transformer." In: *Proc. Interspeech 2019*, pp. 1468–1472.
- Chen, Danqi, Adam Fisch, Jason Weston, and Antoine Bordes (2017a). "Reading Wikipedia to Answer Open-Domain Questions." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879.
- Chen, Hongshen, Xiaorui Liu, Dawei Yin, and Jiliang Tang (2017b). "A survey on dialogue systems: Recent advances and new frontiers." In: *ACM SIGKDD Explorations Newsletter* 19.2, pp. 25–35.
- Chen, Jianbo, Le Song, Martin J Wainwright, and Michael I Jordan (2018a). "Learning to explain: An information-theoretic perspective on model interpretation." In: *arXiv preprint arXiv:1802.07814*.
- Chen, Lu, Cheng Chang, Zhi Chen, Bowen Tan, Milica Gašić, and Kai Yu (2018b). "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.
- Cho, Kyunghyun, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." In:
- Chomsky, Noam (1991). *Lectures on Government and Binding*. Mouton de Gruyter.

- Chronopoulou, Alexandra, Christos Baziotis, and Alexandros Potamianos (2019). "An Embarrassingly Simple Approach for Transfer Learning from Pretrained Language Models." 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)*, pp. 2089–2095.
- Chu, Eric, Deb Roy, and Jacob Andreas (2020). "Are visual explanations useful? a case study in model-in-the-loop prediction." In: *arXiv preprint arXiv:2007.12248*.
- Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning (2019). "What does bert look at? an analysis of bert's attention." In: *arXiv preprint arXiv:1906.04341*.
- Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning (2020). "Electra: Pre-training text encoders as discriminators rather than generators." In: *arXiv preprint arXiv:2003.10555*.
- Cohen, David (1973). "Hindi'apnaa': A Problem in Reference Assignment." In: *Foundations of Language* 10.3, pp. 399–408.
- Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural language processing (almost) from scratch." In: *Journal of machine learning research* 12.ARTICLE, pp. 2493–2537.
- Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni (2018a). "What you can cram into a single vector: Probing sentence embeddings for linguistic properties." In: *arXiv preprint arXiv:1805.01070*.
- Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual language model pretraining." In: *Advances in Neural Information Processing Systems*, pp. 7057–7067.
- Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov (2018b). "XNLI: Evaluating Cross-lingual Sentence Representations." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485. DOI: [10.18653/v1/D18-1269](https://doi.org/10.18653/v1/D18-1269). URL: <https://www.aclweb.org/anthology/D18-1269>.
- Connell, Louise and Mark T Keane (2006). "A model of plausibility." In: *Cognitive Science* 30.1, pp. 95–120.

- Croskerry, Pat (2009). "A universal model of diagnostic reasoning." In: *Academic medicine* 84.8, pp. 1022–1028.
- Cuayahuitl, Heriberto, Simon Keizer, and Oliver Lemon (2015). "Strategic Dialogue Management via Deep Reinforcement Learning." In: *NIPS'15 Workshop on Deep Reinforcement Learning*.
- Dale, Robert (2019). "Law and word order: Nlp in legal tech." In: *Natural Language Engineering* 25.1, pp. 211–217.
- Das, Arun and Paul Rad (2020). "Opportunities and challenges in explainable artificial intelligence (xai): A survey." In: *arXiv preprint arXiv:2006.11371*.
- Das, Rajarshi, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum (2018). "Multi-step Retriever-Reader Interaction for Scalable Open-domain Question Answering." In: *International Conference on Learning Representations*.
- Daume III, Hal and Daniel Marcu (2006). "Domain adaptation for statistical classifiers." In: *Journal of Artificial Intelligence Research* 26, pp. 101–126.
- DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace (2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models." In: *ACL*.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019a). "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. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://www.aclweb.org/anthology/N19-1423>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019b). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *NAACL*.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019c). "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.
- Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning." In: *arXiv preprint arXiv:1702.08608*.
- Ebrahimi, Javid, Anyi Rao, Daniel Lowd, and Dejing Dou (2018). "HotFlip: White-Box Adversarial Examples for Text Classification." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36.
- 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. URL: <http://www.aclweb.org/anthology/W17-5526>.
- Elshawi, Radwa, Mouaz Al-Mallah, and Sherif Sakr (2019). "On the interpretability of machine learning-based model for predicting hypertension." In: *BMC Med Inform Decis Mak.* 19.
- Eriksson, Linn and Nicole Nguyen (2019). "The Swedish occupational register 2017." In: *Yrkesstrukturen i Sverige*.
- Ethayarajh, Kawin and Dan Jurafsky (2020). "Utility Is in the Eye of the User: A Critique of NLP Leaderboard Design." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853.
- Evans, J St BT, Julie L Barston, and Paul Pollard (1983). "On the conflict between logic and belief in syllogistic reasoning." In: *Memory & cognition* 11.3, pp. 295–306.
- Evans, J St and Handley BT. "SJ and Harper, C. 2001." In: *Necessity, possibility and belief: A study of syllogistic reasoning. Quarterly Journal of Experimental Psychology A* 54 (), pp. 935–958.
- Evans, Jonathan St BT (2003). "In two minds: dual-process accounts of reasoning." In: *Trends in cognitive sciences* 7.10, pp. 454–459.

- Evans, Jonathan St BT (2008). "Dual-processing accounts of reasoning, judgment, and social cognition." In: *Annu. Rev. Psychol.* 59, pp. 255–278.
- Evans, Jonathan St BT and Jodie Curtis-Holmes (2005). "Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning." In: *Thinking & Reasoning* 11.4, pp. 382–389.
- Evans, Jonathan St BT and Keith Ed Frankish (2009). *In two minds: Dual processes and beyond*. Oxford University Press.
- Fan, Angela, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli (2019). "Eli5: Long form question answering." In: *arXiv preprint arXiv:1907.09190*.
- Fan, Angela, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel (2020). "Generating Fact Checking Briefs." In: *arXiv preprint arXiv:2011.05448*.
- Fazio, Lisa K and Elizabeth J Marsh (2008). "Slowing presentation speed increases illusions of knowledge." In: *Psychonomic Bulletin & Review* 15.1, pp. 180–185.
- Fedus, William, Barret Zoph, and Noam Shazeer (2021). *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. arXiv: 2101.03961 [cs.LG].
- Feng, Shi and Jordan Boyd-Graber (2019a). "What can AI do for me? evaluating machine learning interpretations in cooperative play." In: *IUI*, pp. 229–239.
- Feng, Shi and Jordan Boyd-Graber (2019b). "What can AI do for me? evaluating machine learning interpretations in cooperative play." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 229–239.
- Fiebrink, Rebecca and Marco Gillies (2018). *Introduction to the special issue on human-centered machine learning*.
- Flowerdew, John, Michael H Long, Jack C Richards, et al. (1994). *Academic listening: Research perspectives*. Cambridge University Press.
- Forgues, Gabriel, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay (2014). "Bootstrapping dialog systems with word embeddings." In: *Nips, modern machine learning and natural language processing workshop*. Vol. 2.



- Fort, Karèn and Alain Couillault (2016). "Yes, we care! results of the ethics and natural language processing surveys." In: *international Language Resources and Evaluation Conference (LREC) 2016*.
- Furnham, Adrian and Hua Chu Boo (2011). "A literature review of the anchoring effect." In: *The journal of socio-economics* 40.1, pp. 35–42.
- Galley, Michel, Chris Brockett, Alessandro Sordani, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan (2015). "deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets." In: *arXiv preprint arXiv:1506.06863*.
- Gardent, Claire and Bonnie Webber (2001). "Towards the Use of Automated Reasoning in Discourse Disambiguation." In: *Journal of Logic, Language and Information* 10.
- 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.
- Goebel, Randy, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger (2018). "Explainable AI: the new 42?" In: *International cross-domain conference for machine learning and knowledge extraction*. Springer, pp. 295–303.
- Goel, Vinod and Oshin Vartanian (2011). "Negative emotions can attenuate the influence of beliefs on logical reasoning." In: *Cognition and Emotion* 25.1, pp. 121–131.
- Goh, Gary SW, Sebastian Lapuschkin, Leander Weber, Wojciech Samek, and Alexander Binder (2020). "Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution." In: *arXiv preprint arXiv:2004.10484*.
- Goldberg, Yoav (2019). "Assessing BERT's syntactic abilities." In: *arXiv preprint arXiv:1901.05287*.



- Goldstine, Herman H and Adele Goldstine (1946). "The electronic numerical integrator and computer (eniac)." In: *Mathematical Tables and Other Aids to Computation* 2.15, pp. 97–110.
- González, Ana Valeria, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard (2020). "Type B Reflexivization as an Unambiguous Testbed for Multilingual Multi-Task Gender Bias." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2637–2648.
- Gonzalez, Ana, Isabelle Augenstein, and Anders Søgaard (2018). "A strong baseline for question relevancy ranking." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4810–4815.
- Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2014). "Explaining and harnessing adversarial examples." In: *arXiv preprint arXiv:1412.6572*.
- Greenwald, Anthony G and Linda Hamilton Krieger (2006). "Implicit bias: Scientific foundations." In: *California law review* 94.4, pp. 945–967.
- Grosz, Barbara, Aravind Joshi, and Scott Weinstein (1995). "Centering: A Framework for Modeling the Local Coherence of Discourse." In: *Computational Linguistics* 21.
- Gruber, Sebastian (2019). "LIME and Sampling." In: *Limitations of ML Interpretability*. Ed. by Christoph Molnar. Chap. 13.
- Guillory, Andrew and Jeff A Bilmes (2011). "Simultaneous learning and covering with adversarial noise." In: *ICML*.
- Gulyaev, Pavel, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, and Mikhail Burtsev (2020). "Goal-oriented multi-task bert-based dialogue state tracker." In: *arXiv preprint arXiv:2002.02450*.
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger (2017). "On calibration of modern neural networks." In: *arXiv preprint arXiv:1706.04599*.
- Ham, Donghoon, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim (2020). "End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2." In: *ACL*.
- Hase, Peter and Mohit Bansal (2020). "Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" In: *arXiv preprint arXiv:2005.01831*.

- Heath, Chip and Amos Tversky (1991). "Preference and belief: Ambiguity and competence in choice under uncertainty." In: *Journal of risk and uncertainty* 4.1, pp. 5–28.
- Heine, Bernd (2005). "On reflexive forms in creoles." In: *Lingua* 115, pp. 201–257.
- Henderson, James, Oliver Lemon, and Kallirroi Georgila (2008). "Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets." In: *Computational Linguistics* 34.4, pp. 487–511.
- 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*. URL: <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/44018.pdf>.
- 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. URL: <http://www.aclweb.org/anthology/W14-4337>.
- Henderson, Matthew, Blaise Thomson, and Steve Young (2013). "Deep neural network approach for the dialog state tracking challenge." In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 467–471.
- 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. URL: <https://www.aclweb.org/anthology/W/W14/W14-4340.pdf>.
- 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 Society*, pp. 123–129.
- Hinze, Scott R, Daniel G Slaten, William S Horton, Ryan Jenkins, and David N Rapp (2014). "Pilgrims sailing the Titanic: Plausibility effects on memory for misinformation." In: *Memory & Cognition* 42.2, pp. 305–324.
- Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

- Hoffman, Robert R, Shane T Mueller, Gary Klein, and Jordan Litman (2018). "Metrics for explainable AI: Challenges and prospects." In: *arXiv preprint arXiv:1812.04608*.
- Holzinger, Andreas, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell (2017). "What do we need to build explainable AI systems for the medical domain?" In: *arXiv preprint arXiv:1712.09923*.
- Honselaar, Wim (1986). "REFLECTIONS ON THE RUSSIAN REFLEXIVE POSSESSIVE PRONOUN." In: *Studies in Slavic and General Linguistics* 8, pp. 235–248.
- Höök, Kristina (2000). "Steps to take before intelligent user interfaces become real." In: *Interacting with computers* 12.4, pp. 409–426.
- Hooker, Sara, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim (2019). "A Benchmark for Interpretability Methods in Deep Neural Networks." In: *Advances in Neural Information Processing Systems* 32. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 9737–9748. URL: <http://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks.pdf>.
- Horn, Robert E. (1995). "The Turing Test." In: *Parsing the Turing Test*. Ed. by Robert Epstein, Gary Roberts, and Grace Beber. Springer, pp. 73–88.
- Hornbæk, Kasper and Antti Oulasvirta (2017). "What is interaction?" In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 5040–5052.
- Horvitz, Eric (1999). "Principles of mixed-initiative user interfaces." In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166.
- Hosseini-Asl, Ehsan, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher (2020). "A simple language model for task-oriented dialogue." In: *arXiv preprint arXiv:2005.00796*.
- Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Atariyan, and Sylvain Gelly (2019). "Parameter-Efficient Transfer Learning for NLP." In: *ICML*.
- 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.

- Howard, Jeremy and Sebastian Ruder (2018). “Universal language model fine-tuning for text classification.” In: *arXiv preprint arXiv:1801.06146*.
- Huang, Yangsibo, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora (2020). “Texthide: Tackling data privacy in language understanding tasks.” In: *arXiv preprint arXiv:2010.06053*.
- Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl (2020). “Social Biases in NLP Models as Barriers for Persons with Disabilities.” In: *arXiv preprint arXiv:2005.00813*.
- Indyk, Piotr and Rajeev Motwani (1998). “Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality.” In: *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. STOC '98. Dallas, Texas, USA: ACM, pp. 604–613. ISBN: 0-89791-962-9. DOI: [10.1145/276698.276876](https://doi.org/10.1145/276698.276876). URL: <http://doi.acm.org/10.1145/276698.276876>.
- Jain, Sarthak and Byron C. Wallace (June 2019). “Attention is not Explanation.” 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. 3543–3556. DOI: [10.18653/v1/N19-1357](https://doi.org/10.18653/v1/N19-1357). URL: <https://www.aclweb.org/anthology/N19-1357>.
- Jia, Robin and Percy Liang (2017). “Adversarial Examples for Evaluating Reading Comprehension Systems.” In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031.
- Jiang, Jing and ChengXiang Zhai (2007). “Instance weighting for domain adaptation in NLP.” In: *Proceedings of ACL*.
- Jiao, Xiaoqi, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Lintan Li, Fang Wang, and Qun Liu (2019). “TinyBERT: Distilling BERT for Natural Language Understanding.” In:
- Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (2019). “SpanBERT: Improving Pre-training by Representing and Predicting Spans.” In: *arXiv preprint arXiv:1907.10529*.
- Joshi, Mandar, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer (2017). “Triviaqa: A large scale distantly supervised

- challenge dataset for reading comprehension." In: *arXiv preprint arXiv:1705.03551*.
- Ju, Toldova S, A Roytberg, AA Ladygina, MD Vasilyeva, IL Azerkovich, M Kurzukov, G Sim, DV Gorshkov, A Ivanova, A Nedoluzhko, et al. (2014). "RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian." In: *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue"*, pp. 681–694.
- Judd, Charles H (1908). "The relation of special training and general intelligence." In: *Educational review* 36, pp. 28–42.
- Kahneman, Daniel (2003). "A perspective on judgment and choice: mapping bounded rationality." In: *American psychologist* 58.9, p. 697.
- Kakade, Sham M (2002). "A natural policy gradient." In: *Advances in neural information processing systems*, pp. 1531–1538.
- Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih (2020). "Dense Passage Retrieval for Open-Domain Question Answering." In: *arXiv preprint arXiv:2004.04906*.
- Karthikeyan, Kaliyaperumal, Zihan Wang, Stephen Mayhew, and Dan Roth (2019). "Cross-lingual ability of multilingual bert: An empirical study." In: *International Conference on Learning Representations*.
- Kelly, Daniel and Erica Roedder (2008). "Racial cognition and the ethics of implicit bias." In: *Philosophy Compass* 3.3, pp. 522–540.
- Khosla, Aditya, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba (2012). "Undoing the damage of dataset bias." In: *European Conference on Computer Vision*. Springer, pp. 158–171.
- Kim, Seokhwan, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. (2019). "The eighth dialog system technology challenge." In: *arXiv preprint arXiv:1911.06394*.
- Kincaid, J. Peter, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom (1975). "Derivation of new readability formulas for navy enlisted personnel." In: *Research Branch Report*, pp. 8–75.

- Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization." In: *ICLR*. arXiv: 1412.6980. URL: <http://arxiv.org/abs/1412.6980>.
- Kiparsky, Paul (2001). "Disjoint reference and the typology of pronouns." In: *More than Words. A Festschrift for Dieter Wunderlich*.
- Klaczynski, Paul A, David H Gordon, and James Fauth (1997). "Goal-oriented critical reasoning and individual differences in critical reasoning biases." In: *Journal of Educational Psychology* 89.3, p. 470.
- Klauer, KC, J Musch, and B Naumer (2000). "On belief bias in syllogistic reasoning." In: *Psychological Review* 107.
- Koh, Pang Wei and Percy Liang (2017). "Understanding Black-box Predictions via Influence Functions." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, pp. 1885–1894. URL: <http://proceedings.mlr.press/v70/koh17a.html>.
- Kopper, Philipp (2019). "LIME and Neighborhood." In: *Limitations of ML Interpretability*. Ed. by Christoph Molnar. Chap. 13.
- Kos, Miriam, Theo Vosse, Daniëlle van den Brink, and Peter Hagoort (2010). "About Edible Restaurants: Conflicts between Syntax and Semantics as Revealed by ERPs." In: *Frontiers in Psychology* 1.222.
- Kotonya, Neema and Francesca Toni (2020). "Explainable automated fact-checking for public health claims." In: *arXiv preprint arXiv:2010.09926*.
- Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. (2019). "Natural questions: a benchmark for question answering research." In: *Transactions of the Association for Computational Linguistics* 7, pp. 453–466.
- Lage, Isaac, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez (2019). "An evaluation of the human-interpretability of explanation." In: *arXiv preprint arXiv:1902.00006*.
- Lage, Isaac, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez (2018). "Human-in-the-loop interpretabil-

- ity prior." In: *Advances in Neural Information Processing Systems*, pp. 10159–10168.
- Lamm, Matthew, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins (2020). *QED: A Framework and Dataset for Explanations in Question Answering*. arXiv: 2009.06354 [cs.CL].
- Larsen, M, H Holt, and MR Larsen (2016). "15:16 Et kønsopdelt arbejdsmarked: Udviklingstræk, konsekvenser og forklaringer." In: *SFI: Det Nationale Forskningscenter for Velfærd* 170.
- Laugel, Thibault, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki (2018). "Defining Locality for Surrogates in Post-Hoc Interpretability." In: *arXiv preprint arXiv:1806.07498*.
- Leahy, Wayne and John Sweller (2016). "Cognitive load theory and the effects of transient information on the modality effect." In: *Instructional Science* 44.1, pp. 107–123.
- Lee, John D and Katrina A See (2004). "Trust in automation: Designing for appropriate reliance." In: *Human factors* 46.1, pp. 50–80.
- Lee, Kenton, Ming-Wei Chang, and Kristina Toutanova (2019a). "Latent Retrieval for Weakly Supervised Open Domain Question Answering." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096.
- Lee, Kenton, Ming-Wei Chang, and Kristina Toutanova (July 2019b). "Latent Retrieval for Weakly Supervised Open Domain Question Answering." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6086–6096. DOI: 10.18653/v1/P19-1612. URL: <https://www.aclweb.org/anthology/P19-1612>.
- Lei, Tao, Regina Barzilay, and Tommi Jaakkola (2016). "Rationalizing Neural Predictions." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117.
- Lertvittayakumjorn, Piyawat and Francesca Toni (2019). "Human-grounded Evaluations of Explanation Methods for Text Classification." 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)*, pp. 5198–5208.



- Li, Jiwei, Thang Luong, and Dan Jurafsky (2015). "A Hierarchical Neural Autoencoder for Paragraphs and Documents." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1, pp. 1106–1115.
- Li, Jiwei, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky (2017a). "Adversarial Learning for Neural Dialogue Generation." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2157–2169.
- Li, Lihong, Jason D. Williams, and Suhrud Balakrishnan (2009). "Reinforcement learning for dialog management using least-squares Policy iteration and fast feature selection." In: *INTER-SPEECH*. URL: [https://www.isca-speech.org/archive/archive\\_papers/interspeech\\_2009/papers/i09\\_2475.pdf](https://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_2475.pdf).
- Li, Xiujun, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz (2017b). "End-to-End Task-Completion Neural Dialogue Systems." In: *IJCNLP*.
- Li, Yitong, Timothy Baldwin, and Trevor Cohn (2018). "Towards robust and privacy-preserving text representations." In: *arXiv preprint arXiv:1805.06093*.
- Liu, Bing, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck (2017). "End-to-end optimization of task-oriented dialogue model with deep reinforcement learning." In: *arXiv preprint arXiv:1711.10712*.
- Liu, Bing, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck (2018). "Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems." In: *Proceedings of NAACL-HLT*, pp. 2060–2069.
- Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau (2016). "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132.
- Liu, Hui, Qingyu Yin, and William Yang Wang (2019). "Towards Explainable NLP: A Generative Explanation Framework for Text Classification." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5570–5581.



- Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith (June 2019a). "Linguistic Knowledge and Transferability of Contextual Representations." 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. 1073–1094. DOI: [10.18653/v1/N19-1112](https://doi.org/10.18653/v1/N19-1112). URL: <https://www.aclweb.org/anthology/N19-1112>.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019b). "Roberta: A robustly optimized bert pretraining approach." In: *arXiv preprint arXiv:1907.11692*.
- Lødrup, Helge, Miriam Butt, and Tracy Holloway King (2011). "Norwegian possessive pronouns: Phrases, words or suffixes." In: *Proceedings of the LFG11 Conference*. Oslo: CSLI Publications, pp. 339–359.
- Lopez, Marc Moreno and Jugal Kalita (2017). "Deep Learning applied to NLP." In: *arXiv preprint arXiv:1703.03091*.
- Lowe, Ryan, Michael Noseworthy, Iulian V. Serban, Nicolas A.-Gontier, Yoshua Bengio, and Joelle Pineau (2017a). "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses." In: *ACL*.
- Lowe, Ryan, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau (2017b). "Towards an automatic turing test: Learning to evaluate dialogue responses." In: *arXiv preprint arXiv:1708.07149*.
- Lowe, Ryan, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau (2016). "On the evaluation of dialogue systems with next utterance classification." In: *arXiv preprint arXiv:1605.05414*.
- Luan, Yi, Yangfeng Ji, and Mari Ostendorf (2016). "LSTM based conversation models." In: *arXiv preprint arXiv:1603.09457*.
- Lund, Randall J (1991). "A comparison of second language listening and reading comprehension." In: *The modern language journal* 75.2, pp. 196–204.
- Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions." In: *Advances in neural information processing systems*, pp. 4765–4774.

- MacKenzie, I Scott (2012). "Human-computer interaction: An empirical research perspective." In:
- Manzini, Thomas, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black (2019). "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings." In: *arXiv preprint arXiv:1904.04047*.
- Markovits, Henry and Guilaine Nantel (1989). "The belief-bias effect in the production and evaluation of logical conclusions." In: *Memory & cognition* 17.1, pp. 11–17.
- Marsh, ElizabethJ and Sharda Umanath (2013). "Knowledge neglect: Failures to notice contradictions with stored knowledge." In: *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*.
- May, Chandler, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger (June 2019a). "On Measuring Social Biases in Sentence Encoders." 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. 622–628. DOI: [10.18653/v1/N19-1063](https://doi.org/10.18653/v1/N19-1063). URL: <https://www.aclweb.org/anthology/N19-1063>.
- May, Chandler, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger (2019b). "On measuring social biases in sentence encoders." In: *arXiv preprint arXiv:1903.10561*.
- McCorduck, Pamela (2004). *Machines who think*.
- McCoy, R Thomas, Ellie Pavlick, and Tal Linzen (2019). "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference." In: *ACL*.
- Mehri, Shikib, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi (2019). "Pretraining methods for dialog context representation learning." In: *arXiv preprint arXiv:1906.00414*.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems* 26, pp. 3111–3119.
- Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences." In: *Artificial Intelligence* 267, pp. 1–38.

- Min, Sewon, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer (2020). "AmbigQA: Answering Ambiguous Open-domain Questions." In: *arXiv preprint arXiv:2004.10645*.
- Mishra, Saumitra, Bob L Sturm, and Simon Dixon (2017). "Local Interpretable Model-Agnostic Explanations for Music Content Analysis." In: *ISMIR*, pp. 537–543.
- Mittelstadt, Brent, Chris Russell, and Sandra Wachter (2019). "Explaining explanations in AI." In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288.
- Mou, Lili, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin (2016). "How transferable are neural networks in nlp applications?" In: *arXiv preprint arXiv:1603.06111*.
- 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. URL: <https://aclanthology.info/papers/P15-2130/p15-2130>.
- Mrkšić, Nikola, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young (2016). "Neural belief tracker: Data-driven dialogue state tracking." In: *arXiv preprint arXiv:1606.03777*.
- 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. URL: <http://aclweb.org/anthology/P17-1163>.
- Mrkšić, Nikola and Ivan Vulić (2018). "Fully Statistical Neural Belief Tracking." In: *Proceedings of ACL*, pp. 108–113. URL: <http://aclweb.org/anthology/P18-2018>.
- 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. URL: <http://www.aclweb.org/anthology/Q17-1022>.
- Munro, Robert (2020). *Human-in-the-Loop Machine Learning*. Manning.

- Murdoch, W James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu (2019). "Interpretable machine learning: definitions, methods, and applications." In: *arXiv preprint arXiv:1901.04592*.
- Mynatt, Clifford R, Michael E Doherty, and Ryan D Tweney (1977). "Confirmation bias in a simulated research environment: An experimental study of scientific inference." In: *Quarterly Journal of Experimental Psychology* 29.1, pp. 85–95.
- Narang, Sharan, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan (2020a). *WT5?! Training Text-to-Text Models to Explain their Predictions*. arXiv: 2004.14546 [cs.CL].
- Narang, Sharan, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan (2020b). "WT5?! Training Text-to-Text Models to Explain their Predictions." In: *arXiv preprint arXiv:2004.14546*.
- Narayanan, Menaka, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez (2018). "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation." In: *arXiv preprint arXiv:1802.00682*.
- National Bureau of Statistics (2004). *Women and Men in China: Facts and figures*. URL: <http://www.stats.gov.cn/english/Statisticaldata/OtherData/200509/U020150722579392934100.pdf>.
- Newstead, Stephen E, Paul Pollard, Jonathan St BT Evans, and Julie L Allen (1992). "The source of belief bias effects in syllogistic reasoning." In: *Cognition* 45.3, pp. 257–284.
- Ng, Nathan, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov (2019). "Facebook FAIR's WMT19 News Translation Task Submission." In: *Proc. of WMT*.
- Nguyen, Dong (June 2018a). "Comparing Automatic and Human Evaluation of Local Explanations for Text Classification." 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. 1069–1078. DOI: 10.18653/v1/N18-1097. URL: <https://www.aclweb.org/anthology/N18-1097>.
- Nguyen, Dong (2018b). "Comparing automatic and human evaluation of local explanations for text classification." In: *Proceedings of the 2018 Conference of the North American Chapter of*

- the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1069–1078.
- Nivre, Joakim, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. (2017). “Universal Dependencies 2.1.” In:
- Norman, Donald A (1994). “How might people interact with agents.” In: *Communications of the ACM* 37.7, pp. 68–71.
- Nouri, Elnaz and Ehsan Hosseini-Asl (2018). “Toward Scalable Neural Dialogue State Tracking.” In: *NeurIPS 2018, 2nd Conversational AI workshop*.
- Olhede, Sofia C and Patrick J Wolfe (2018). “The growing ubiquity of algorithms in society: implications, impacts and innovations.” In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128, p. 20170364.
- Osada, Nobuko (2004). “Listening comprehension research: A brief review of the past thirty years.” In: *Dialogue* 3.1, pp. 53–66.
- PAIR, Google (2019). *The People + AI Guidebook*. <https://pair.withgoogle.com/guidebook/>. Accessed: 2021-1-28.
- Pan, Sinno Jialin and Qiang Yang (2010). “A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*.” In: 22 (10): 1345–1359.
- Papakyriakopoulos, Orestis, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco (2020). “Bias in word embeddings.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 446–457.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei jing Zhu (2002). “BLEU: a Method for Automatic Evaluation of Machine Translation.” In: pp. 311–318.
- Papini, Matteo, Matteo Pirota, and Marcello Restelli (2017). “Adaptive batch size for safe policy gradients.” In: *Advances in Neural Information Processing Systems*, pp. 3591–3600.
- Paranjape, Bhargavi, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer (2020). “An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction.” In: *arXiv preprint arXiv:2005.00652*.

- 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*. URL: <https://openreview.net/pdf?id=BJJsrmfCZ>.
- Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti (2014). "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk." In: *Behavior research methods* 46.4, pp. 1023–1031.
- Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation." In: *Proceedings of the 2014 conference on 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 (2018). "Deep contextualized word representations." In: *arXiv preprint arXiv:1802.05365*.
- Pietquin, Olivier and Helen Hastie (2013). "A survey on metrics for the evaluation of user simulations." In: *The knowledge engineering review* 28.1, p. 59.
- Platt, John et al. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." In: *Advances in large margin classifiers* 10.3, pp. 61–74.
- Poerner, Nina, Hinrich Schütze, and Benjamin Roth (July 2018). "Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 340–350. DOI: [10.18653/v1/P18-1032](https://doi.org/10.18653/v1/P18-1032). URL: <https://www.aclweb.org/anthology/P18-1032>.
- Popat, Kashyap, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum (2017). "Where the truth lies: Explaining the credibility of emerging claims on the web and social media." In: *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1003–1012.
- Popat, Kashyap, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum (2018). "Declare: Debunking fake news and false claims using evidence-aware deep learning." In: *arXiv preprint arXiv:1809.06416*.

- Poursabzi-Sangdeh, Forough, Daniel G Goldstein, Jake M Hoffman, Jennifer Wortman Vaughan, and Hanna Wallach (2018). "Manipulating and measuring model interpretability." In: *arXiv preprint arXiv:1802.07810*.
- Prates, Marcelo OR, Pedro H Avelar, and Luís C Lamb (2018). "Assessing gender bias in machine translation: a case study with Google Translate." In: *Neural Computing and Applications*, pp. 1–19.
- Pujari, Arun K, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar (2019). "Debiasing gender biased hindi words with word-embedding." In: *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pp. 450–456.
- Qudar, M and VIJAY Mago (2020). "A survey on language models." In:
- Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). *Improving language understanding by generative pre-training*.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language models are unsupervised multitask learners." In: *OpenAI blog* 1.8, p. 9.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2019). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." In: *arXiv preprint arXiv:1910.10683*.
- Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher (2019). "Explain Yourself! Leveraging Language Models for Commonsense Reasoning." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942.
- Rajpurkar, Pranav, Robin Jia, and Percy Liang (2018). "Know What You Don't Know: Unanswerable Questions for SQuAD." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789.
- Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392.



- 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. URL: <http://www.aclweb.org/anthology/P18-2069>.
- Rashid, Al Mamunur, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl (2002). "Getting to know you: learning new user preferences in recommender systems." In: *Proceedings of the 7th international conference on Intelligent user interfaces*, pp. 127–134.
- Rasmussen, Jens, Annelise Mark Pejtersen, and Len P Goodstein (1994). "Cognitive systems engineering." In:
- 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. URL: <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46399.pdf>.
- Rei, Marek and Anders Søgaard (June 2018). "Zero-Shot Sequence Labeling: Transferring Knowledge from Sentences to Tokens." 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. 293–302. DOI: 10.18653/v1/N18-1027. URL: <https://www.aclweb.org/anthology/N18-1027>.
- Ren, Hang, Weiqun Xu, Yan Zhang, and Yonghong Yan (2013). "Dialog state tracking using conditional random fields." In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 457–461.
- 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.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). "" Why should i trust you?" Explaining the predictions of any classifier." In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016b). "Model-agnostic interpretability of machine learning." In: *arXiv preprint arXiv:1606.05386*.



- Riedl, Mark O (2019). "Human-centered artificial intelligence and machine learning." In: *Human Behavior and Emerging Technologies* 1.1, pp. 33–36.
- Ritter, Alan, Colin Cherry, and William B Dolan (2011). "Data-driven response generation in social media." In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 583–593.
- Robnik-Šikonja, Marko and Marko Bohanec (2018). "Perturbation-based explanations of prediction models." In: *Human and machine learning*. Springer, pp. 159–175.
- Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2021). "A primer in bertology: What we know about how bert works." In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866.
- Rosenthal, Sara and Kathleen McKeown (2011). "Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 763–772.
- Rouse, William B (1986). "On the value of information in system design: A framework for understanding and aiding designers." In: *Information Processing & Management* 22.3, pp. 217–228.
- Rouse, William B and William B Rouse (1991). *Design for success: A human-centered approach to designing successful products and systems*. Vol. 2. Wiley-Interscience.
- Ruder, Sebastian (2019). "Neural transfer learning for natural language processing." PhD thesis. NUI Galway.
- Ruder, Sebastian, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf (2019). "Transfer Learning in Natural Language Processing." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18.
- 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. DOI: [10.18653/v1/N18-2002](https://doi.org/10.18653/v1/N18-2002). URL: <https://www.aclweb.org/anthology/N18-2002>.

- SCB, Statistics Sweden (2018). *Women and Men in Sweden: Facts and figures*. URL: [https://www.scb.se/contentassets/4550eaae793b46309da2aad79691e0201\\_2017b18\\_br\\_x10br1801eng.pdf](https://www.scb.se/contentassets/4550eaae793b46309da2aad79691e0201_2017b18_br_x10br1801eng.pdf).
- Samek, W., A. Binder, G. Montavon, S. Lapuschkin, and K. Müller (2017). "Evaluating the Visualization of What a Deep Neural Network Has Learned." In: *IEEE Transactions on Neural Networks and Learning Systems* 28.11, pp. 2660–2673.
- Schatzmann, Jost, Kallirroi Georgila, and Steve Young (2005). "Quantitative evaluation of user simulation techniques for spoken dialogue systems." In: *6th SIGdial Workshop on DIS-COURSE and DIALOGUE*.
- Schmidt, Philipp and Felix Biessmann (2019). "Quantifying Interpretability and Trust in Machine Learning Systems." In: *arXiv preprint arXiv:1901.08558*.
- Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization." In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.
- Sennrich, Rico, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams (Sept. 2017). "The University of Edinburgh's Neural MT Systems for WMT17." In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 389–399. DOI: [10.18653/v1/W17-4739](https://doi.org/10.18653/v1/W17-4739). URL: <https://www.aclweb.org/anthology/W17-4739>.
- Serban, Iulian V, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau (2015). "Hierarchical neural network generative models for movie dialogues." In: *arXiv preprint arXiv:1507.04808* 7.8, pp. 434–441.
- Serban, Iulian Vlad, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau (2016). "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models." In: *AAAI*. Vol. 16, pp. 3776–3784.
- Serban, Iulian Vlad, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio (2017). "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues." In: *AAAI*, pp. 3295–3301.
- Shah, Kirti and Bikas Sinha (1989). "4 Row-Column Designs." In: *Lecture Notes in Statistics* 54, pp. 66–84.

- Shang, Lifeng, Zhengdong Lu, and Hang Li (2015b). “Neural Responding Machine for Short-Text Conversation.” In: *ACL*.
- Shang, Lifeng, Zhengdong Lu, and Hang Li (2015a). “Neural Responding Machine for Short-Text Conversation.” In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1, pp. 1577–1586.
- Sharma, Shikhar, Layla El Asri, Hannes Schulz, and Jeremie Zumer (2017). “Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation.” In: *arXiv preprint arXiv:1706.09799*.
- Shen, Xiaoyu, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long (2017). “A Conditional Variational Framework for Dialog Generation.” In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2, pp. 504–509.
- Shibata, Masahiro, Tomomi Nishiguchi, and Yoichi Tomiura (2009). “Dialog system for open-ended conversation using web documents.” In: *Informatica* 33.3.
- Shneiderman, Ben (2020). “Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems.” In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10.4, pp. 1–31.
- Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). “Learning Important Features Through Propagating Activation Differences.” In: *ICML*.
- Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). “Deep inside convolutional networks: Visualising image classification models and saliency maps.” In: *arXiv preprint arXiv:1312.6034*.
- Singh, Jasdeep, Bryan McCann, Caiming Xiong, and Richard Socher (2019). “BERT is Not an Interlingua and the Bias of Tokenization.” In: *EMNLP*.
- 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. URL: <https://dl.acm.org/citation.cfm?id=1622410>.

- Slack, Dylan, Sorelle A Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy (2019). "Assessing the local interpretability of machine learning models." In: *arXiv preprint arXiv:1902.03501*.
- Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts (2013). "Recursive deep models for semantic compositionality over a sentiment treebank." In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.
- Song, Congzheng, Thomas Ristenpart, and Vitaly Shmatikov (2017). "Machine learning models that remember too much." In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 587–601.
- Song, Yiping, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang (2016). "Two are Better than One: An Ensemble of Retrieval- and Generation-Based Dialog Systems." In: *arXiv preprint arXiv:1610.07149*.
- Sordoni, Alessandro, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie (2015a). "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion." In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, pp. 553–562.
- Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan (2015b). "A neural network approach to context-sensitive generation of conversational responses." In: *arXiv preprint arXiv:1506.06714*.
- Sousa, David A (2002). *How the Brain Learns/Como Aprende el Cerebro*. Corwin Press.
- Stamatatos, Efstathios (2009). "A survey of modern authorship attribution methods." In: *Journal of the American Society for Information Science and Technology* 60.3, pp. 538–556.
- Stanovich, Keith E and Richard F West (2008). "On the relative independence of thinking biases and cognitive ability." In: *Journal of personality and social psychology* 94.4, p. 672.
- Steinbach, Markus (1998). "Middles in German." PhD thesis. Humboldt University.
- Stent, Amanda, Matthew Marge, and Mohit Singhai (2005). "Evaluating evaluation methods for generation in the pres-

- ence of variation." In: *international conference on intelligent text processing and computational linguistics*. Springer, pp. 341–351.
- Stent, Amanda, Rashmi Prasad, and Marilyn Walker (2004). "Trainable sentence planning for complex information presentations in spoken dialog systems." In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 79–86.
- Stoykova, Velislava (2012). "The inflectional morphology of Bulgarian possessive and reflexive-possessive pronouns in Universal Networking Language." In: *Procedia Technology* 1, pp. 400–406.
- Subramanian, Sandeep, Raymond Li, Jonathan Pilault, and Christopher Pal (2019). "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models." In: *arXiv preprint arXiv:1909.03186*.
- Sugiyama, Hiroaki, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami (2013). "Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures." In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 334–338.
- 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, Siqui Yao, Xueyang Wu, and Kai Yu (2016). "Hybrid Dialogue State Tracking for Real World Human-to-Human Dialogues." In: *INTERSPEECH*, pp. 2060–2064.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks." In: *arXiv preprint arXiv:1703.01365*.
- 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. URL: <https://www.aclweb.org/anthology/N06-1012>.
- Sutton, Richard S. and Andrew G. Barto (1998). *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press. ISBN: 0262193981. URL: <http://www.incompleteideas.net/book/first/the-book.html>.

- Swanson, Kyle, Lili Yu, and Tao Lei (2020). "Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport." In: *arXiv preprint arXiv:2005.13111*.
- Sweller, John (2011). "Cognitive load theory." In: *Psychology of learning and motivation*. Vol. 55. Elsevier, pp. 37–76.
- Sysoev, A, I Andrianov, and A Khadzhiiskaia (2017). "Coreference resolution in russian: State-of-the-art approaches application and evolvment." In: *Proceedings of International Conference Dialogue-2017*, pp. 327–338.
- Tan, Yi Chern and L Elisa Celis (2019). "Assessing social and intersectional biases in contextualized word representations." In: *Advances in Neural Information Processing Systems*, pp. 13230–13241.
- Taylor, Wilson L (1953). "'Cloze procedure': A new tool for measuring readability." In: *Journalism Bulletin* 30.4, pp. 415–433.
- Thompson, Irene and Joan Rubin (1996). "Can strategy instruction improve listening comprehension?" In: *Foreign Language Annals* 29.3, pp. 331–342.
- Thorndyke, EL and RS Woodworth (1901). "The influence of improvement in one mental function upon the efficiency of other functions." In: *Psychological Review* 8, pp. 247–261.
- Tian, Ran, Shashi Narayan, Thibault Sellam, and Ankur Parikh (2020). "Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation." In: *Arxiv*, p. 1910.08684.
- Trippas, Dries and Simon J Handley (2018). "The parallel processing model of belief bias: Review and extensions." In:
- Tukey, John W (1949). "Comparing individual means in the analysis of variance." In: *Biometrics*, pp. 99–114.
- Turing, Alan (1950). "Computing machinery and intelligence." In: *Mind* 59.236, pp. 433–433.
- Tversky, Amos and Daniel Kahneman (1974). "Judgment under uncertainty: Heuristics and biases." In: *science* 185.4157, pp. 1124–1131.
- Vasconcelos, Nuno and Andrew Lippman (1999). "Learning from user feedback in image retrieval systems." In: *Advances in neural information processing systems* 12, pp. 977–986.

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need." In: *Advances in neural information processing systems*, pp. 5998–6008.
- Vinyals, Oriol and Quoc V Le (2015). "A Neural Conversational Model." In:
- Vlassopoulos, Georgios (2019). *Decision Boundary Approximation: A new method for locally explaining predictions of complex classification models*. Tech. rep. University of Leiden.
- Wall, Emily, Leslie M Blaha, Lyndsey Franklin, and Alex Endert (2017). "Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics." In: *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, pp. 104–115.
- Wallace, Eric, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh (2019). "Universal Adversarial Triggers for Attacking and Analyzing NLP." 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)*, pp. 2153–2162.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman (2018). "Glue: A multi-task benchmark and analysis platform for natural language understanding." In: *arXiv preprint arXiv:1804.07461*.
- Wang, Danding, Qian Yang, Ashraf Abdul, and Brian Y Lim (2019). "Designing theory-driven user-centric explainable AI." In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15.
- 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. URL: <https://www.aclweb.org/anthology/W/W13/W13-4067.pdf>.
- Weaver, Lex and Nigel Tao (2001). "The optimal reward baseline for gradient-based reinforcement learning." In: *UAI*.
- Webb, Geoffrey I, Michael J Pazzani, and Daniel Billsus (2001). "Machine learning for user modeling." In: *User modeling and user-adapted interaction* 11.1-2, pp. 19–29.
- Webster, Kellie, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford (Aug. 2019). "Gendered Ambiguous Pronoun



- (GAP) Shared Task at the Gender Bias in NLP Workshop 2019." In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, pp. 1–7. DOI: [10.18653/v1/W19-3801](https://doi.org/10.18653/v1/W19-3801). URL: <https://www.aclweb.org/anthology/W19-3801>.
- Weitz, Katharina, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André (2019). "' Do you trust me?' Increasing user-trust by integrating virtual agents in explainable AI interaction design." In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 7–9.
- Weizenbaum, Joseph (1966). "ELIZA—a computer program for the study of natural language communication between man and machine." In: *Communications of the ACM* 9.1, pp. 36–45.
- Wen, Haoyang, Yijia Liu, Wanxiang Che, Libo Qin, and Ting Liu (2018). "Sequence-to-Sequence Learning for Task-oriented Dialogue with Dialogue State Representation." In: *COLING*.
- Wen, TH, D Vandyke, N Mrkšić, M Gašić, LM Rojas-Barahona, PH Su, S Ultes, and S Young (2017). "A network-based end-to-end trainable task-oriented dialogue system." In: *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*. Vol. 1, pp. 438–449.
- Weston, Jason, Emily Dinan, and Alexander Miller (2018). "Retrieve and Refine: Improved Sequence Generation Models For Dialogue." In: *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pp. 87–92.
- White, Michael and Ted Caldwell (1998). "EXEMPLARS: A practical, extensible framework for dynamic text generation." In: *Natural Language Generation*.
- Wickramasinghe, Chathurika S, Daniel L Marino, Javier Grandio, and Milos Manic (2020). "Trustworthy AI Development Guidelines for Human System Interaction." In: *2020 13th International Conference on Human System Interaction (HSI)*. IEEE, pp. 130–136.
- Wiegrefe, Sarah and Yuval Pinter (Nov. 2019). "Attention is not not Explanation." 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 Com-



- putational Linguistics, pp. 11–20. DOI: [10.18653/v1/D19-1002](https://doi.org/10.18653/v1/D19-1002). URL: <https://www.aclweb.org/anthology/D19-1002>.
- Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu (2015). “Towards universal paraphrastic sentence embeddings.” In: *arXiv preprint arXiv:1511.08198*.
- Williams, Adina, Andrew Drozdov\*, and Samuel R Bowman (2018). “Do latent tree learning models identify meaningful structure in sentences?” In: *Transactions of the Association for Computational Linguistics* 6, pp. 253–267.
- Williams, Jason D, Kavosh Asadi, and Geoffrey Zweig (2017). “Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning.” In: *arXiv preprint arXiv:1702.03274*.
- 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. ISSN: 0885-2308. DOI: <https://doi.org/10.1016/j.csl.2006.06.008>. URL: <http://www.sciencedirect.com/science/article/pii/S0885230806000283>.
- Williams, Jason D and Geoffrey Zweig (2016). “End-to-end lstm-based dialog control optimized with supervised and reinforcement learning.” In: *arXiv preprint arXiv:1606.01269*.
- 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. URL: <http://aclweb.org/anthology/W13-4068>.
- Williams, Jason, Antoine Raux, and Matthew Henderson (2016). “The dialog state tracking challenge series: A review.” In: *Dialogue & Discourse* 7.3, pp. 4–33.
- 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. URL: <http://www.aclweb.org/anthology/W13-4065>.
- Williams, Ronald J and Jing Peng (1991). “Function optimization using connectionist reinforcement learning algorithms.” In: *Connection Science* 3.3, pp. 241–268. URL: <https://www.tandfonline.com/doi/abs/10.1080/09540099108946587>.
- Wu, Chien-Sheng, Steven Hoi, Richard Socher, and Caiming Xiong (2020). “Tod-bert: Pre-trained natural language un-

- derstanding for task-oriented dialogues." In: *arXiv preprint arXiv:2004.06871*.
- Xu, Wei (2019). "Toward human-centered AI: a perspective from human-computer interaction." In: *interactions* 26.4, pp. 42–46.
- Yang, Qian (2017). "The role of design in creating machine-learning-enhanced user experience." In: *2017 AAAI spring symposium series*.
- Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning (2018). "Hotpotqa: A dataset for diverse, explainable multi-hop question answering." In: *arXiv preprint arXiv:1809.09600*.
- Yeom, Samuel, Irene Giacomelli, Matt Fredrikson, and Somesh Jha (2018). "Privacy risk in machine learning: Analyzing the connection to overfitting." In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, pp. 268–282.
- Yoshino, Koichiro, Takuya Hiraoka, Graham Neubig, and Satoshi Nakamura (2016). "Dialogue state tracking using long short term memory neural networks." In: *Proceedings of Seventh International Workshop on Spoken Dialog Systems*, pp. 1–8.
- Zhang, Hainan, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng (2018). "Reinforcing Coherence for Sequence to Sequence Model in Dialogue Generation." In: *IJCAI*.
- Zhang, Yizhe, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan (2019). "Dialogpt: Large-scale generative pre-training for conversational response generation." In: *arXiv preprint arXiv:1911.00536*.
- Zhang, Yu, Mark Lewis, Michael Pellon, and Phillip Coleman (2007). "A Preliminary Research on Modeling Cognitive Agents for Social Environments in Multi-Agent Systems." In: *AAAI Fall Symposium*.
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang (2019). "Gender bias in contextualized word embeddings." In: *arXiv preprint arXiv:1904.03310*.
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (June 2018). "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. DOI: [10.18653/v1/N18-2003](https://doi.org/10.18653/v1/N18-2003). URL: <https://www.aclweb.org/anthology/N18-2003>.

Zhao, Tiancheng and Maxine Eskenazi (Jan. 2016). “Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning.” In: pp. 1–10. DOI: [10.18653/v1/W16-3601](https://doi.org/10.18653/v1/W16-3601).

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. URL: <http://www.aclweb.org/anthology/P18-1135>.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.” In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.