UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

User-centric Natural Language Processing

Permalink

https://escholarship.org/uc/item/7nm5p7gd

Author

Majumder, Bodhisattwa Prasad

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org

Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

User-centric Natural Language Processing

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Bodhisattwa Prasad Majumder

Committee in charge:

Professor Julian McAuley, Chair
Professor Taylor Berg-Kirkpatrick
Professor Garrison W. Cottrell
Professor Arya Mazumder
Professor Lawrance Saul
Professor Sameer Singh

2023

Copyright

Bodhisattwa Prasad Majumder, 2023

All rights reserved.

The Dissertation of Bodhisattwa Prasad Majumder is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To the community that fosters innovations in computer science and social science

TABLE OF CONTENTS

Dissertation Approval Page iii

Dedication iv

Table of Contents v

List of Figures ix

List of Tables xii

Acknowledgements xv

Vita xviii

Abstract of the Dissertation xix

Chapter 1 Introduction 1
- 1.1 User-centric AI 1
- 1.2 Knowledge Grounding 3
- 1.3 Generating Explanations 5
- 1.4 Engaging in Interactions 6
- 1.5 Summary 8

Chapter 2 Background 10
- 2.1 Commonsense Knowledge 12
  - 2.1.1 Definition of Commonsense 12
  - 2.1.2 Resources 13
- 2.2 Language Generation Tasks Requiring Background Knowledge 15
- 2.3 Sequence Transduction Models 16
  - 2.3.1 Pre-training 16
- 2.4 Large Language Models 17
- 2.5 Conclusion 17

Chapter 3 COMPAC: Persona-grounded Dialog with Commonsense Expansions 18
- 3.1 Introduction 19
- 3.2 Persona Grounded Dialog 21
- 3.3 Persona Expansion 22
  - 3.3.1 COMET 22
  - 3.3.2 Paraphrasing 23
- 3.4 Common sense and Persona Aligned Chatbot (COMPAC) 24
  - 3.4.1 Persona Choice Prior 25
  - 3.4.2 Generator Network 26
  - 3.4.3 Learning and Inference 27

3.5 Experiments 29
3.5.1 Baselines 29
3.5.2 Comparison of Dialog Quality 30
3.5.3 Human Evaluation for Dialog Generation 32
3.5.4 Fine-grained Persona Grounding 32
3.5.5 Controllable Generation 34
3.5.6 Qualitative Analysis 36
3.6 Conclusion 36

Chapter 4 INTERVIEW: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding 38
4.1 Introduction 38
4.2 INTERVIEW: A Media Dialog Corpus 41
4.2.1 Comparison with Other Datasets 41
4.3 INTERVIEW Discourse Analysis 42
4.3.1 Dialog Patterns 42
4.3.2 Question Types as Dialog Acts 43
4.3.3 Knowledge Grounding 45
4.4 Modeling Media Dialog 48
4.4.1 Knowledge Grounded Generator 49
4.4.2 Predicting Look-ahead Dialog Patterns 50
4.4.3 Predicting Question types 50
4.5 Experiments 50
4.5.1 Effect of Knowledge Grounding 52
4.5.2 Effect of Auxiliary Tasks 53
4.5.3 Human Evaluation 54
4.6 Conclusion 55

Chapter 5 PABST: Unsupervised Enrichment of Persona-grounded Dialog with Background Stories 56
5.1 Introduction 56
5.2 Unsupervised Persona Enrichment with Background Stories 58
5.2.1 Retrieving Relevant Stories 59
5.2.2 Gradient-based Inference 59
5.3 Experiments 61
5.3.1 Automatic Evaluation 63
5.3.2 Human Evaluation 63
5.4 Conclusion 65

Chapter 6 POKI: Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection 66
6.1 Introduction 67
6.2 Post-hoc Knowledge for Dialog 69

6.2.1 Parametric knowledge sources 70
6.2.2 Non-parametric knowledge sources 70
6.3 Unsupervised Knowledge Injection in Generated Dialog 71
6.3.1 Relevance-Redundancy Tradeoff for Knowledge Selection 71
6.3.2 Gradient-based Constrained Decoding for Knowledge Injection 73
6.3.3 Unsupervised Ranking of Candidate Final Responses 74
6.4 Experimental Setup 75
6.4.1 Scenarios and Datasets 75
6.4.2 Baselines and Ablations 75
6.5 Results and Discussion 77
6.5.1 Automatic Evaluation 77
6.5.2 Human Evaluation 79
6.5.3 User Study for Effectiveness of Knowledge Injection 80
6.5.4 Discussion 82
6.6 Conclusion 83

Chapter 7 Improving Clarification Question Generation using Global Knowledge 84
7.1 Introduction 84
7.2 Problem Setup and Scenarios 86
7.3 Approach 87
7.3.1 Identifying Missing Information 88
7.3.2 Generating Useful Questions 90
7.4 Experiments 91
7.4.1 Datasets 91
7.4.2 Baselines and Ablations 93
7.4.3 Evaluation Metrics 93
7.4.4 Experimental Results 95
7.5 Analysis 100
7.6 Conclusion 102

Chapter 8 REXC: Knowledge Grounded Self-rationalization via Extractive and Natural Language Explanations 103
8.1 Introduction 103
8.2 REXC 106
8.2.1 Extractive Rationales via Binary Latent Variables 106
8.2.2 Knowledge about an Extractive Rationale 106
8.2.3 Knowledge Selection 107
8.2.4 NLE Generation and Task Prediction 108
8.2.5 Training 109
8.3 Experiments 110
8.4 Results 112
8.4.1 Evaluating the Quality of the Explanations 112
8.4.2 Task Performance 115
8.4.3 Zero-shot NLEs 115

8.4.4 Generative vs. Retrieval-based Knowledge Module 116
8.5 Evaluating Faithfulness 116
8.5.1 Faithfulness of the NLEs 116
8.5.2 Faithfulness of the ERs and Knowledge Snippets 119
8.6 Summary and Outlook 119

Chapter 9 INTERFAIR: Debiasing with Natural Language Feedback for Fair Interpretable Predictions 121
9.1 Introduction and Background 121
9.2 Controlling Bias Exposure via Rationales 122
9.2.1 Rationales and Energies 122
9.2.2 Training with Energy-based Constraint 123
9.3 Natural Language (NL) Feedback with INTERFAIR 124
9.3.1 Parsing NL feedback 125
9.3.2 Modifying Task Rationales 126
9.4 Experiments and Results 126

Chapter 10 Related Work 130
10.1 Related Work for COMPAC 130
10.2 Related work for INTERVIEW 131
10.3 Related Work for PABST 132
10.4 Related Work for POKI 133
10.5 Related Work for Clarification Question Generation 134
10.6 Related Work for REXC 135

Chapter 11 Conclusion and Future Outlook 138
11.1 Future Work for Knowledge Grounding 138
11.2 Future Work for Generating Explanations 139
11.3 Future Work for Interactive Systems 139

Bibliography 141

LIST OF FIGURES

Figure 1.1. Interactive explainability as the foundation of user-centric AI systems 3

Figure 1.2. Post-hoc Knowledge Injection (POKI) in a dialog model that was trained on limited knowledge 4

Figure 1.3. REXC bridges extractive rationales and abstractive explanations using background knowledge to improve plausibility, accuracy, and attribution. 6

Figure 1.4. INTERFAIR, a new interactive paradigm of controllable debiasing. Users can update the model's belief about sensitive information to mitigate bias while maintaining the task performance. 7

Figure 2.1. State-of-the-art generation models struggle to generate a factual text [IV et al., 2019], respond a user's query [Welleck et al., 2019], or generate a story with plausible story-line [Guan et al., 2020]. 11

Figure 3.1. State-of-the-art models struggle to respond a user's query, where generating an engaging response depends on commonsense reasoning. 19

Figure 3.2. Expansions of an original persona via (a) human rewrite, (b) paraphrase, and (c) COMET. 20

Figure 3.3. COMPAC samples a persona sentence from the prior and generates the response conditioned on the dialog context and sampled persona. The inference network is used only during training. 26

Figure 4.1. Our dialog model incorporates grounding documents alongside dialog history. We also leverage the dialog patterns and interrogative positioning by the host via auxiliary losses. 39

Figure 4.2. Example conversation from INTERVIEW with annotated discourse analysis. Text highlighted in blue indicates the question of interest, uttered by the host. The dialog triplet is marked in red. 40

Figure 4.3. (a) Bar plot depicts test perplexity for linking algorithms: None (no grounding), TF-IDF, and PL/PL3 which indicate probabilistic linking with reassignment at every 1/3 epochs, respectively. Plotting validation perplexity by epoch shows that PL3 converges faster to the optimum (b). 46

Figure 4.4. Knowledge grounded generator model with two discourse-specific auxiliary tasks for media dialog 48

Figure 5.1. We enrich agent personas with ‘background stories’ from an existing corpus. We propose a gradient-based technique which encourages the generated response to be fluent with the dialog history, minimally different from the retrieved story, and consistent with the persona. 57

Figure 6.1. Augmenting initial response from an existing dialog model with relevant external knowledge leads to more *engaging* and *informative* responses improving the success in achieving the conversational goal (here, finding a fun activity). 67

Figure 6.2. Pipeline of POKI: It first retrieves **post-hoc knowledge**, then the most relevant and diverse knowledge snippets are **selected** from the retrieved set. Each selected snippet is individually injected through a **constrained decoding** to generate a candidate response. The final response is selected via a **ranking** step. 69

Figure 6.3. POKI converses with a user who is looking for some restaurant options (left column). In each turn, a knowledge snippet (right column) is injected into an initial response (middle column). 78

Figure 7.1. Test-time behaviour of our model in a Community-QA setup. 1. We obtain a local schema, 2. the global schema, and 3. estimate the *missing* schema for the clarification question. 4. A BART model generates a question and 5. a PPLM model further tunes it to be more *useful*. 85

Figure 7.2. Dependency tree and paths showing how we obtain schema triples for a sentence: *“Will this bag hold a gaming laptop and an iPad?”* (from Figure 7.1). 89

Figure 7.3. Results of a pairwise comparison (on usefulness criteria) between our model and baseline generated question on (a) 300 randomly sampled product descriptions from the Amazon test set, (b) 150 randomly sampled dialogs from the Ubuntu test set as judged by humans. 97

Figure 7.4. Average BLEU score difference between classes having longer (> 200 (median) words) and shorter descriptions; larger (> 200 (median) key-phrases) and shorter global schema for the Amazon dataset. Lower differences indicate more invariance toward the available information. 101

Figure 8.1. Illustrative examples for REXC on (a) natural language and (b) vision-language tasks. 104

Figure 8.2. Architecture of REXC. The knowledge module is frozen, while the rest of the modules are trained jointly with the signals from the NLEs and outputs. Deliverables from REXC are in blue. 107

Figure 8.3. Examples of NLEs and ERs generated from RExC along with selected knowledge snippets vs. those from the previous SOTA for the correct predictions for COSe and VCR. 113

Figure 8.4. Feature importance agreement. Left: Accuracy when important features (solid) vs when random features (dotted) are occluded. Right: Simulatabilities important features (solid) vs when random features (dotted) are occluded. 117

Figure 8.5. Robustness equivalence analysis when noise is added to the (a) input and (b) selected knowledge snippets. In each pair, the left chart shows % of stable (unflipped) labels (solid), and accuracy (dashed). The right chart in a pair depicts the simulatability of NLEs. 118

Figure 9.1. **Interactive setup for INTERFAIR** where user provides feedback to update bias rationales that in turn updates task rationales and subsequent task prediction. . . . 124

LIST OF TABLES

Table 3.1. Summary of notation used in this chapter 25

Table 3.2. Dialog quality metrics on the PERSONA-CHAT test set. PPL=Perplexity, D-1/2=% of distinct uni- and bi-grams. 29

Table 3.3. Pairwise comparison (in %) between responses generated by COMPAC vs. other baselines (og: original, par: paraphrase) as well as the Gold response. The highest entries are **bolded**. 30

Table 3.4. Sample generations by different models. 31

Table 3.5. Assessment of persona grounding with and without inference network using the DNLI entailment set. Human evaluation (eval.) was conducted to measure the relevance when an expanded persona is chosen–all entries are statistically significant. 33

Table 3.6. Conditional generation performance on the PERSONA-CHAT test set to show the similarity between generated responses and grounding persona sentences. We omit GPT2-based models since they do not select a particular persona sentence for grounding. 34

Table 3.7. Controlled generation with edited persona 35

Table 3.8. Examples showing correct and incorrect persona choices in various dialog contexts by COMPAC model. COMPAC is capable of choosing a correct persona sentence (original or expanded) but sometimes the prior network fails to sample an appropriate one (third case). 36

Table 4.1. Comparative media dialog dataset statistics. *RadioTalk does not contain full conversations 41

Table 4.2. F1 Performance of question-type classifier models on the test set. 45

Table 4.3. Performance on auxiliary tasks: Dialog Pattern prediction and Question Type prediction 49

Table 4.4. Metrics on generated interviewer responses on test set. NPO/NEO = Noun-phrase/Named entity overlap with context (C) and gold (G); QR = Question rate. NIDF is a measure of specificity [See et al., 2019]. QR, NPO, NEO are measured in percentages. 51

Table 4.5. Pairwise comparison (in %) between responses generated by our best model (including both discourse analysis auxiliary tasks) vs. responses generated by other baselines as well as the Gold response. The highest entries are **bolded**. 51

Table 4.6. Sample generated response on NFL's new rule. When we add discourse specific losses, the models generate questions that bears more coherence to the context as well as ask clarifying questions. 52

Table 5.1. Diversity metrics on the PersonaChat test set. D-1/2 is the % of distinct uni- and bi-grams. ENTR is the geometric mean of n-gram entropy. Grad. Inf. is the unsupervised gradient-based decoding as opposed to Nucleus sampling [Holtzman et al., 2020]. 61

Table 5.2. Pairwise comparison (% win/loss cases, ties not reported) between responses from PABST and from other baselines as well as gold. All differences are significant with $p < 0.05$. Cohen's Kappa [Cohen, 1960] for sensibility and engagement were 0.79 and 0.82 respectively. 62

Table 5.3. Generations from different models. 62

Table 6.1. Automatic metrics on the test set of MultiWoZ. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$). 76

Table 6.2. Automatic metrics on the test set of Wizard-of-Wikipedia. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$). 76

Table 6.3. Pairwise comparison (% win/loss cases, tie not reported) between responses from POKI and from other baselines as well as ground truth. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$). κ denotes Cohen's Kappa [Cohen, 1960] between a pair of annotators. 77

Table 6.4. Real-time user study with average # of turns for successful goal completion, % of time the goal was achieved, % of success cases users were helped by an *additional* knowledge (Know) that was not explicitly asked to reach their goal, and if users would like to use the system in future. 81

Table 6.5. Evaluation for the quality of the knowledge snippets for random and DPP-based selection. 82

Table 6.6. Mean and std. error of clock-time taken per token 82

Table 7.1. Product description from amazon.com paired with a clarification question generated by our model. 86

Table 7.2. Number of data instances in the train, validation and test splits of Amazon and Ubuntu datasets (Both datasets are in English.) 92

Table 7.3. Automatic metric results on the full test set of Amazon. The difference between bold and non-bold numbers is statistically significant with $p < 0.001$. 94

Table 7.4. Automatic metric results the full test set of Ubuntu. The difference between bold and non-bold numbers is statistically significant with $p < 0.001$...... 94

Table 7.5. Human judgment results (0-1) on 300 randomly sampled descriptions from the Amazon test set 96

Table 7.6. Human judgment results (0-1) on 150 randomly sampled dialog contexts from Ubuntu test set 97

Table 7.7. Model generations for an example product from Amazon and an example dialog context from Ubuntu.. 99

Table 7.8. Missing information overlap (in %) between missing schema and output generations 99

Table 8.1. Our tasks: three Natural Language and two Vision-Language. 110

Table 8.2. **Task performance (Acc.) and NLE quality for the (a) NL and (b) VL tasks.** NLE Automatic metrics: METEOR, BERTScore, BLEURT, and NLE human evaluation metrics: e-ViL score, Yes/No %s. **Bold** are the best numbers ($p < 0.001$). Underline indicates best task performance for models with any explanations. 111

Table 8.3. **ER quality.** Comparison of previous SOTA models [DeYoung et al., 2020a] for rationale extraction vs. REXC for ER quality. Best numbers are in **bold**......... 113

Table 8.4. **Comprehensiveness (Comp.) and Sufficiency (Suff.)** metrics for ERs and selected knowledge snippets generated by REXC vs. random ERs and knowledge snippets 120

Table 9.1. **NL feedback parser.** Parse examples for IID and compositional (Comp) splits; parsing accuracy on IID, compositional splits and overall test set. GPT-Neo has 2.7B parameters, GPT-J has 6B parameters............................. 127

Table 9.2. **Evaluation** for task accuracy (Acc. ↑), bias (F1 ↓), and faithfulness for task rationales: Comprehensiveness (Compre. ↑) and Sufficiency (Suff. ↓) 128

ACKNOWLEDGEMENTS

I am incredibly appreciative of the chance I was given to pursue my academic objectives. I would not have been able to accomplish this feat without the steadfast support and assistance of numerous people.

I would like to express my deepest gratitude to Julian McAuley, my advisor, for his unwavering encouragement, guidance, and multiple chats over boba throughout my dissertation journey. On multiple occasions, his valuable advice has helped me become an independent researcher and alleviated my confidence in approaching other professional opportunities beyond thesis research.

My heartfelt appreciation goes out to Taylor Berg-Kirkpatrick, Gary Cottrell, Arya Mazumder, Lawrance Saul, and Sameer Singh, the other members of my dissertation committee, for providing me with indispensable suggestions and encouragement that significantly contributed to shaping my research direction.

I was fortunate to be supported by the generous funds from NSF Award #1750063, and an unrestricted gift from MeetElise. My research was recognized by the Qualcomm Innovation Fellowship, Adobe Research Fellowship, and Friends of the International Center Fellowship, UC San Diego—to whom I am deeply thankful.

Following, I would not have been able to give my research shape without the contribution of the many mentors I was lucky to have during my internships: Marc Najork (Google AI), Sandeep Tata (Google AI), Navneet Potti (Google AI), James Wendt (Google AI), Qi Zhao (Google AI), Sudha Rao (Microsoft Research), Michel Galley (Microsoft Research), Y-Lan Boureau (Meta AI), Asli Celikyilmaz (Meta AI), Peter Clark (AI2), Bhavana Dalvi (AI2), Oyvind Tafjord (AI2), and Niket Tandon (AI2). I must express my ultimate gratitude to Harshit Surana, Animesh Mukherjee, and Pawan Goyal, whose support was instrumental in me taking the maiden step of my Ph.D. career.

My Ph.D. journey would have been much wary had it not been for my amazing collaborators: Harsh Jhamtani, Oana-Maria Camburu, Thomas Lukasiewicz, Vered Shwartz, Shuyang

Li, Jianmo Ni, Henry Mao, Zexue He, Zhouhang Xie, Sophia Sun, and all other labmates from my lab (The Julian McAuley Lab)—help and guidance from all of them have been extremely invaluable all throughout.

Eventually, it would have been very difficult to sustain without the continual support of my friends-turned-family: Sumit Da, Sujoy, Aranya Da, Digbalay Da, Rajrup, Agnimitra, Ragini, Brato Da, Shouvik Da, Varun, Rob, Sanandita, Veltu Da, Subho, Diya, Buria Di, Ronty, and Mama-Mami from San Diego. And a special thanks goes to Sanchaita, a constant throughout the good times and the bad.

Finally, without the sacrifices, guidance, and blessings of my parents, I would not have had the opportunity to achieve my goals and wishes, and for that, I am forever grateful.

I am also grateful to my co-authors who kindly approved the following publications and material to be included in my dissertation:

Chapter 3, in part, is a reprint of the material as it appears in “Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions” by Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley, which was published in *Empirical Methods in Natural Language Processing*, 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in “Interview: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding” by Bodhisattwa Prasad Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley, which was published in *Empirical Methods in Natural Language Processing*, 2020. The dissertation author was one of the primary investigators and author of this paper.

Chapter 5, in part, is a reprint of the material as it appears in “Unsupervised Enrichment of Persona-grounded Dialog with Background Stories” by Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, Harsh Jhamtani, which was published in *Association for Computational Linguistics, Main*, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in part, is a reprint of the material as it appears in “Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection” by Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley, which was published in *Association for Computational Linguistics*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 7, in part, is a reprint of the material as it appears in “Ask what’s missing and what’s useful: Improving Clarification Question Generation using Global Knowledge” by Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, Julian McAuley, which was published in *North American Chapter of the Association for Computational Linguistics*, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 8, in part, is a reprint of the material as it appears in “Knowledge-grounded Self-rationalization via Extractive and Natural Language Explanations” by Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, Julian McAuley, which was published in *International Conference on Machine Learning*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 9, in part, is a reprint of the material as it appears in “InterFair: Debiasing with Natural Language Feedback for Fair Interpretable Predictions” by Bodhisattwa Prasad Majumder*, Zexue He*, Julian McAuley, which is being prepared for submission. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2015	B.Tech in Electronics & Telecommunication Engineering, Jadavpur University
2017	Post Graduate Diploma in Business Analytics, IIT Kharagpur
2017–2018	Statistical Analyst, Walmart Labs
2020	M.S. in Computer Science and Engineering, University of California, San Diego
2023	Ph.D. in Computer Science and Engineering, University of California, San Diego

PUBLICATIONS

See the full list of publications on my website: https://www.majumderb.com/.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley, "Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions", *Empirical Methods in Natural Language Processing*, 2020.

Bodhisattwa Prasad Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley, "Interview: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding", *Empirical Methods in Natural Language Processing*, 2020.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, Julian McAuley, "Ask what's missing and what's useful: Improving Clarification Question Generation using Global Knowledge", *North American Chapter of the Association for Computational Linguistics*, 2021.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, Harsh Jhamtani, "Unsupervised Enrichment of Persona-grounded Dialog with Background Stories", *Association for Computational Linguistics, Main*, 2021.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley, "Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection", *Association for Computational Linguistics*, 2022.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, Julian McAuley, "Knowledge-grounded Self-rationalization via Extractive and Natural Language Explanations", *International Conference on Machine Learning*, 2022.

Bodhisattwa Prasad Majumder*, Zexue He*, Julian McAuley, "InterFair: Debiasing with Natural Language Feedback for Fair Interpretable Predictions", *presented in Workshop on Interactive Learning for Natural Language Processing, NeurIPS*, 2022.

ABSTRACT OF THE DISSERTATION

User-centric Natural Language Processing

by

Bodhisattwa Prasad Majumder

Doctor of Philosophy in Computer Science

University of California San Diego, 2023

Professor Julian McAuley, Chair

Artificial Intelligence (AI) systems that use language generation models hold incredible promise to assist humans to perform complex decision-making tasks. State-of-the-art language generation models can produce engaging content, reason about the world, and retrieve relevant information for an information-seeking task. However, these models often ignore sparse, long-tail knowledge about individual users, cultural subtleties, and domain-specific knowledge, preventing end users from reaping the full benefit of the scale. In this dissertation, we redesign AI systems to start with individual needs.

Ideally, a user-centric AI system must be grounded in the real-world, produce faithful chains of reasoning to explain its prediction, and align with the user's preferences. We elevate existing AI systems with *knowledge*, *explanations*, and *interactions* and develop both training-time and post-hoc techniques to make these systems user-centric. We show additional knowledge-grounding promotes user success in achieving conversational goals while using a conversational AI system. We demonstrate that AI explanations, when attributed to world knowledge, render them to be faithful and consistent. Finally, we discover that user-centric interventionist approach can help users obtain more equitable predictions backed by faithful explanations as compared to a black-box counterpart. In summary, our research establishes that increased effectiveness, explainability, and equitability can be achieved through knowledge-grounding and user-centric approaches to personalize AI models—a long-standing goal of artificial general intelligence.

Chapter 1

Introduction

Artificial Intelligence (AI) systems that use large language models (LLMs), such as GPT-3[1], have made incredible progress in performing complex tasks that humans otherwise do effortlessly. Current AI models can reason about the world, follow instructions during navigation, and generate multimedia content with remarkable efficiency. However, these AI models often ignore the long tail of the information, such as user preferences, cultural nuances, and domain knowledge, blocking end users from reaping the full benefit of the scale. Consequently, it is vital to update these AI models to make them more user-centric. My research stands at the heart of the following question:

What can we achieve by redesigning AI systems to start with individual needs?

1.1 User-centric AI

Users continuously seek to access information to make a decision, reduce uncertainty, or accomplish a task. However, today's AI systems can hardly assist in the above scenarios. First, current systems highly focus on high-frequency information, ignore the long-tail (e.g., cultural nuances), and become outdated or irrelevant in many situations. Second, current methods are often not interpretable, limiting users from fully trusting the prediction. Even systems that produce explanations primarily focus on low-level interpretations (e.g., just showing the most

---

[1] https://en.wikipedia.org/wiki/GPT-3

toxic words to explain a tweet as toxic), failing to address the complex reasoning associated with social and historical context. Third, existing AI systems only solve a narrow class of problems (that have a unique *correct* answer) and cannot adapt to cases where outputs are personalized or decisions are subjective (e.g., computational ethics).

Existing techniques for interpretable ML typically focus on producing (for example) attention maps, visualizations of weight vectors, etc. These types of interpretations demonstrate the features that were useful to make a prediction, but they require considerable expertise to interpret. For example, an attention-map over image regions might be useful as an interpretation from the perspective of a machine learning expert, but such an interpretations is generally not useful to the general public or an end user who is not familiar with how an image classifier works. As such, existing interpretable ML techniques are mostly useful as diagnostic tools (i.e., to assist experts) rather than something useful by laypeople.

Furthermore, existing techniques often assume a single, objective explanation of a prediction, i.e., that there is a unique, ‘correct’ answer or prediction that should be generated by an interpretable ML system. This is an acceptable assumption e.g. in simple classification scenarios, but is likely not effective in settings where (1) predictions depend on context which must be included in the explanation; or (2) predictions need to be personalized to a user or explained in terms of highly subjective factors.

Both of the above issues are particularly notable in user-facing applications. Examples include recommendation algorithms, personalized health, and personalized education. For such applications it is useful to provide explanations interpretable by laypeople, and those explanations should be personalized to each individual user, e.g. an interpretable recommender system should explain why the content was surfaced to them in particular.

Ideally, a user-centric AI system must be aware of the surrounding world (*relevant*), produce faithful explanations (*trustworthy*), and align with the user’s preferences (*adaptive*) [Miller, 2019]. My research concisely addresses the above through **interactive explainability**, realized via three interwoven pillars:

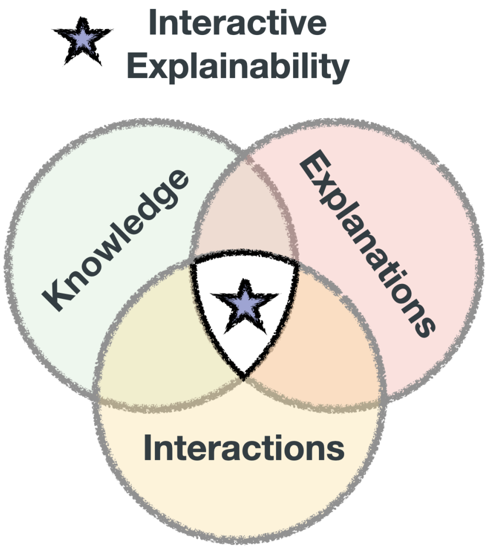


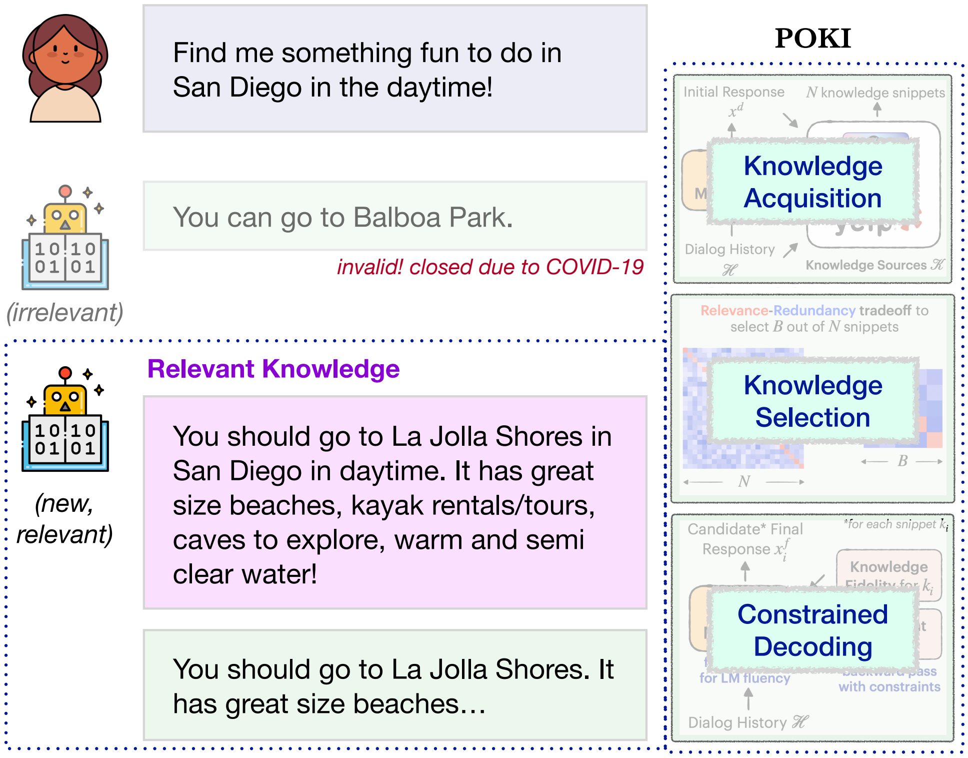
Figure 1.1. Interactive explainability as the foundation of user-centric AI systems

1. Knowledge: Discovering and deducing knowledge from context using external resources or via clarifying conversations,
2. Explanations: Enhancing machine predictions and their explanations by aligning the model's reasoning process with world and personal knowledge,
3. Interactions: Enabling users to critique and update model beliefs to align predictions and explanations to their personal, social, and subjective contexts.

1.2 Knowledge Grounding

Knowledge in its surface form is either propositional (facts) or the perception of skills, objects, and events (commonsense). However, required contextual knowledge greatly varies based on the personal preferences of the user, subjectivity associated with the context, and availability of resources to acquire such knowledge.

When AI models act as *experts* in a knowledge-seeking scenario (e.g., seeking recommendations, obtaining explanations), they often ignore subjective preferences or become limited to the knowledge they are trained on. We discovered that a model trained on persona-grounded



Find me something fun to do in San Diego in the daytime!

You can go to Balboa Park.

You should go to La Jolla Shores in San Diego in daytime. It has great size beaches, kayak rentals/tours, caves to explore, warm and semi clear water!

You should go to La Jolla Shores. It has great size beaches…

POKI
invalid! closed due to COVID-19
(irrelevant)
Relevant Knowledge
(new, relevant)
Initial Response x^d
N knowledge snippets
Knowledge Acquisition
Dialog History $\mathcal{H}$
Knowledge Sources $\mathcal{K}$
Relevance-Redundancy tradeoff to select B out of N snippets
Knowledge Selection
B
N
Candidate* Final Response x^f_i
*for each snippet k_i
Knowledge Fidelity for k_i
Constrained Decoding
for LM fluency
with constraints
Dialog History $\mathcal{H}$

Figure 1.2. Post-hoc Knowledge Injection (POKI) in a dialog model that was trained on limited knowledge

conversations could not deduce implicit knowledge from the dialog context. For example, humans can easily infer that *if someone likes hiking, they may love nature or want to be fit*; in contrast, traditional generative models fail to acknowledge it.

To remedy this, we devised a training-time knowledge augmentation paradigm that expands a textual context into possible inferential knowledge and then augments therein. Critically, we used a retrieve-generate framework that first uses external commonsense knowledge graphs (e.g., ATOMIC [Sap et al., 2019]), web-scale corpora (e.g., Yelp Reviews), or generative models (e.g., GPT3, COMET [Bosselut et al., 2019]) as knowledge sources and then augments the retrieved knowledge in the dialog model using variational learning. We found that humans predominantly prefer our generated responses as they are highly diverse, attributable, and controllable with input persona.

Training-time knowledge augmentation requires continuous fine-tuning to keep models up-to-date, resulting in higher carbon footprints. We developed post-hoc approaches to knowledge acquisition and injection for existing dialog models to make the process more lightweight and greener. It also applies when the underlying model's parameters are not accessible or updatable. This time, we retrieved the additional knowledge and used a *post-hoc* gradient-based method (POKI) to inject new knowledge into a generated dialog response both at inference time. We showed (see Figure 1.2) that when we injected up-to-date knowledge (e.g., post-COVID travel regulations) in an existing dialog model (let's say, trained in pre-COVID time), users could efficiently (re)use it to reach their conversational goals (e.g., planning their travel in 2022).

We further realized that knowledge-augmentation techniques detailed so far will still be ineffective when the context is ambiguous. For example, suppose a user is looking for a *travel recommendation*. Several nuances (e.g., number of travelers, location, and transport preferences) can change the knowledge requirement given the same context. We later developed a question generation framework to estimate the missing information, pose relevant and useful clarifying questions to reduce ambiguity, and gather knowledge that aligns with user preferences.

Finally, we tested our knowledge-augmentation techniques at a scale of millions of users. We showed a 65% improvement in user-satisfaction and more than 180% increase in user-engagement by making a dialog agent knowledge-aware and up-to-date. Our proposal for knowledge-grounded dialog system was selected in the finals of Amazon Alexa Prize, as 1 of 10 teams from 300 international participants, and awarded $250,000.

1.3 Generating Explanations

Machines often achieve near human-level performance in many tasks by mostly employing a different approach than humans. Hence, it is crucial to understand the model's underlying reasoning for better scientific understanding and improved trust. Moreover, to make models more user-centric, these explanations must be comprehensive, personalized, and attributable to world knowledge.

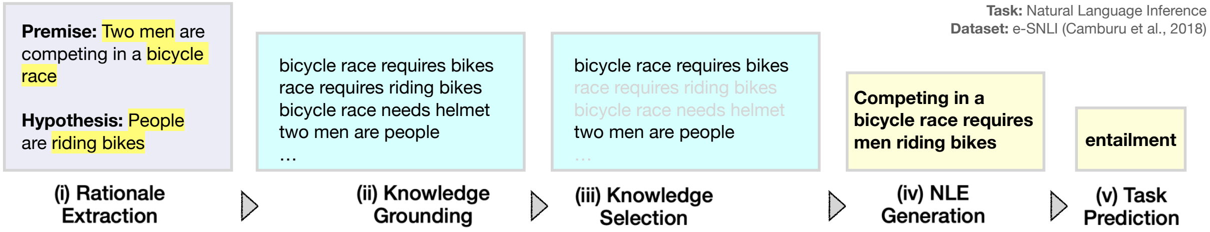

Task: Natural Language Inference
Dataset: e-SNLI (Camburu et al., 2018)
Premise: Two men are competing in a bicycle race
Hypothesis: People are riding bikes
(i) Rationale Extraction
bicycle race requires bikes
race requires riding bikes
bicycle race needs helmet
two men are people
...
(ii) Knowledge Grounding
bicycle race requires bikes
race requires riding bikes
bicycle race needs helmet
two men are people
...
(iii) Knowledge Selection
Competing in a bicycle race requires men riding bikes
(iv) NLE Generation
entailment
(v) Task Prediction

Figure 1.3. REXC bridges extractive rationales and abstractive explanations using background knowledge to improve plausibility, accuracy, and attribution.

We primarily worked with expressive forms of explanations, such as rationales (predictive parts of input) or natural language explanations, that could provide more accessibility to users and cover subjective contexts. Upon investigation, we found that existing explainable models often lack background knowledge, affecting task performance and explanation quality. We showed adequate knowledge grounding for three natural language tasks and two vision-language tasks could improve the quality of the explanations to be state-of-the-art (in REXC, see Figure 1.3). Additionally, we achieved the best task performance across all equivalent explainable models—indicating that REXC closes the critical gap between task performance and explainability.

Knowledge-grounding to improve explanation quality further gave rise to several emergent properties of the explanations: factuality, robustness, and faithfulness—critical for enabling the user to take actions based on the explanations. We observed that generated explanations exhibit a high degree of faithfulness; a similar observation was made for T5-based self-rationalizing models [Wiegreffe et al., 2021a]. Similarly, these models are more robust to knowledge-based adversarial attacks (e.g., change of entity, negations) than not knowledge-grounded models. This effect is more pronounced in domain-specific applications, such as e-commerce, where several state-of-the-art explainable models succumb to the issue of hallucination.

1.4 Engaging in Interactions

Despite our best efforts to make AI systems knowledgeable and explainable, data around us will still be inherently biased and limited by its origin. The models we build will be less than

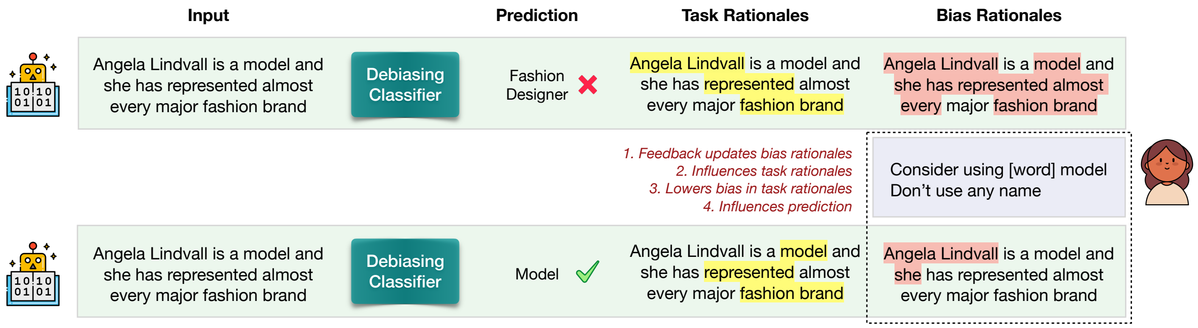

Input
Prediction
Task Rationales
Bias Rationales
Angela Lindvall is a model and she has represented almost every major fashion brand
Debiasing Classifier
Fashion Designer
Angela Lindvall is a model and she has represented almost every major fashion brand
Angela Lindvall is a model and she has represented almost every major fashion brand
1. Feedback updates bias rationales
2. Influences task rationales
3. Lowers bias in task rationales
4. Influences prediction
Consider using [word] model
Don't use any name
Angela Lindvall is a model and she has represented almost every major fashion brand
Debiasing Classifier
Model
Angela Lindvall is a model and she has represented almost every major fashion brand
Angela Lindvall is a model and she has represented almost every major fashion brand

Figure 1.4. INTERFAIR, a new interactive paradigm of controllable debiasing. Users can update the model's belief about sensitive information to mitigate bias while maintaining the task performance.

perfect. Moreover, humans also constantly modify their expectations from the AI models. Hence, the gold solution remains to make the user a part of the learning paradigm. To this end, designing, building, and evaluating interactive models can propel our progress toward anthropomorphic AI systems. My proposal was supported by Adobe Research Fellowship (2022) and Qualcomm Innovation Fellowship (2020).

We merely understand black-box models and how they encode complex social contexts into model parameters. In contrast, we are first to show that rationales are useful in exposing the model's understanding of complex social contexts. We discovered that most debiasing models remove sensitive information too harshly disregarding task performance. In [He et al., 2022], I proposed remedies to constrain the model explanations to be minimally biased but retain the high task performance. We build a critiquing framework to improve debiasing performance bringing users into the loop (INTERFAIR). I observed that the trade-off between task performance and bias mitigation greatly varies between users [Yaghini et al., 2021] and is often hard to achieve via purely learning from data [Zhang et al., 2018a]. Figure 1.4 shows how users can modify the amount of bias (here, removing the gendered name) in model explanations to balance bias mitigation and prediction accuracy.

1.5 Summary

While chapter 1 introduces the motivation behind a user-centric AI systems, chapter 2 lays the foundation of various knowledge sources and text generative models—the building blocks of our proposed user-centic systems.

Chapter 3, 4, 5, 6, and 7 broadly concerns knowledge grounding.

Chapter 3 discusses a variational learning framework to capture commonsense implications of input persona in a persona-grounded dialog agent using richer expansions obtained from existing commonsense knowledge bases.

Chapter 4 includes the first large-scale analysis of discourse in media dialog ("Interview" - 105K conversations) and its impact on generative modeling of dialog turns, with a focus on interrogative patterns and use of external knowledge.

Chapter 5 discusses an unsupervised gradient-based rewriting framework to adapt potential background stories to an existing persona-grounded dialog.

Chapter 6 proposes a post-hoc knowledge-injection technique that first retrieves and selects a diverse set of relevant knowledge snippets and further inject them into an initial response from an exisiting dialog model. Enriching dialog responses at decoding time with external knowledge (without re-training the existing models) promotes achieving conversational goals.

Chapter 7 introduces a two-stage framework that 1) estimates missing information from the global knowledge of similar contexts, and 2) conditionally generates useful questions using gradient-based decoding based on a usefulness scorer at the inference time.

Chapter 8 invests in generating explanations. It proposes a unified framework to map extractive rationales and abstractive natural language explanations (NLE) of AI models using commonsense. In this chapter, we establish new state-of-the-art in NLE generation, rationale extraction and predictive task performance.

Chapter 9 hints at the possibility of engaging users in interactions with an AI system. We tackle the issue of fairness in debiasing (i.e. balancing task performance and bias mitigation) and argue that it is subjective and difficult to learn from data. Using an interactive setup, we enable users to provide feedback and achieve a better balance, supported by controllable explanations.

Chapter 10 refers to the related work for all novel work introduced in previous chapter. Chapter 11 concludes the dissertation while discussing the novel possibilities and future outlook towards user-centric AI systems.

Chapter 2

Background

Language generation is one of the core tasks in natural language processing that has seen tremendous performance enhancement in recent years. On the other hand, achieving commonsense reasoning by machines has been a long-standing goal of artificial intelligence. Despite an increased focus on language generation, recent works revealed that state-of-the-art generative systems often struggle to generate plausible text consistent with the real-world. With the availability of good-quality datasets and strong computational models, we have made progress towards machine reasoning. In this chapter, we provide a background of various downstream generation tasks that require commonsense and external knowledge.

Commonsense knowledge, such as knowing that “nature has a healing effect over mind" or “ice makes the road slippery," is crucial for everyday navigation in the real world [Davis and Marcus, 2015]. But making machines to understand such human-like commonsense is yet to be fully achieved. Still, the tremendous advancement of artificial intelligence has shown that machines are somewhat capable of understanding real-world concepts. More so, they are capable of generating content that can easily be perceived by a human. Language generation is one such machine capability that has been investigated recently with great interest. While many previous works mainly focused on commonsense understanding, a few targeted to incorporate commonsense in generative systems. This is particularly useful since generated text should be meaningful, factual, and consistent with the real world.

Language Modeling:
Barrack’s wife is Hillary
The capital of India is the city
St. Louis is a city in the state of Oldham

Dialog Generation:
Bot: Today, I went to the central park with my dog.
User: I am not an animal lover.
Bot: Me too. I don’t have a pet.

Story Generation:
Harry shot Leo and tried to run away. The night was dark
and scary. (…) Harry invited Leo for dinner.

Figure 2.1. State-of-the-art generation models struggle to generate a factual text [IV et al., 2019], respond a user’s query [Welleck et al., 2019], or generate a story with plausible story-line [Guan et al., 2020].

While the literature on language generation is rich, recent works [IV et al., 2019, Welleck et al., 2019, Guan et al., 2020] showed that these systems are often not useful in real-world settings, mainly due to the inconsistency and factual incoherence. For example, a machine-generated dialog will not be helpful if it becomes inconsistent with its already uttered facts. Similarly, a generated story will not be accepted well if the plot outline lacks commonsense or temporal reasoning. Incorporating commonsense was particularly challenging due to a lack of commonsense resources, lack of strong computational models, and improper representation techniques, to name a few. See Figure 2.1 for erroneous generations by state-of-the-art models across various language generation tasks.

Recent studies have slowly started focusing on addressing the above issues. A great effort is made to standardize the commonsense knowledge bases and expands them to a reasonable scale. Strong pre-trained language models have shown how to accomplish reasonable performance in generation tasks; those were impossible even a few years back [Brown et al., 2020]. Multi-task learning, and graph neural networks have provided efficient ways to infuse external knowledge and obtain better representations of concepts. We believe that this is a timely survey that tries to cover various recent approaches towards generating language with commonsense.

Here, we first define commonsense and discuss various available resources to aid downstream generation tasks. After giving a preview of generation tasks that require commonsense, we discuss a way of incorporating commonsense in a generative system: using sequence transduction models.

2.1 Commonsense Knowledge

In his work on bounded rationality, Kahneman [Kahneman, 2003] defined three cognitive systems – perception, intuition, and reasoning. Perception and intuition are fast processes that get initiated automatically but can be realized in parallel. For example, a human can trigger a series of intuitive mechanisms at a time and can process them simultaneously. These two cognitive, though, come with a cost of slow learning more than the real world's association. On the other hand, the reasoning process is slow that generally involves a series of intermediate steps to resolve. This process is more controlled though it can be learned directly from observing rules and axioms.

The knowledge acquired to perform these three cognitive tasks is often termed as commonsense knowledge. This knowledge is usually associated with real-world observations and elements, hence often interchangeable terms as real-world knowledge too. As commonsense understanding is crucial and fundamental to human intelligence, it is believed to be instrumental for machine intelligence.

2.1.1 Definition of Commonsense

Commonsense can be defined as the necessary level of practical knowledge and reasoning concerning everyday situations and events commonly shared with *most* of the people[1]. For example, it is acceptable to keep the closet door open, but it is not acceptable to keep the fridge

---

[1] https://homes.cs.washington.edu/~msap/acl2020-commonsense/

door open, as it might spoil the food inside. There are always counterexamples to such scenarios such a fridge door can be kept open when it is empty – but these cases are rare and not ‘common’ to ‘most’ of the people.

2.1.2 Resources

Commonsense knowledge is acquired through the world’s human experiences and generally reasoned by an individual’s mental model. But to incorporate commonsense knowledge into machine intelligence, we need tangible resources that a machine can rely on or draw inference from. Commonsense resources are generally represented as a bank of knowledge that enables machines to reason about the world [Kintsch, 1988].

Existing Knowledge Bases are represented in two formats – 1. in symbolic logic (e.g. LISP-style logic) and 2. in natural language. Representations stored in symbolic format can be extracted automatically, whereas knowledge represented in natural language requires extra supervision. Even though they are of high quality, symbolic representation is often restricted to the experts, whereas knowledge represented in natural language can be accessible to non-experts. We will show later that knowledge represented in natural language is more useful than symbolic representations for specific downstream tasks. Here, we will briefly survey existing commonsense resources used in various capacities in the downstream tasks.

NELL

NELL is a dynamic bank of knowledge that ‘reads’ the web and represents facts into a symbolic representation [Carlson et al., 2010, Mitchell et al., 2015]. NELL has collected over 50 million facts with associated confidence in the facts extracted. Primarily NELL works on employing an information extraction algorithm that finds out entities or nouns and tried to relate them via a defined relation. For example: ‘**San Diego Zoo** is a popular **travel attraction**’ – where San Diego Zoo is attributed as a travel attraction. While these facts are useful for historical consistency and correctness, NELL also contains various attributes of a range of objects in the world, such as fruits, food, chemicals, species, etc. closer to commonsense knowledge we defined earlier.

OpenCyc

OpenCyc (4.0) is another symbolic knowledge base that holds 239,000 concepts and 2,093,00 facts about them [Lenat and Lisp]. OopenCyc is extensive to gather facts from the web, much like NELL, but mainly focuses on establishing specific and detailed ontology for implicit and commonsense knowledge.

ConceptNet

Unlike NELL and OpenCyc, ConcepNet holds meanings of words and concepts that people use in natural language [Speer et al., 2017]. This semantic network is built by a pipeline of crowd-sourcing and automatic extraction, spanning over multiple languages. It represents the general world knowledge using a graph representation (i.e. nodes as concepts and edges as relations) involved in understanding the meanings behind the words people use. Speer et al. [2017] provides an example of such knowledge: 'A **net** is used for **catching fish**.' Initial works [Turney, 2013, Speer et al., 2017] showed that when combined with distributional semantics of words i.e., word embeddings, the semantic meaning inferred from ConceptNet boosts language understanding tasks such as solving analogies.

ATOMIC

Going beyond word meaning, ATOMIC tries to capture world events and connect them via casual relationships [Sap et al., 2019]. Like ConceptNet, ATOMIC is also represented using natural language. The inferrential knowledge presented in ATOMIC makes it a very powerful commonsense resource for downstream tasks where reasoning is involved. Specifically, ATOMIC provides tuples that belong to nine relation types spanning over cause-effect interrelations between events: `oEffect`, `oReact`, `oWant`, `xAttr`, `xEffect`, `xIntent`, `xNeed`, `xReact`, and `xWant`—where a prefix '`x`' indicates an effect or cause on the person and '`o`' denotes the same on others. For example: If PersonX pays PersonY a **compliment**, PersonY will be **flattered**.

Beyond these common resources and knowledge bases, we have seen an influx of new datasets targeted towards commonsense reasoning such as SWAG [Zellers et al., 2018], HellaSwag [Zellers et al., 2019b], etc. These resources can be additionally used to guide the machine to acquire commonsense while performing a downstream task such as text generation.

2.2 Language Generation Tasks Requiring Background Knowledge

Text generation or natural language generation is one of the core tasks that has seen a sharp increase in interest. A recent breakthrough has enabled machines to generate text based on existing works (e.g usually written by a human) or from scratch. While fluency has been the primary focus of text generation algorithms, there was a little focus on achieving factual or commonsensical consistency in the generated text. Recently, some works explicitly focused on maintaining commonsense in the generated text [Bosselut et al., 2019, Guan et al., 2020, Zhou et al., 2018, Mao et al., 2019]. Here we will briefly list a set of language generation tasks that require commonsense to be preserved in the generation.

Story generation is the task of automatically producing compelling creative writing. Various structural elements in a story should adhere to some level of commonsense or logic, such as a story outline must follow a situation logic; a set of subevents must follow a temporal logic [Guan et al., 2020]. Similarly, other creative language generation tasks such as simile generation, sarcasm generation [Chakrabarty et al., 2020a] demands commonsense to be even valid.

Semantic understanding for dialog generation is often related to understanding underlying commonsense knowledge. Especially when off-the-shelf generative models do not perform reasoning based on the dialog context, it is crucial to explicitly signal the model with such knowledge. We see a prevalence of commonsense knowledge bases as an external resource to aid effective dialog generation in open-domain chitchat settings [Wu et al., 2020a, Young et al., 2018, Wu et al., 2020b].

In literature, we also find other generation tasks such as caption generation [Fang et al., 2020] and summarization [Amplayo et al., 2018] that utilize commonsense knowledge bases as an external resource for the respective generative models.

2.3 Sequence Transduction Models

The task of language generation often relies on the primary task of language modeling. A language model learns to assign the probability of a sequence of tokens $X = x_1, \ldots, x_T$ using the factorization rule of conditional probability:

$$P(X) = \prod_{t=1}^{T} P(x_t | x_{1:t-1}). \tag{2.1}$$

The model is trained using a cross-entropy loss between ground truth tokens and predicted probabilities given current tokens for the next time-step. We aim to incorporate commonsense in the generative system since the language model only focuses on language fluency and often lacks logical consistency. In the neural era, sequence transduction models are generally used for language generation tasks, which mimics a language model's working on the target given the conditional input.

2.3.1 Pre-training

[Bosselut et al., 2019] is one of the first works that incorporated commonsense knowledge as part of pre-training. COMET or Commonsense Transformer is a pre-training framework that adapts to construct commonsense knowledge bases using a language model that takes a seed set of knowledge tuples as the input. COMET is trained on tuples derived from both ConceptNet and ATOMIC knowledge bases, we discussed in Section 2.1.2. The pre-training objective is to generate new nodes and edges for the commonsense knowledge graph leveraging the language model representation learned.

COMET takes input as a tuple of tokens directly obtained from the knowledge graphs where the tuples are in the $\{s, r, o\}$ format. The COMET aims to generate o, the object, given the subject (s) and the relation (r). COMET uses OpenAI-GPT [Radford et al., 2018] as the generative model. To honor the original specification of GPT, COMET concatenates the subject and the relation and learns to generate follow-up commonsense knowledge in the form of $\{o\}$. To remove noise and superficial input, the authors masked the tokens that are not entities or relations.

The loss function for COMET is the conditional log-likelihood, with the target being the object (o) tokens. COMET uses ATOMIC and ConceptNet as the seed knowledge bases. The experiment shows that when the language model is initialized with pre-trained weights, the generation quality is higher.

Sequence transduction models are becoming increasingly popular due to the recent advent of powerful pre-trained language models. Multi-tasking is a reasonable approach to infuse commonsense into the generator system, but the process is explicit. There are debates that a weak auxiliary task may not be sufficient to change the decoder parameters pre-learned on a significantly large dataset. Also, current transformer based sequence-to-sequence models do not provide flexibility in architectural modifications; hence, more investigations are needed in that direction.

2.4 Large Language Models

2.5 Conclusion

In this chapter, we survey the recent advances in the intersection of language generation, commonsense reasoning, and knowledge grounding. We introduced a high-level view of various commonsense resources and how they can be used in language generation tasks. We then list various text generation tasks where the consistency of commonsense and world knowledge in the generated text is incredibly important. We discussed sequence transduction models and the evolution of the large language models as powerful language generation models.

Chapter 3

COMPAC: Persona-grounded Dialog with Commonsense Expansions

In this chapter, we dive deep into grounding dialog models with external knowledge. Here, we will focus on commonsense knowledge. Dialog models are required to be grounded in user persona, i.e., preferences and personality of the users. However, existing persona-grounded dialog models often fail to capture the simple implications of given persona descriptions, something that humans are able to do seamlessly. For example, state-of-the-art models cannot infer that interest in hiking might imply a love for nature or a longing for a break. In this chapter, we propose to expand available persona sentences using existing commonsense knowledge bases and paraphrasing resources to imbue dialog models with access to an expanded and richer set of persona descriptions. Additionally, we introduce *fine-grained* grounding on personas by encouraging the model to make a discrete choice among persona sentences while synthesizing a dialog response. Since such a choice is not observed in the data, we model it using a discrete latent random variable and use variational learning to sample from hundreds of persona expansions. Our model outperforms competitive baselines on the PERSONA-CHAT dataset in terms of dialog quality and diversity while achieving persona-consistent and controllable dialog generation.

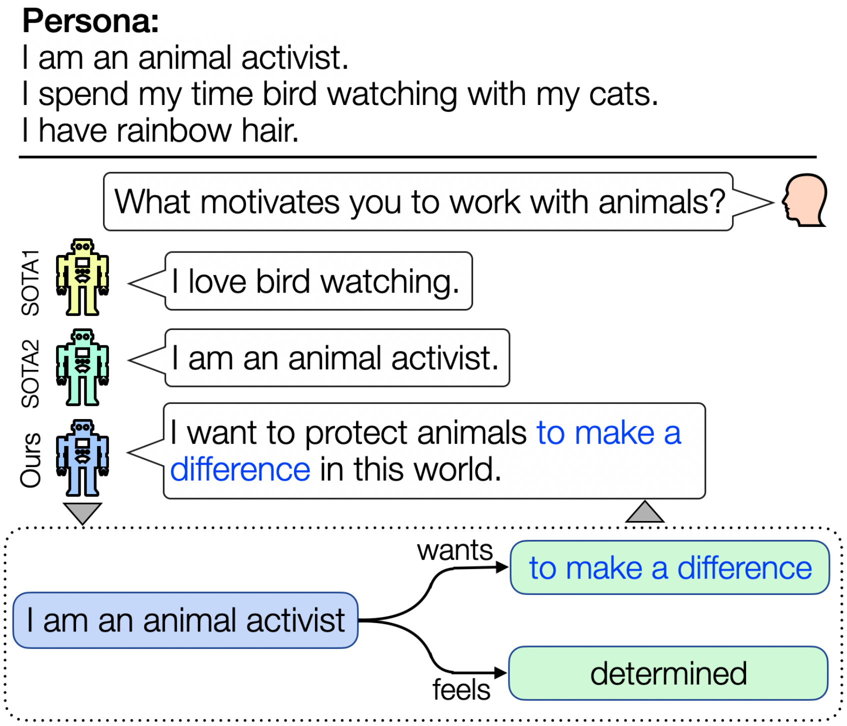


Figure 3.1. State-of-the-art models struggle to respond a user's query, where generating an engaging response depends on commonsense reasoning.

3.1 Introduction

Persona-grounded dialog generation is a 'chit-chat' dialog setup where a dialog agent is expected to communicate based on a given profile [Zhang et al., 2018b]. Many recent works have focused on a popular benchmark dataset for this task: PERSONA-CHAT [Zhang et al., 2018b] that provides personas as a set of sentences along with each dialog (example in Figure 3.1). However, a careful analysis of state-of-the-art (SOTA) models reveals that they often struggle to respond to contexts that do not closely match given persona sentences, even when the implications might be obvious to a human.

For example, in Figure 3.1, the user asks an *indirect* question to the bot related to one of its persona sentences: *I am an animal activist*. SOTA1, which concatenates all persona sentences with dialog history and finetunes a pre-trained generative model (e.g. GPT2) [Wolf et al., 2019], fails to infer implied commonsense from the dialog context and conditions on

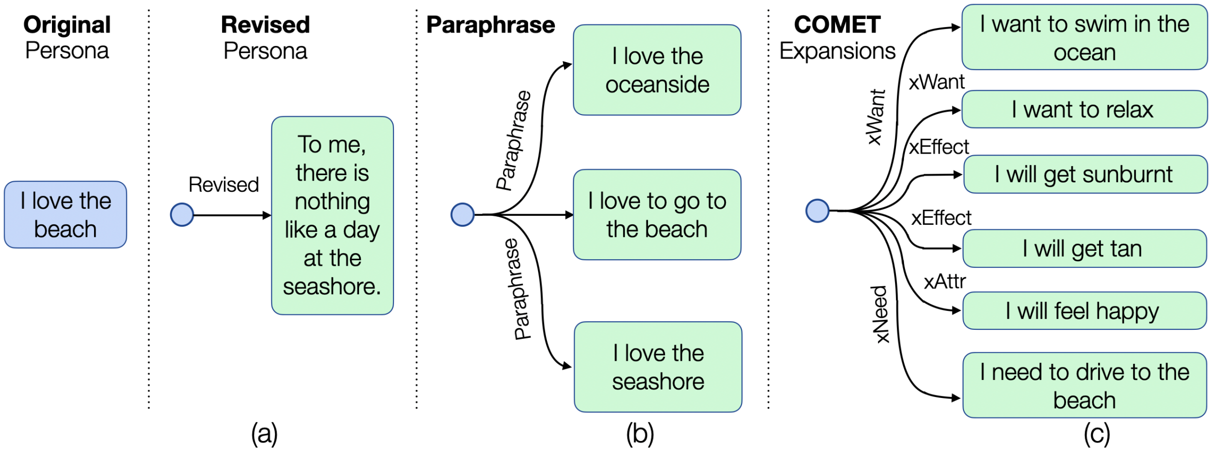


Figure 3.2. Expansions of an original persona via (a) human rewrite, (b) paraphrase, and (c) COMET.

an incorrect persona. SOTA2, which separately selects a persona sentence given the dialog history [Lian et al., 2019] manages to choose the correct persona but merely copies it as the final response. Neither approach is in general capable of responding to context that goes beyond what is explicitly mentioned in the available persona sentences, which limits consistent and interesting conversation. The goal of our model is to understand that being 'an animal activist' may imply that the person wants 'to make a difference' via their activity towards animals and synthesizes a context-consistent and engaging response.

In this chapter, we focus on making persona-grounded chatbots more consistent with personas and implicit dialog context. We present a framework to expand available persona sentences to their commonsense implications by using an existing commonsense knowledge base or paraphrasing resources (see Section 3.3). We endow our dialog model with these expansions directly rather than requiring the model to learn them from scratch for being context-consistent. We find that expansions derived from a commonsense knowledge base are more useful to provide engaging contextual information compared to other expansion sources.

We further propose a **Com**mon Sense and **P**ersona **A**ligned **C**hatbot[1] (**COMPAC**) which models choices over the *expanded* persona set via a discrete latent random variable (See Sec-

---

[1] Code is available at – https://github.com/majumderb/compac.

tion 3.4) as *fine-grained* persona grounding. Even though it is tractable to marginalize over all expansions, that would require a forward pass through the dialog generator for each outcome which is prohibitively slow during training. Instead, to accommodate hundreds of persona expansions, we train the model by optimizing a lower bound on the log-likelihood. We use amortized variational inference by approximating the true posterior using an inference network that eventually provides useful inductive bias. Particularly, we show that our Bayesian formulation for the fine-grained persona grounding was essential as simply providing expanded knowledge does not help the model generate better responses.

We also outperform competitive baselines in all dialog quality metrics as well as human evaluations which find COMPAC to be engaging and coherent. We demonstrate that COMPAC learns to be consistent with the dialog context with accurate persona grounding especially in the presence of commonsense expansions. Finally, we show that our model can reflect a change in response generation when a grounding persona is modified, indicating the possibility of controllable generation.

3.2 Persona Grounded Dialog

We use a popular benchmark dataset: PERSONA-CHAT [Zhang et al., 2018b] for our persona-grounded dialog generation task. It contains 10,907 dialogs between pairs of speakers where each speaker follows their own persona; 968 dialogs are used for validation and 1,000 for testing. Each speaker is described by 3-5 persona sentences. (e.g. ‘I love the beach’, ‘My mother is a medical doctor’). Out of 1,155 total unique personas, 100 are used for validation and 100 for testing.

The task of persona-grounded dialog generation is: given a dialog history H and grounding persona sentences S, we must predict the next utterance x (Summary of notations in Table 3.1). Hence a dialog model should maximize the likelihood $p(x|H,S)$. From the PERSONA-CHAT dataset, we use 131,438 utterances for training the dialog model, 15,602 for validation, and 15,024 for testing.

3.3 Persona Expansion

Persona sentences used in persona-grounded dialogs are instances of world events that often imply real-world consequences or richer information. For example, ‘I love surfing’ naturally implies that the person might be ‘adventurous’ or ‘loves the outdoors’. Similarly, it also means that the person wants ‘to go to the beach’ regularly. Inferring these *expansions* from the original fact is non-trivial without additional commonsense knowledge.

Zhang et al. [2018b] found evidence that having human written interpretations of a persona sentence via rephrasing often helps in providing novel information in persona grounding. While obtaining such expansions by manual rewriting is expensive, here we explore two automatic ways to generate them at scale and separately evaluate them on the downstream dialog modeling task.

3.3.1 COMET

COMET [Bosselut et al., 2019] is a framework that generates rich and diverse commonsense expansions of a given world event. It is a finetuned version of a pre-trained GPT2 [Radford, 2018] model on a pre-existing commonsense knowledge graph such as ATOMIC [Sap et al., 2019] that can generate novel nodes (events) and edges (relations), as seen in Figure 3.2c. Specifically, ATOMIC provides tuples that belong to nine relation types spanning over cause-effect interrelations between events: `oEffect`, `oReact`, `oWant`, `xAttr`, `xEffect`, `xIntent`, `xNeed`, `xReact`, and `xWant`—where a prefix ‘`x`’ indicates an effect or cause on the person and ‘`o`’ denotes the same on others. While we tried COMET finetuned on an alternative commonsense knowledge base (e.g.) ConceptNet, not all of the expansions were appropriate to describe a persona, mainly because we observe that persona sentences are *event*-like (‘I love to go to the beach’) as opposed to *concepts* such as ‘beach’. For more details on COMET and ATOMIC we refer the reader to [Bosselut et al., 2019] and [Sap et al., 2019] respectively.

We use the COMET framework to generate expansions for each persona sentence along the nine relation types that ATOMIC provides. We obtain different samples while decoding via beam search from COMET for more diverse and unique expansions, as shown in Figure 3.2c. We preprocess these expansions to add suitable prefixes to make them similar to the original persona. For example, expansions relating to `xWant` and `xAttr` are prefixed with ‘I want’ and ‘I am’ respectively. For each persona sentence, we generate 5 expansions per relation, i.e., in total we will obtain $5 \times 9 = 45$ expansions per persona sentence.

3.3.2 Paraphrasing

To explore alternative sources for generating commonsense expansions beyond COMET, we consider paraphrasing persona sentences. Paraphrases of a sentence convey almost the same meaning to a listener as the original. Often paraphrases use synonymous phrases or manipulate word-syntax of the original sentence, which implicitly involves both context comprehension and world knowledge [Zeng et al., 2019]. We obtain these in two ways:

Paraphrase Network To generate paraphrases at scale, we use an off-the-shelf paraphrasing system based on back-translation [Xie et al., 2019b, Federmann et al., 2019] with pre-trained language translation models. We make use of `En-Fr` and `Fr-En` pre-trained translation models as the components for back-translation.[2] While we tried other language pairs, the `En-Fr` pair proved the most satisfactory based on qualitative analysis on 500 samples. We generate 5 paraphrases per persona sentence, which readily provides more lexical and syntactic variants as shown in Figure 3.2b.

Manual Paraphrasing To compare with other expansions, we reuse manually written revised versions of persona sentences provided with PERSONA-CHAT [Zhang et al., 2018b] though these are limited to only one paraphrase per sentence. We call them **revised** for short (see Figure 3.2a).

---

[2] https://github.com/google-research/uda

3.4 Common sense and Persona Aligned Chatbot (COMPAC)

To infuse commonsense context in persona-grounded dialog generation, we imbue our dialog model with the expanded persona set instead of only original personas S. But these persona expansions lead to hundreds of new sentences as opposed to only a few given persona sentences which makes it infeasible to encode using a single transformer, as was done in prior works [Wolf et al., 2019]. Additionally, encoding all persona sentences as a single text input leads to a lack of interpretability i.e., it is not clear which persona sentence was used by the model in generating a particular response.

Instead, we propose **COMPAC**: **Com**mon Sense and **P**ersona **A**ligned **C**hatbot that allows us to make a *fine-grained* choice of a persona sentence to generate the target response. Let C denote a list of expended personas, derived from S (including S itself). We further add a null persona $\varnothing$ in C considering that some utterances can purely condition on the dialog context. We are interested in modeling the conditional $p(x|H,C) = p(z|H,C)p(x|z,H,C)$ where $z \in \{1,2,\ldots,|C|\}$ is a latent discrete random variable, unobserved in the data. Given the dialog history H, first we sample a particular persona sentence C_z from a *prior network* $p_\theta(z|H)$ (see Figure 6.2). Next, as depicted in Figure 6.2, the dialog response x is sampled from a *generator network* $p_\phi(x|H,C_z)$ by conditioning on the history H and chosen persona sentence C_z.

In the generative model described above, the latent variable z is a discrete random variable which points to a single persona sentence. This decision (of conditioning on a single persona sentence) was based on the observation that most dialog responses in the datasets under consideration are relevant to only one persona sentence. It is possible to allow for multiple persona sentences by defining z to pick a subset of $|C|$ persona sentences instead of picking a single sentence. We leave this as a possible future extension.

Table 3.1. Summary of notation used in this chapter

S	Set of original persona sentences
C	Set of expanded persona sentences (includes S and a null persona $\varnothing$)
H	Dialog history with alternative turns from each speaker
x	Target utterance
z	Discrete latent random variable $\in \{1,2,\ldots,|C|\}$
e	Mean of RoBERTa subword embeddings as an encoder
t_k	Expansion type for k-th expansion
f_i	i-th feature function for prior network; $i \in \{1,2,3\}$
θ	Parameters for prior network $p_\theta(z|H,C)$
ϕ	Parameters for generator network $p_\phi(x|H,C_z)$
α	Parameters for inference network $p_\alpha(z|x,H,C)$

3.4.1 Persona Choice Prior

The dialog history H can hold cues regarding which persona sentence might be applicable given the context. For example, in Figure 6.2 the historical context suggests that 'following fashion trends' can be a consequence of 'being fashionable'.

We encode both the dialog history H and persona sentence C_k by averaging RoBERTa subword embeddings [Liu et al., 2019] as $e(H)$ and $e(C_k)$. We use an implementation from HuggingFace for RoBERTa[3] with `roberta-base` as the pretrained model. Then we parameterize the prior $p_\theta(z|H,C)$ as a log-linear model with the following features:

Dialog history We obtain $f_1(H,C_k)$: a scalar feature using a bilinear product $\langle e(H), e(C_k)\rangle$ to align the persona sentences with the dialog history.

Expansion types Each k-th persona expansion corresponds to an expansion type t_k. In the case of COMET, these types are the nine commonsense relations provided by ATOMIC (see Section 3.3.1). For paraphrased expansions, we annotate each as type `paraphrase` and the original persona sentences as `original`. We consider two additional features with expansion types: (a) $f_2(t_k)$ that represents a global preference over the relation type embedded via a type embedding layer; and (b) $f_3(t_k,H)$ that appends the expansion type embedding with dialog

---

[3] https://huggingface.co/transformers/model_doc/roberta.html

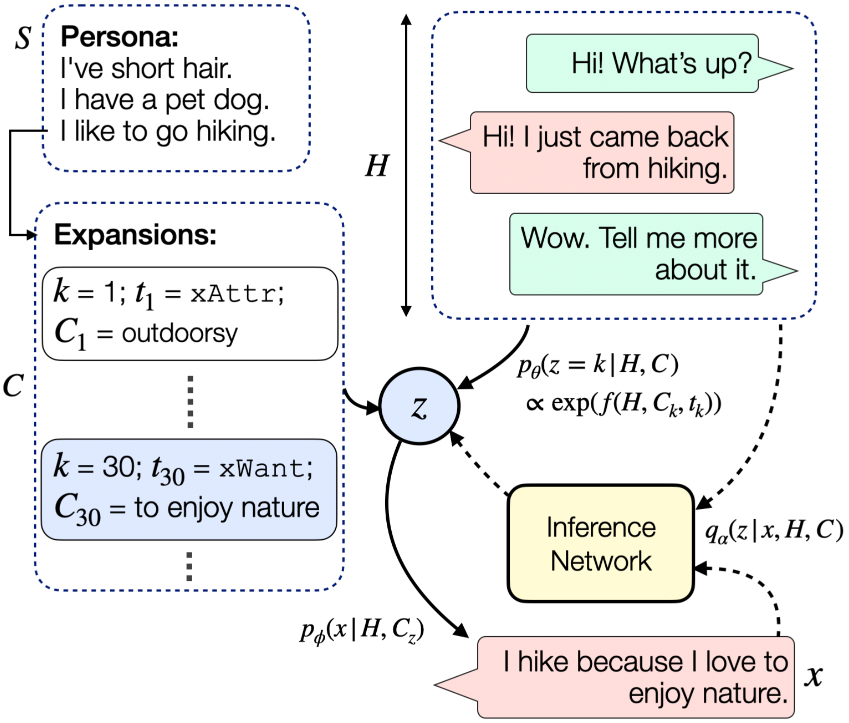

S
Persona:
I've short hair.
I have a pet dog.
I like to go hiking.
Expansions:
$k = 1; t_1 =$ xAttr;
C_1 = outdoorsy
$k = 30; t_{30} =$ xWant;
C_{30} = to enjoy nature
C
H
Hi! What's up?
Hi! I just came back from hiking.
Wow. Tell me more about it.
z
$p_\theta(z = k \mid H, C) \propto \exp(f(H, C_k, t_k))$
Inference Network
$q_\alpha(z \mid x, H, C)$
$p_\phi(x \mid H, C_z)$
I hike because I love to enjoy nature.
x

Figure 3.3. COMPAC samples a persona sentence from the prior and generates the response conditioned on the dialog context and sampled persona. The inference network is used only during training.

history encoding $e(H)$, followed by a linear layer to obtain a real-valued score for history-specific preference over the expansion type.

The dimension of the expansion type embedding was set to 5. Finally, the prior model can be represented concisely as $p_\theta(z = k|H,C) \propto \exp(f(H,C_k,t_k))$, where $f(H,C_k,t_k)$ is the sum $\lambda_1 * f_1(H,C_k) + \lambda_2 * f_2(t_k) + \lambda_3 * f_3(t_k,H)$ with λ_i's are trainable parameters.

3.4.2 Generator Network

Following prior work [Wolf et al., 2019], we use pre-trained GPT2 [Radford, 2018] (Transformer with 12 layers, 768 hidden size, 12 heads— `gpt2-small`[4]) to generate dialog responses given the dialog history H, with the selected persona sentence C_z prepended to it. In the case of C_z being the null persona, an empty string is prepended. We further append the target response x to the combined context $(C_z; H)$, and feed the sequence to GPT2, after tokenization.

---

[4] https://github.com/huggingface/transfer-learning-conv-ai

To distinguish between persona tokens, history tokens, and target response tokens, we use segment indicators—{`Persona`, `Speaker1`, `Speaker2`}—for which corresponding embeddings are learned via a separate segment embedding layer in the model. We add the segment embedding to the corresponding token embedding in the model input layer. To obtain the conditional likelihood $p_\phi(x|H,C_z)$, we only consider the target tokens for cross-entropy loss calculation.

Wolf et al. [2019] also leveraged incorrect responses given a dialog history from PERSONA-CHAT as negative samples in an auxiliary loss to encourage the correct candidate to obtain the highest likelihood compared to the incorrect ones. However, we did not find any improvement using this loss in COMPAC.

3.4.3 Learning and Inference

Our training data $\mathscr{D}$ consists of instances of dialog history H and ground truth dialog responses x. We train our model parameters θ and ϕ to maximize the likelihood of the target dialog response x given the dialog history: $\log p(x|H,C;\theta,\phi)$ totalled over $\mathscr{D}$. Since the discrete random variable z is unobserved in the training data, we must marginalize over z to compute the desired likelihood $p(x|H;\theta,\phi)$:

$$\log p(x|H;\theta,\phi) = \log \mathbb{E}_{z\sim p_\theta(z|H)}[p_\phi(x|z,H)];$$

where we drop C from the conditionals for simplicity.

Inference Network Note that the number of persona expansions is typically in the range 150-250, and thus it is computationally expensive to marginalize over the entire selection space of z during training. We instead optimize a variational lower bound (ELBO) of $\log p(x|H;\theta,\phi)$

given as

$$\mathbb{E}_{z \sim q_\alpha(z|H)}[\log p_\phi(x|z,H)]$$
$$- KL(q_\alpha(z|x,H)||p_\theta(z|H)),$$

where we use the inference network $q_\alpha(z|x,H)$ to compute the approximate posterior [Kingma and Welling, 2014]. In our initial experiments, we observe that using an inference network leads to better perplexity values than using samples from the prior.

The architecture of the inference network is similar to that of the prior network, a log-linear model. Along with the features related to dialog history and expansion types, we additionally include another scalar feature: a bilinear product $\langle x, C_k \rangle$ between the encoded persona and ground truth response x encoded with RoBERTa embeddings to align the persona choice according to the target utterance.

Optimization The parameters of the generator network (ϕ) and prior network (θ) are trained directly via back-propagation. Since z is a discrete latent variable, we use REINFORCE [Williams, 1992] to train the inference network parameters α. However, the REINFORCE estimator often suffers from high variance. To reduce the variance, we found it useful to (1) use a moving average baseline [Zhao et al., 2011]; and (2) regularize the prior network by penalizing the entropy of the output categorical distribution. To avoid KL mode collapse, we use KL-annealing [Bowman et al., 2016] where we linearly increase the weight of the KL term beginning from 0 to 1 as training progresses.

Decoding At decoding time, we first sample k from the prior $p_\theta(z|H,C)$, and then C_k is fed to the generator network. Following previous work [Wolf et al., 2019], we use nucleus sampling [Holtzman et al., 2020] (with $p = 0.95$) to decode the final response from the probabilities produced by the generator. We also found that high-temperature sampling from the prior often

Table 3.2. Dialog quality metrics on the PERSONA-CHAT test set. PPL=Perplexity, D-1/2=% of distinct uni- and bi-grams.

System	PPL	BLEU-1	BLEU-2	D-1	D-2
Original					
Per-CVAE	48.37	0.19	0.11	0.03	0.21
LIC + KS	30.50	0.18	0.07	0.07	0.24
GPT2	21.46	1.42	0.78	0.05	0.11
COMPAC-original	19.56	3.24	1.31	0.15	0.25
Paraphrased					
GPT2-revised	21.01	1.54	0.97	0.13	0.25
GPT2-paraphrase	21.57	1.61	0.86	0.16	0.35
COMPAC-revised	18.12	3.52	0.99	0.48	0.65
COMPAC-paraphrase	17.09	3.83	**1.87**	0.56	0.85
COMET					
GPT2-COMET	21.12	1.62	0.81	0.21	0.39
COMPAC	**16.21**	**4.12**	1.82	**0.87**	**1.07**

leads to more diverse generation.

3.5 Experiments

We conduct our experiments based on the following desiderata: (1) Do persona expansions help to generate high quality and diverse responses? (2) Does COMPAC achieve accurate persona grounding given a dialog context? (3) Does COMPAC enable persona-consistent and controllable generation?

3.5.1 Baselines

To demonstrate the efficacy of COMPAC, we compare it with three competitive baselines on the PERSONA-CHAT dataset:

1. **Per-CVAE:** A CVAE model that exploits persona sentences for diverse generation with an external memory [Song et al., 2019b]
2. **LIC + KS:** The best performing transformer model (Lost in Conversation i.e., LIC) in terms of human evaluation in the ConvAI2 NeurIPS competition [Dinan et al., 2019a]

Table 3.3. Pairwise comparison (in %) between responses generated by COMPAC vs. other baselines (og: original, par: paraphrase) as well as the Gold response. The highest entries are **bolded**.

COMPAC vs.	**GPT2**		**LIC + KS**		**GPT2-COMET**		**COMPAC-og**		**COMPAC-par**		**Gold**	
Metric ↓	win	loss	win	loss	win	loss	win	loss	win	loss	win	loss
Fluency	**81.2***	5.1	**83.2***	6.7	**90.5***	2.3	**68.0**	26.0	**65.0**	19.4	40.1	**42.5**
Engagement	**90.5***	3.3	**87.4**	5.9	**97.6***	0.5	**86.5***	10.5	**81.5***	10.5	**62.1***	30.5
Relevance	**78.2***	4.8	**78.0***	7.7	**93.2***	1.8	**65.5***	18.5	**62.1**	15.6	32.8	**54.6***

combined with a knowledge-selection (KS) mechanism Lian et al. [2019] to achieve state-of-the-art results on PERSONA-CHAT;

3. **GPT2:** Finetuned GPT2 on PERSONA-CHAT just by concatenating all persona sentences along with dialog history [Wolf et al., 2019] to obtain the best automatic metric in the ConvAI2 competition.

A minimal version of COMPAC is also considered, **COMPAC-original**, which only uses the original persona, for a direct comparison with other model architectures that only use the original persona. Furthermore, to justify the choice of fine-grained persona grounding for an effective utilization of persona expansions, we also consider baseline versions of GPT2 trained with each of the expansion strategies: **GPT2-revised**, **GPT2-paraphase**, and **GPT2-COMET**. To show that COMPAC can work with persona expansions derived from various sources, we compare with versions of COMPAC trained with paraphrase-based expansions: **COMPAC-revised** and **COMPAC-paraphrase**. By default, COMPAC indicates it is trained with COMET expansions.

3.5.2 Comparison of Dialog Quality

We measure perplexity for language modeling performance, and BLEU-1 [Papineni et al., 2002a] and BLEU-2 [Vedantam et al., 2015] scores between generated and gold utterances to measure the fidelity of the generated responses. Given our goal of generating engaging responses with novel information, we deem it important to consider the diversity in the generated responses which we measure using D-1 and D-2 (percentage of distinct uni- and bi-grams respectively) [Li

Table 3.4. Sample generations by different models.

Persona: I enjoy listening to classical music. I'm a Hindu. My favorite color is red.
User: Hi, recently I have got interests in religion.
GPT2: Hi! How are you?
COMPAC-original: I'm a Hindu.
COMPAC-revised: Hi! I am a Hindu too.
COMPAC-paraphrase: That's great. I am religious.
COMPAC: That's great. I regularly go to the temple and learn about Hinduism.

et al., 2016a].

Table 3.2 shows that COMPAC outperforms three competitive baselines when trained on the original persona in all quality metrics indicating the efficacy of our architecture. Moreover, when combined with persona expansions, we observe a modest 3-8 point decrease in perplexity and a large improvement in both BLEU and diversity scores which confirms that COMPAC successfully leverages the persona expansions to improve dialog quality. COMPAC trained with COMET expansions achieves the best performance both in terms of fidelity and diversity which shows that COMET expansions help the model to respond to implicit context with commonsense and to explore novel information. But with revised personas, we find that both COMPAC and GPT2 provide marginal performance gains, mirroring the observation from [Zhang et al., 2018b]. Finally we observe gradual degradation in performance when we trivially finetune GPT2 with paraphrase and COMET expansions. Note that GPT-2 could have implicitly learned to focus on a single persona attribute. However, the COMPAC performs better suggesting that fine-grained persona grounding acts as a useful inductive bias in effectively utilizing larger expansion sets.

3.5.3 Human Evaluation for Dialog Generation

Automatic evaluation of dialog systems is still notoriously unreliable [Liu et al., 2016, Novikova et al., 2017] and such systems should be evaluated by human users. Hence, we perform pairwise comparisons between responses generated our best system, COMPAC trained on COMET expansions, and responses generated by four strong baselines: GPT2, GPT2-COMET, COMPAC-original, COMPAC-paraphrase (the best COMPAC model with paraphrase expansions). We also consider the gold responses for comparison. We conduct a human evaluation with 100 test examples on three aspects critical for practical use: (1) **Fluency** measures whether the generated output is fluent (in English); (2) **Engagement** measures whether the generated response is engaging or interesting; and (3) **Relevance** measures whether the generated output is relevant with respect to the dialog history.

Table 3.3 shows that human annotators found responses generated by COMPAC trained with COMET expansions more engaging as compared to responses from all the baselines as well as the gold responses by statistically significant margins. Entries with * denote significance with $p < 0.05$ from bootstrap tests on 1000 subsets of size 50. This confirms our hypothesis that COMET expansions were helpful in adding novel content. Human judges also found that despite a significant drop in perplexity, COMPAC was not more fluent than COMPAC-original and COMPAC-paraphrase with statistical significance, indicating similar language modeling performance. We find the inter-annotator agreement, as measured by Cohen's kappa [Cohen, 1960], for fluency, engagement, and relevance were 0.62, 0.71, and 0.73 respectively.

3.5.4 Fine-grained Persona Grounding

Next we want to investigate the extent of COMPAC's ability to ground the response generation with a fine-grained persona choice as a probing experiment. Specifically, we want to measure whether our model can choose a coherent persona from the available persona sentences given the dialog context. Note that in persona-grounded chitchat, not all utterances are tied to a

Table 3.5. Assessment of persona grounding with and without inference network using the DNLI entailment set. Human evaluation (eval.) was conducted to measure the relevance when an expanded persona is chosen–all entries are statistically significant.

System	**Persona Entailment**		**Human**
	Prior	**Inference Network**	**eval.**
Original			
COMPAC-original	25.5	79.3	–
Paraphrased			
COMPAC-revised	20.6	78.9	40.6
COMPAC-paraphrase	27.8	87.3	67.8
COMET			
COMPAC	**37.9**	**96.4**	**87.3**

personas and could be purely based on dialog context. We find that 44% of the time the model selects the null persona ($\varnothing$) and conditions only on the dialog history. To assess the persona grounding for the remaining (56%) utterances, we perform (a) a persona entailment experiment, and (b) human evaluation.

Persona Entailment We adapt the Dialogue Natural Language Inference (DNLI) dataset [Welleck et al., 2019] and collect persona-utterance pairs that belong to an *entailment* relation. This results in a subset of 4,613 utterances with associated ground truth persona sentences in our test set. Next, we obtain a persona sentence by performing argmax over the prior $p_\theta(z|H,C)$ as well as the inference network $q_\alpha(z|x,H,C)$ from our COMPAC models and calculate accuracy with the ground truth persona. For models that use expanded personas, we track the original persona from the retrieved expansion for accuracy calculation. Table 3.5 shows that COMPAC with COMET achieves the most accurate persona grounding suggesting that inference networks can approximate the true posterior better when a commonsense persona is available for grounding. In the case of the prior, a better entailment accuracy than random chance $(1/5)$ confirms our choice of the history-conditioned prior network rather than a uniform prior.

Human Evaluation Since DNLI does not entail expanded personas, we conduct a human evaluation to judge the relevance of a chosen persona *expansion* sampled from the inference

Table 3.6. Conditional generation performance on the PERSONA-CHAT test set to show the similarity between generated responses and grounding persona sentences. We omit GPT2-based models since they do not select a particular persona sentence for grounding.

System	Unigram Overlap			BERT
	Recall	**Precision**	**F1**	**Score**
Original				
LIC + KS	10.4	34.2	15.3	–
COMPAC-original	14.9	39.1	21.6	57.2
Paraphrased				
COMPAC-revised	15.2	40.3	22.1	58.1
COMPAC-paraphrase	17.8	42.2	25.1	72.9
COMET				
COMPAC	**21.4**	**48.9**	**29.8**	**78.8**

network. Specifically, we ask: *Is this knowledge relevant to the given dialog history?*—with options as ‘Yes’, ‘No’, and ‘Uncertain’—and with 100 examples for each COMPAC variant that uses expanded personas. The inter-annotator agreement, as measured by Cohen’s kappa was 0.76. Again, Table 3.5 shows that models with COMET expansions can choose the most relevant persona sentence which corroborates our claim in persona entailment experiments. On average, we noticed that COMPAC with COMET expansions prefers to choose expanded personas 87% of the time out of all non-null persona choices. This reduces to 62% in the case COMPAC-paraphrase. In contrast, COMPAC-revised tends to select an original persona over an expansion more often.

3.5.5 Controllable Generation

Controllable generation of persona-grounded dialog can help to generalize the dialog agent to newer persona details just by changing the grounding in the conditional generator. While controllable text generation with a desired attribute has gained interest recently [Dathathri et al., 2020, Kong et al., 2019], we investigate the possibility of controlling generation with a desired persona and measure the performance of the conditional generator. For this, we observe a set of knowledge overlap metrics—the unigram recall/precision/F1 scores–from Dinan et al. [2019b]

Table 3.7. Controlled generation with edited persona

Performance	Example
Presence of changed entity **86%**	**Changing the key entity** Before: My favorite color is red After: My favorite color is **green**
	Conversation: User: What is your favorite color? Bot: My favorite color is **green**
BERT score with unedited persona: 46.2 edited persona: **74.6**	**Swapping with another expansion** Before: I want to swim in the ocean After: I want to buy a beach **umbrella**
	Conversation: User: What do you do at beaches? Bot: I will buy an **umbrella** at the beach

and BERT score [Zhang et al., 2020] for semantic similarity between the generated responses and the persona retrieved. Table 3.6 shows that conditional generation is strongest when COMPAC is trained with COMET suggesting commonsense expansions are more appropriate to the dialog context in influencing the response generation.

Next, we create a diagnostic dataset of 100 examples where we manually edit the persona by changing an entity in a persona sentence or swapping the selected persona expansion with another relevant one (See examples in Table 3.7) to directly measure controllability in response generation. We observe that COMPAC can successfully reflect the entity-change in the generated response based on the change in the persona grounding 86% of the time. For a swapped persona expansion, we also see a higher BERT score (74.6) between the edited persona and newly generated response as opposed to a lower score (46.2) with the unedited persona. Together with the qualitative examples in Table 3.7 this suggests that COMPAC supports controllable generation with contextually modified personas.

Table 3.8. Examples showing correct and incorrect persona choices in various dialog contexts by COMPAC model. COMPAC is capable of choosing a correct persona sentence (original or expanded) but sometimes the prior network fails to sample an appropriate one (third case).

Verdict	**Persona**	**Dialog History**	**Persona Choice (z)**	**Response**
Uses Original Persona	1. I'm the youngest of five. 2. I work at the hospital as a nurse. 3. I've pink hair.	What do you do for work?	I work at the hospital as a nurse.	I am a nurse.
Uses Expanded Persona	1. I just want to have fun with my friends. 2. I don't drink or do drugs or anything. 3. I am 19 and I cannot wait to move out of my parents home.	Are you enjoying life?	I just want to have fun with my friends → wants to have a party	Not really. I want to have a party.
Uses Improper Persona	1. I make a million dollars a year. 2. I'm married and have three kids. 3. I'm a baseball player.	I find it hard to support my family working at a bar. What about you?	Null persona (∅)	I enjoy my life.

3.5.6 Qualitative Analysis

Table 5.3 shows responses from different models for a sample dialog context. Qualitatively, we find that COMPAC with COMET expansions responds to the context with commonsense using novel content from a commonsense expansion (being Hindu → to learn about Hinduism), where other responses remain generic or incoherent. In Table 3.8, we illustrate responses generated by the COMPAC model along with the underlying persona choice sampled from the prior network. Cases show that COMPAC successfully chooses an original or an expanded persona sentence, as appropriate, but also defaults to the null persona (∅) that leads to a bland response.

3.6 Conclusion

In this work, we showed that expanding persona sentences with commonsense helps a dialog model to generate high-quality and diverse persona-grounded responses. Moreover, we found that *fine-grained* persona grounding is crucial to effectively condition on a large pool of commonsense persona expansions, which further provided additional controllability in conditional generation.

While our expansions are limited by the performance of COMET or paraphrase systems, we envision future work to train the dialog model end-to-end along with the expansion generation. As future work, we would like extend the prior network to sample more than one persona sentences by expanding the sample space of the discrete random variable to generate more interesting responses.

Chapter 3, in part, is a reprint of the material as it appears in “Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions” by Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley, which was published in *Empirical Methods in Natural Language Processing*, 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 4

INTERVIEW: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding

In this chapter, we will continue with the task of grounding dialog models with external knowledge. We perform the first large-scale analysis of discourse in media dialog and its impact on generative modeling of dialog turns, with a focus on interrogative patterns and use of external knowledge. Discourse analysis can help us understand modes of persuasion, entertainment, and information elicitation in such settings, but has been limited to manual review of small corpora. We introduce INTERVIEW—a large-scale (105K conversations) media dialog dataset collected from news interview transcripts—which allows us to investigate such patterns at scale. We present a dialog model that leverages external knowledge as well as dialog acts via auxiliary losses and demonstrate that our model quantitatively and qualitatively outperforms strong discourse-agnostic baselines for dialog modeling—generating more specific and topical responses in interview-style conversations.

4.1 Introduction

Much of the news, information, and punditry the general public listens to and reads consists of *media dialog*—a category of open-domain conversations between an interviewer and interviewee centered on world events and situational context. A system for modeling

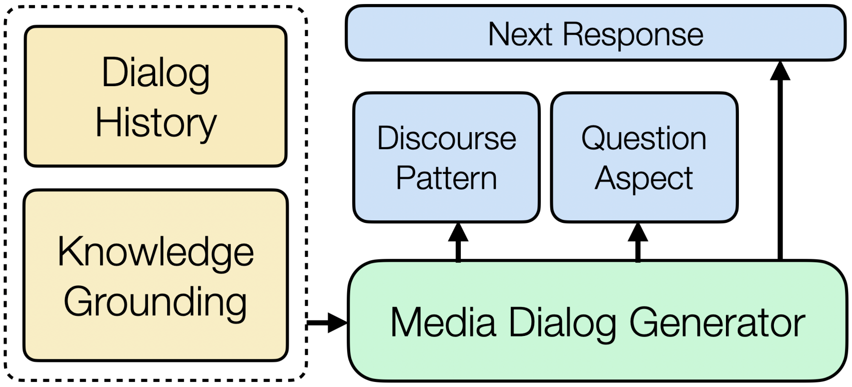
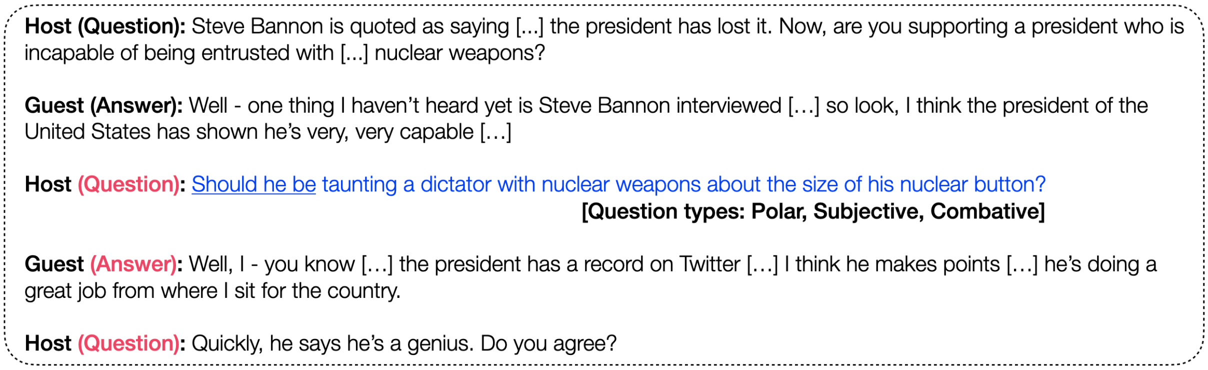


Figure 4.1. Our dialog model incorporates grounding documents alongside dialog history. We also leverage the dialog patterns and interrogative positioning by the host via auxiliary losses.

media dialog from the perspective of one of these roles can help us better understand how media persuades and informs the public [Southwell et al., 2018]. Thus, while recent work in dialog modeling has focused on goal-oriented [Bordes et al., 2017], spontaneous [Shao et al., 2017], or synthetic open-domain chit-chat [Li et al., 2017, Dinan et al., 2019b, Gopalakrishnan et al., 2019b], we aim to analyze discourse patterns in media dialog and their impact on dialog modeling.

Media dialog differs linguistically and in purpose from unstructured, spontaneous conversation such as open-domain chitchat, and both the topical content and interlocutor intent are heavily influenced by the social, cultural, and temporal setting [Weizman, 2008]. The study of media dialog has traditionally focused on individual and manual review of small-scale (<200K word) news corpora [Bednarek, 2006, van Dijk, 2011], and we see an opportunity to scale some forms of discourse analysis to tens of thousands of such documents. In this work, we perform the first large-scale automatic analysis of structural components (response-type patterns) and question type categorization on media dialog, specifically for English news interviews. We show that predicting discourse features can improve generative dialog modeling performance, demonstrating the degree to which discourse structure impacts an interviewer's choice of response type and content. News interviews are also heavily situation-grounded and contextualized by past events and world knowledge. We explore methods to associate each conversation with



Host (Question): Steve Bannon is quoted as saying [...] the president has lost it. Now, are you supporting a president who is incapable of being entrusted with [...] nuclear weapons?

Guest (Answer): Well - one thing I haven't heard yet is Steve Bannon interviewed […] so look, I think the president of the United States has shown he's very, very capable […]

Host (Question): Should he be taunting a dictator with nuclear weapons about the size of his nuclear button?
[Question types: Polar, Subjective, Combative]

Guest (Answer): Well, I - you know […] the president has a record on Twitter […] I think he makes points […] he's doing a great job from where I sit for the country.

Host (Question): Quickly, he says he's a genius. Do you agree?

Figure 4.2. Example conversation from INTERVIEW with annotated discourse analysis. Text highlighted in blue indicates the question of interest, uttered by the host. The dialog triplet is marked in red.

a selection of world facts, and show that by modeling interviewers as *knowledge-grounded* speakers mediating a conversation we are able to generate relevant and specific utterances fitting their role.

Our main contributions in this work are:

1. We collect a dataset of 105K media dialogs (23K two-party dialogs)[1] encompassing two decades of National Public Radio (NPR) radio programs, on which we conduct extensive experiments;

2. We present a probabilistic framework to link a dialog with facts from a large corpus of grounding documents and show that it improves downstream dialog modeling performance compared to a strong TF-IDF baseline;

3. We introduce two auxiliary losses to guide utterance generation in a media dialog setting: look-ahead dialog structure prediction and question-attribute prediction[2]. We show that these losses significantly improve generation quality via automatic and human metrics.

---

[1] https://www.kaggle.com/shuyangli94/interview-npr-media-dialog-transcripts

[2] Code: https://github.com/MEDIA-DIALOG/interview-media-analysis

Table 4.1. Comparative media dialog dataset statistics. *RadioTalk does not contain full conversations

Dataset	Structured	# Dialogs	# Turns	# Words
RadioTalk [Beeferman et al., 2019]	✗	5.98 M*	116 M	2.9 B
TAL [Mao et al., 2020]	✓	663	163,808	7.4 M
INTERVIEW 2P	✓	23,714	454,739	21.7 M
INTERVIEW	✓	105,848	3,199,856	126.7 M

4.2 INTERVIEW: A Media Dialog Corpus

We collect a new dataset of 105K multi-party interview transcripts for 7 programs on National Public Radio (NPR)[3] over 20 years (1999–2019). These transcripts contain in total 3M turns comprising 7.5M sentences (127M words) from 184K speakers, of which 287 are interviewers. To investigate host-mediated media dialog, we curate a subset, **INTERVIEW 2P**, with **two roles**: an *interviewer* and a *guest*, comprising 23K two-party conversations encompassing 455K turns, with 1.24M sentences and 21.7M words. In these two-party conversations, each speaker takes an average of nine turns per dialog. Guests tend to speak longer on their turns, with 1.6x as many sentences spoken and 2x as many words per turn. Meanwhile, hosts ask five times as many questions as guests, with 40% of their dialog turns containing questions. When asking questions, hosts and guests use interrogative forms [See et al., 2019] at the same rate (65%).

4.2.1 Comparison with Other Datasets

Open-domain dialog datasets have traditionally focused on either spontaneous (e.g. telephone calls) or goal-oriented conversation, and there is a paucity of English-language *media dialog* datasets—that is, dialog corpora comprising semi-structured conversations for the purpose of information elicitation and presentation. The closest such datasets are This American Life [Mao et al., 2020], a dataset of several hundred long-form expository podcast episodes, and RadioTalk [Beeferman et al., 2019], which comprises over one million ten-minute snippets of

---

[3] https://www.npr.org/

talk radio transcripts. While these corpora are derived from broadcast media, episodes of the former contain a broad range of expository speakers who are not professional journalists, while the latter dataset is constructed via an automated transcription system with a 13%+ word error rate and does not contain full conversations (segments from radio conversations are transcribed). We compare INTERVIEW statistics to other English media dialog datasets in Table 4.1.

Traditional media dialogs (e.g. news interviews) comprise a significant body of media consumed by the general public and we believe there is value in the large-scale study of such media. Efforts to collect and transcribe broadcast news span the world, from the French EPAC corpus [Estève et al., 2010] to Arabic and Chinese news manually transcribed via the GALE program [Cohen, 2007]. To our knowledge, no attempt has yet been made to analyze the discourse patterns or trends in such data—these datasets have primarily been used to support the development of automatic speech recognition, transcription, and machine translation systems. Early efforts to collect English-language broadcast conversation transcripts [Placeway et al., 1997] similarly aimed to build smaller, high-quality parallel corpora for speech transcription. The large-scale study of discourse in media dialog is not supported in such corpora, and the INTERVIEW corpus enables such analysis at scale for English-language media.

4.3 INTERVIEW Discourse Analysis

We tackle three aspects of discourse analysis that can be scaled to INTERVIEW: 1) Dialog patterns that emerge through new interviews; 2) Large scale annotation of interviewer question types (dialog acts); and 3) Obtaining grounding documents that provide situational context for a news interview. We study these discourse features in context of English broadcast news interviews.

4.3.1 Dialog Patterns

The news interview setting revolves around sets of questions and answers—naively, one may assume the interviewer to be the sole questioner. However, media dialog has steadily deviated from this rigid structure, tending toward the broadly conversational [Fairclough, 1988]. Each participant may be at turns jovial, inquisitive, and critical, and this is reflected in question-answer patterning. Heritage [1985] frames the analysis of media discourse in terms of the *third-turn receipt*, where 1) they ask a question; 2) the interviewee responds; and 3) the interviewer chooses how to proceed. We are motivated by this, as well as studies of *question-response-confirmation* patterns in spontaneous dialog [Van Hekken and Roelofsen, 1982]. We focus on discourse patterns in **response type triplets** beginning with an interviewer (host) question.

We define a triplet as $\{r_1, r_2, r_3\}$ where the response type at utterance i is a question or an answer: $r_i \in \{\mathrm{Q}, \mathrm{A}\}$. By imposing a binary label on each utterance, we are able to efficiently mine all occurrences of each of eight possible host-guest-host patterns across our 23K dialogs. We find that a structured interrogative Q-A-Q pattern comprises 27% of all cases, while 20% of the time the host poses a non-interrogative third response (Q-A-A). Guests respond to questions with questions of their own only 7% of the time, supporting the theory that interviewers serve as the primary *mediators* in such conversations [Weizman, 2008]. Manual inspection evinces recurring action patterns corresponding to interviewer stance-taking and agendas ranging from cooperative to confrontational. For example, the conversation segment in Figure 8.3 is comprised entirely of Q-A-Q patterns, with the host prompting [Heritage, 1985] the guest, re-contextualizing and refocusing the guest's stance for the benefit of the audience. To leverage the inter-dependence of action choice (question or answer) and stance-taking (implicitly or explicitly via utterance content) [Haddington, 2004], we propose to predict the subsequent response type triplet while modeling an interviewer utterance. We thus explore how utterance phrasing and structure may depend on projected or desired conversation directions.

4.3.2 Question Types as Dialog Acts

In their role as a mediator, interviewers can shape the narrative by posing different *types* of questions to guests. Weizman [2008] posits that this choice of question type is influenced by dialog context and conversation flow. We examine ways to structurally bias our model to take advantage of conversational context in order to ask appropriate interviewer questions. Based on common interviewing guides[4] and linguistic analysis of open-ended questions in a conversational setting [Karttunen, 1977], we define three interrogative aspects (attributes): 1) **Polarity**: determining if the question is yes/no (polar) or open-ended; 2) **Subjectivity**: determining if it demands a factual answer or invites a subjective opinion; and 3) **Combativeness**: whether the question is confrontational or clarifying. Our mode of categorization resembles that of Gnisci and Bonaiuto [2003], who add additional categories that are more relevant to the study of equivocation in confrontational interviews. While previous works have primarily used question polarity and interrogative forms to improve diversity in spontaneous dialog generation [Zhao et al., 2017], we explore how a news interviewer constructs question contents given desired interrogative aspects.

We hired two expert annotators to assess a question based on these three aspects. We provided interviewer questions alongside corresponding dialog histories, and annotators marked the binary presence/absence of each aspect for each question. The first host question from Figure 8.3 would be marked as polar, subjective, and combative, as it asks the guest whether (polar) they endorse (subjective) an intentionally ridiculous statement (combative). We collected 1,000 questions in this manner, each labeled by both annotators. The inter-annotator agreement (Cohen's kappa [Cohen, 1960]) for each of the binary labeling tasks—polar vs. open-ended, subjective vs. objective, combative vs. clarifying—was 0.8 for polarity, 0.72 for subjectivity and 0.7 for combativeness. We observed questions in this sample to be 60.2% polar, 38.7% subjective, and 29.5% combative.

Automatic Classification

We label the remainder of INTERVIEW by training a multi-label classifier, fine-tuning BERT [Devlin et al., 2019] to predict the presence of each attribute in our human-annotated set of questions. We concatenate dialog history and the interviewer question separated by a `[SEP]`

---

[4] http://prndg.org/host-interviewing-tips

Table 4.2. F1 Performance of question-type classifier models on the test set.

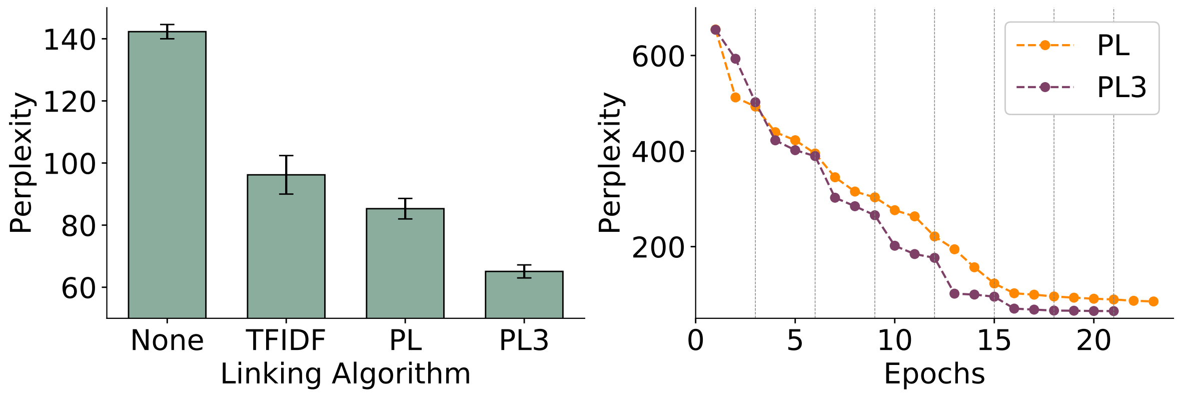
History	Model	Polarity	Combativeness	Subjectivity
No	MLP	55.61	48.91	50.87
	CNN	68.20	57.19	53.91
	LSTM	66.87	49.70	51.96
	BERT	75.31	58.10	66.92
Yes	MLP	68.71	60.81	61.21
	CNN	74.71	65.87	67.98
	LSTM	70.49	60.54	63.09
	BERT	**80.20**	**70.14**	**76.92**

token and prepend a `[CLS]` token. We calculate binary cross entropy loss over a linear projection of the final hidden state of the `[CLS]` token. BERT achieves 80.20, 70.14, and 76.92 F1 scores for polarity, combativeness and subjectivity respectively on the test set in four epochs.

We consider multiple baselines: 1) an MLP model using Bag-of-Words input features; 2) a CNN [Fukushima, 1988] with 2 convolution layers; and 3) a Bi-LSTM [Graves et al., 2005] network with max-pooling of final hidden layers. We initialize all embeddings with BERT embedding vectors. As shown in Table 4.2, BERT achieves the highest F1-score. Including dialog history improves classification performance, confirming that the type of question asked depends on conversational context. This suggests that we may also be able to better predict question content through jointly leveraging the dialog history and question type. Both human annotators and our model find predicting polarity the easiest, and combativeness the most difficult.

4.3.3 Knowledge Grounding

Media dialog is frequently characterized by references to world knowledge, current events, and factual information. This can be learned to some extent in large language models pre-trained on diverse text corpora [Petroni et al., 2019], and such models can act as knowledge stores [Chen et al., 2019]. However, for tasks involving complex reasoning and induction it remains beneficial to provide models with externally linked knowledge [Mitra et al., 2019, Fan et al., 2019]. Specifically for dialog modeling, the Wizard of Wikipedia [Dinan et al., 2019b] and



(a) **(b)**

Figure 4.3. (a) Bar plot depicts test perplexity for linking algorithms: None (no grounding), TF-IDF, and PL/PL3 which indicate probabilistic linking with re-assignment at every 1/3 epochs, respectively. Plotting validation perplexity by epoch shows that PL3 converges faster to the optimum (b).

Topical Chat [Gopalakrishnan et al., 2019b] corpora consist of grounding documents linked with open-domain chit-chat. As such, we explore methods to link *grounding knowledge documents* for each conversation in INTERVIEW, drawn from NPR news articles from the past two decades. We aim to link documents that can best inform conversation content and structure as measured by downstream dialog modeling performance.

TF-IDF Linking

We assess a strong retrieval baseline for grounding document linking, using TF-IDF [Salton and Buckley, 1988] to find relevant documents for each conversation. To support large-scale TF-IDF similarity computation, we use the Lucene-based ElasticSearch [Gormley and Tong, 2015] engine[5] to calculate TF-IDF similarity between full interview texts and the concatenation of the document headline and body, returning the 50 most similar grounding documents for each INTERVIEW conversation. We aim to link documents that would be reasonably relied on by the speakers at the time of the interview, and as such for each interview exclude articles that were published after the interview itself.

---

[5] https://aws.amazon.com/elasticsearch-service/

Algorithm 1. Pseudocode for probabilistic linking

```
Initialize document assignments from TF-IDF priors
while average validation perplexity decreases do
    Learning: Update the model with current assignments for $N$ epochs
    for each $d$ in Dialogs do
        Sample $K$ documents from top 50 TF-IDF priors
        for each $k$ in $K$ do
            Condition each response in the dialog with $k$, and calculate perplexity, aggregate at the dialog level
        end for
        Choose $k$ that yields the lowest perplexity
    end for
    Assignment: Gather all $k$'s for each dialog to update current assignments
end while
```

Probabilistic Linking

While TF-IDF based document linking provides a co-occurence-based similarity measure between documents and conversations, there is no guarantee such linking will improve dialog modeling performance. Thus, we aim to train a linking model such that conditioning on linked documents has a positive effect on dialog modeling performance. We use a two-phase coordinate ascent framework as described in Algorithm 1. In the *Learning* phase, a dialog model is trained based on the available assignments, and its weights are fixed (frozen). Then, in the *Assignment* phase, we compute a re-assignment that maximizes dialog model performance under different possible assignments. Searching over the complete document set is computationally infeasible, so we perform an approximate greedy search over possible documents ordered by their TF-IDF prior score.

We compare the performance of a Transformer [Vaswani et al., 2017a] language model provided with grounding documents assigned by different algorithms in Figure 4.3a. A model without grounding scores by far the worst in terms of perplexity, which indicates that knowledge grounding is important for modeling media dialog. While TF-IDF assignments significantly improve performance compared to no grounding, probabilistic grounding models achieved the best performance. The sudden drops in perplexity values at every third epoch in Figure 4.3b indicates that the model was well-trained based on current assignments before a new assignments

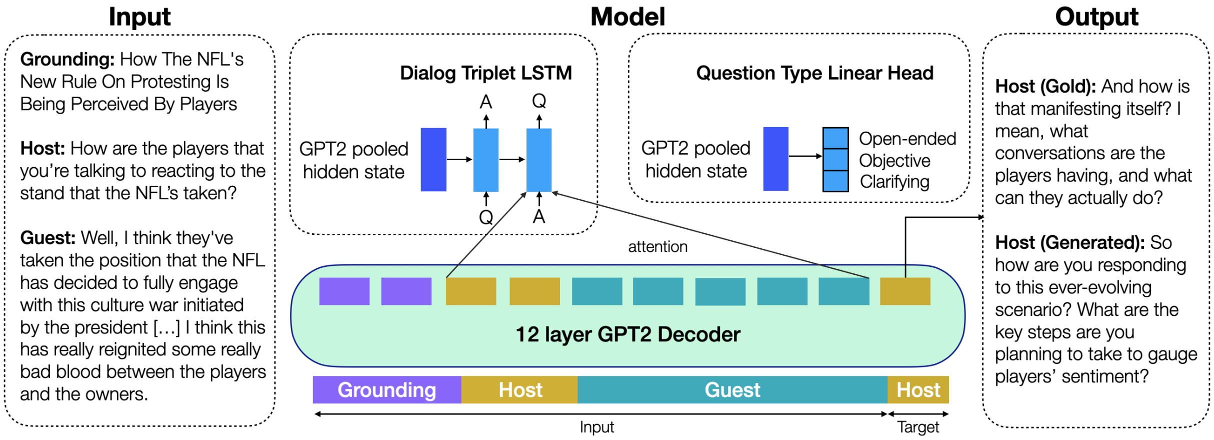

Input
Grounding: How The NFL's New Rule On Protesting Is Being Perceived By Players
Host: How are the players that you're talking to reacting to the stand that the NFL's taken?
Guest: Well, I think they've taken the position that the NFL has decided to fully engage with this culture war initiated by the president […] I think this has really reignited some really bad blood between the players and the owners.
Model
Dialog Triplet LSTM
GPT2 pooled hidden state
A
Q
Q
A
Question Type Linear Head
GPT2 pooled hidden state
Open-ended
Objective
Clarifying
attention
12 layer GPT2 Decoder
Grounding
Host
Guest
Host
Input
Target
Output
Host (Gold): And how is that manifesting itself? I mean, what conversations are the players having, and what can they actually do?
Host (Generated): So how are you responding to this ever-evolving scenario? What are the key steps are you planning to take to gauge players' sentiment?

Figure 4.4. Knowledge grounded generator model with two discourse-specific auxiliary tasks for media dialog

were obtained.

While our articles and conversations come from the same broadcasting source, the NPR interview transcripts generally do not contain links or metadata connecting them with specific grounding documents, and thus there are no ground truth labels available to us. To ascertain that the grounding is relevant, we enlisted two native English speakers who regularly listened to broadcast radio to perform a qualitative evaluation of 100 randomly sampled interview and article pairs. We found that 87% of these pairings are highly relevant, 5% are somewhat relevant and the rest are irrelevant. The inter-annotator agreement measured by Cohen's Kappa was 0.79. The lack of ground truth is something we would argue is not a limitation, rather our probabilistic linking step avoids the dependency on data that is not likely to be available in practice.

4.4 Modeling Media Dialog

A model's ability to learn underlying discourse dynamics is reflected in its performance on downstream tasks. Here, we assess how well our model learns from dialog structure and question-pattern metadata using utterance generation—a simple predictive task that relies on a holistic understanding of grounding knowledge and a dialog history. This serves as an initial measure of understanding of discourse patterns and grounding even if the exact dialog produced

Table 4.3. Performance on auxiliary tasks: Dialog Pattern prediction and Question Type prediction

Model	**Dialog Pattern Pred. Accuracy**	**Question Type Pred. F1**
KGG + Prob. Ground.	38.5	68.8
+ Dialog Pattern	86.3	76.2
+ Question types	**87.9**	**90.5**

can vary.

We treat knowledge-grounded response generation in the media dialog setting as a language modeling task: given a dialog history H and a grounding knowledge document K, we seek to predict the next utterance x by maximizing the likelihood $p(x|H,K)$. The dialog history is composed of turns spoken by both the interviewer and interviewee where each utterance is provided with the role annotation. We only model interviewer (*host*) responses, which aim to moderate the conversation via questions, follow-ups, and acknowledgements. To understand the effect of dialog structure and question types in response modeling, we introduce two *auxiliary losses* to influence generation—a multi-task setup that has seen success in goal-oriented dialog generation [Luan et al., 2017].

4.4.1 Knowledge Grounded Generator

We use a common decoder-only model for knowledge-grounded dialog generation [Gopalakrishnan et al., 2019b]: GPT2 [Radford et al., 2019], a pre-trained Transformer decoder. As model input, we concatenate tokenized grounding documents, dialog history, and the target response. To distinguish each section, we add jointly-learned segment embeddings—{`Grounding`, `Host`, `Guest`}—to each input token. We demonstrate in Table 5.1 that such segment embeddings are essential for this kind of dialog modeling. We only consider target tokens for cross-entropy loss calculation with the conditional likelihood $p(x|H,K)$.

4.4.2 Predicting Look-ahead Dialog Patterns

Following Section 4.3.1, we use a generative model to explore the role of response type triplets in structuring media dialog (stemming from an interviewer utterance [Heritage, 1985]). Following response type triplets defined in Section 4.3.1, we predict the pattern of the dialog triplet beginning with the generated host question as an auxiliary predictive task alongside host utterance generation.

We treat this as a sequence transduction task, employing an LSTM [Hochreiter and Schmidhuber, 1997] decoder with an initial hidden state computed by mean-pooling GPT2 final layer hidden states. Consider $\mathbf{s}_i$ the i-th hidden state from the GPT2 decoder for a length L sequence; now for each hidden state $\mathbf{l}_i$ in the LSTM decoder, we also calculate attention over the GPT2 hidden states, where $\{\mathbf{s}_i\}$ are the keys and values, and $\mathbf{l}_i$ is the query, resulting in an attended vector. We concatenate this attended vector with the LSTM hidden state $\mathbf{l}_i$ and then project it to predict the dialog triplet sequence, maximizing the log-likelihood.

4.4.3 Predicting Question types

We further explore the impact of question types (dialog acts) via another auxiliary task: multi-label classification for host utterance question types [McLeod et al., 2019]. We surmise that accurately predicting question types will help infer question framing and wording, improving generation fidelity. Much like dialog pattern prediction, we use a pooled representation of GPT2 hidden states. We produce a score for each of three question attributes—polarity, combativeness, and subjectivity—via a linear projection and optimize via binary cross-entropy loss.

4.5 Experiments

In our experiments, we seek to answering the following: 1) Does knowledge grounding help generate more topical host responses? 2) Do our two auxiliary discourse losses improve dialog generation performance? 3) Do human raters find responses generated by our model

Table 4.4. Metrics on generated interviewer responses on test set. NPO/NEO = Noun-phrase/Named entity overlap with context (C) and gold (G); QR = Question rate. NIDF is a measure of specificity [See et al., 2019]. QR, NPO, NEO are measured in percentages.

Model	PPL	BLEU	QR	NPOG	NPOC	NEOG	NEOC	NIDF
No Grounding								
Finetuned (FT) GPT2	28.6	15.4	34.2	0.67	0.57	0.92	0.98	0.105
FT GPT2 + Segment	27.5	17.5	49.9	1.70	1.67	1.56	1.55	0.117
Effect of grounding								
MemNet + TF-IDF	26.5	17.8	43.8	1.86	1.63	1.51	1.62	0.187
MemNet + Probabilistic Grounding	25.1	17.7	46.9	1.98	2.31	2.89	3.02	0.197
KGG (TF-IDF)	23.5	18.1	48.5	2.73	3.91	3.01	5.58	0.245
KGG (Probabilistic Grounding)	19.6	19.2	53.6	3.24	4.67	3.44	6.78	0.267
Auxiliary Losses								
+ Dialog Pattern	17.2	**21.0**	56.7	3.52	**6.92**	5.16	**7.85**	0.302
+ Question Types	**15.8**	20.3	**58.9**	**3.67**	6.79	**5.89**	7.79	**0.359**

Table 4.5. Pairwise comparison (in %) between responses generated by our best model (including both discourse analysis auxiliary tasks) vs. responses generated by other baselines as well as the Gold response. The highest entries are **bolded**.

Best Model vs. →	**No Grounding**		**TFIDF**		**Prob. Grounding**		**+ Dialog Pattern**		**Gold**	
Metric ↓	win	loss	win	loss	win	loss	win	loss	win	loss
Relevance to Dialog History	**85.1***	9.2	**86.5***	3.3	**69.1***	27.6	**61.0***	22.4	36.7	**47.4***
Relevance to Grounding	**94.5***	4.3	**91.6***	1.9	**82.5***	4.7	**84.5**	9.5	30.1	**55.2***
Fluency	**97.2***	0.8	**87.1***	7.8	**62.1**	10.1	**58.7**	11.2	20.8	**24.6***

coherent and fluent?

Metrics

To measure the fidelity of generated responses, we compute BPE perplexity and BLEU [Papineni et al., 2002a] between generated and gold utterances. To assess topical accuracy, we calculate the overlap between noun-phrases and named entities in the generated and gold responses. We are also interested in measuring coherence with respect to the context (i.e., grounding documents and dialog history), calculated via the noun-phrase and named entity overlap between generated responses and context. Furthermore, as news interviews are intended to inform audiences, interviewers must ask questions using specific vocabulary and construction. To assess this, we adopt the Normalized Inverse Document Frequency [See et al., 2019] to measure

Table 4.6. Sample generated response on NFL's new rule. When we add discourse specific losses, the models generate questions that bears more coherence to the context as well as ask clarifying questions.

Grounding (PL)	How The NFL's New Rule On Protesting Is Being Perceived By Players
Context	**HOST**: Host: How are the players that you're talking to reacting to the stand that the NFL's taken? **GUEST**: Well, I think they've taken the position that the NFL has decided to fully engage with this culture war initiated by the president [. . .] I think this has really reignited some really bad blood between the players and the owners. **HOST**:
Gold	And how is that manifesting itself? I mean, what conversations are the players having, and what can they actually do?
KGG (TF-IDF)	Can you tell me more about NFL's new rules?
KGG (Probabilistic Linking)	This are some significant changes in NFL's rules. I think the most effect will be on the players.
+ Dialog Pattern	Okay so let's talk about NFL players. What is your gameplan?
+ Question Types	So how are you responding to this ever-evolving scenario? What are the key steps are you planning to take to gauge players' sentiment?

vocabulary specificity via word rarity. Finally since we focus on generating interrogative host responses, we also calculate the percentage of questions asked in the generated responses as a measure of model inquisitiveness.

4.5.1 Effect of Knowledge Grounding

To assess the usefulness of explicit grounding documents, we first compare dialog models that use and do not use such documents in Table 4.4. Using segment embeddings to mark utterance bounds improves all measures of fidelity, signifying that this is a useful way to leverage speaker role information in dialog modeling using GPT2. Models that use external grounding knowledge outperform non-grounded models by 1-8 points on almost all metrics, suggesting that such grounding is an important component of host response generation models. To assess the impact of our knowledge grounded generator (KGG) architecture, we compare performance against a strong Memory Network (MemNet) baseline for knowledge grounded dialog generation

[Dinan et al., 2019b]. We confirm our choice of a GPT2-based KGG, as it outperforms Memory Networks in all quality metrics.

Next, we compare the impact of document assignments made via TF-IDF and our probabilistic linking (PL) method. We once again see improved fidelity, mirroring our observations from Section 4.3.3. Models trained using PL document assignments generate utterances with 19-20% higher noun-phrase and named entity overlap with the gold utterance and context, indicating that PL assignments allow the KGG to more strongly condition on the provided context.

4.5.2 Effect of Auxiliary Tasks

In this experiment, we investigate how predicting dialog patterns and question types impacts the specificity and fidelity of generated host responses. Each auxiliary loss contributes a significant improvement (1-2 points) in perplexity but affects fidelity and topicality in different ways.

With dialog pattern prediction, we observe that generated responses are more coherent with respect to conversational context, seeing 8% and 48% improvements in noun phrase and named entity overlap with dialog history, respectively. This supports the sociolinguistic observation that the interviewer's choice of utterance (i.e., whether to ask a question, and response content) depends on the discourse structure toward which they aim to guide the conversation [Heritage, 1985]. Our results suggest that biasing a dialog model to predict future discourse structure can encourage it to more effectively leverage the past dialog structure (from the conversation history). We confirm in Table 4.3 that this model can predict look-ahead dialog patterns with 86.3% test-set accuracy. In light of findings that vanilla dialog models may not condition well on conversation context [Sankar et al., 2019], our results suggest one possible direction toward improving contextual language modeling for dialog with inherent structure, such as media dialog.

When we add question-type-prediction loss, we see a significant drop in perplexity and improved fidelity. As expected, by inducing our model to predict the question attributes for the target utterance, our model achieves the highest inquisitiveness (58% question rate). It can also accurately predict question types, with 90.5% macro-averaged test set F1 score. Our results suggest that as the model learns to categorize the interviewer response via specific attributes, it simultaneously learns to generate responses with more specific wording. Table 4.6 contains representative generations from our best model as well as other baselines, showing that when we add additional discourse specific losses, our model appropriately captures the interviewer's clarifying intent and conversation direction.

4.5.3 Human Evaluation

Automatic evaluation of dialog generation quality is still unreliable [Liu et al., 2016, Novikova et al., 2017], and thus we provide evaluation by human users. We perform pairwise comparisons between responses generated by our best system and those generated by four strong baselines: the best model with no grounding, KGG with TF-IDF, KGG with PL, and KGG with dialog pattern prediction. We also compare against the gold response. Our human evaluation study measures three aspects of response quality on 100 test examples: 1) How relevant the response is with respect to **dialog history**; 2) How relevant the response is with respect to **grounding documents**; and 3) Whether the generated response is **fluent** English.

We observe in Table 4.5 that human judges prefer responses generated by our best model (with both discourse analysis auxiliary tasks) to baselines by statistically significant margins in almost every case. Entries with * denote significance with $p < 0.05$ from bootstrap tests on 1000 subsets of size 50. This indicates that dialog structure and question types are highly useful for generative modeling in a media dialog setting—specifically news interviews. Human raters also found that despite a significant drop in perplexity when adding the question-type prediction loss, the two versions of discourse-conditioned models had similar fluency, indicating similar language modeling performance. We observe an inter-annotator agreement (Cohen's kappa) of 0.79, 0.92, and 0.73 for relevance to dialog history, grounding documents, and fluency, respectively.

4.6 Conclusion

In this work, we perform the first large-scale analysis of discourse patterns in media dialog, using a new dataset of 23K annotated news interview transcripts: INTERVIEW. Our results mirror findings from linguistic studies of news interviews [Weizman, 2008, Heritage, 1985]. We demonstrate that adding auxiliary tasks for discourse pattern and interrogative type prediction helps model such media dialog. We observe that responses depend heavily on external knowledge, and present a probabilistic framework for linking factual documents with a conversation. While we focus on discourse *pattern* analysis, INTERVIEW also supports analysis of temporal patterns in interviewing, argumentation, and knowledge grounding in long conversations.

Chapter 4, in part, is a reprint of the material as it appears in “Interview: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding” by Bodhisattwa Prasad Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley, which was published in *Empirical Methods in Natural Language Processing*, 2020. The dissertation author was one of the primary investigators and author of this paper.

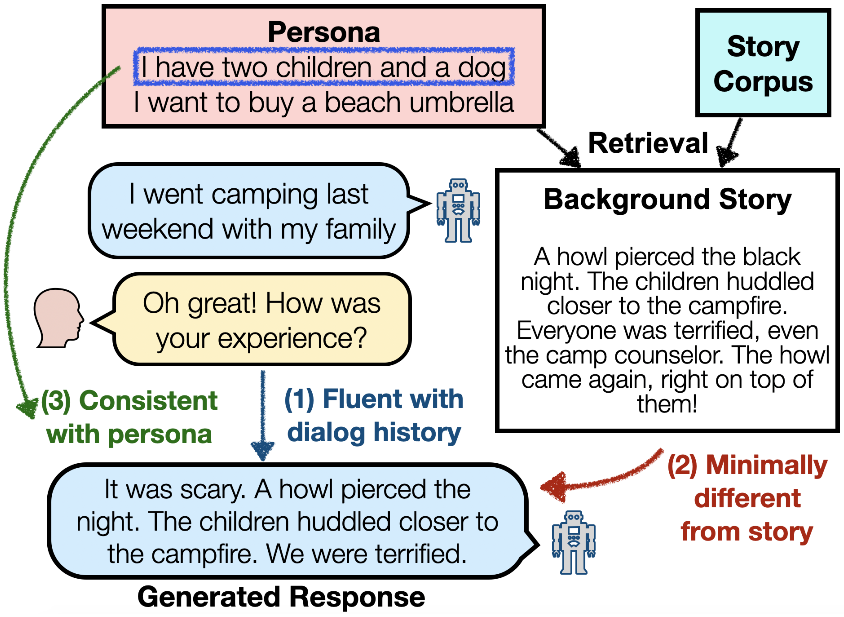
Chapter 5

PABST: Unsupervised Enrichment of Persona-grounded Dialog with Background Stories

Humans often refer to personal narratives, life experiences, and events to make a conversation more engaging and rich. While persona-grounded dialog models are able to generate responses that follow a given persona, they often miss out on stating detailed experiences or events related to a persona, often leaving conversations shallow and dull. In this chapter, we equip dialog models with 'background stories' related to a persona by leveraging fictional narratives from existing story datasets (e.g. ROCStories). Since current dialog datasets do not contain such narratives as responses, we perform an unsupervised adaptation of a retrieved story to generate a dialog response using a gradient-based rewriting technique. Our proposed method encourages the generated response to be *fluent* (i.e., highly likely) with the dialog history, *minimally different* from the retrieved story to preserve event ordering and *consistent* with the original persona. We demonstrate that our method can generate responses that are more diverse and are rated more engaging and human-like by human evaluators compared to outputs from existing dialog models.

5.1 Introduction

Humans often rely on specific incidents and experiences while conversing in social contexts [Dunbar et al., 1997]. Responses from existing chitchat dialog agents often lack such



Persona
I have two children and a dog
I want to buy a beach umbrella
Story Corpus
Retrieval
I went camping last weekend with my family
Background Story
A howl pierced the black night. The children huddled closer to the campfire. Everyone was terrified, even the camp counselor. The howl came again, right on top of them!
Oh great! How was your experience?
(3) Consistent with persona
(1) Fluent with dialog history
(2) Minimally different from story
It was scary. A howl pierced the night. The children huddled closer to the campfire. We were terrified.
Generated Response

Figure 5.1. We enrich agent personas with 'background stories' from an existing corpus. We propose a gradient-based technique which encourages the generated response to be fluent with the dialog history, minimally different from the retrieved story, and consistent with the persona.

specific details. To mitigate this, some prior work has looked into assigning personas to dialog agents [Zhang et al., 2018b, Majumder et al., 2020a]. However, persona descriptions are often shallow and limited in scope, and while they lead to improvements response specificity, they still lack the level of detail with which humans share experiences.

In this work, we propose methods to enrich dialog personas with relevant background events using fictional narratives from existing story datasets such as ROCStories [Mostafazadeh et al., 2016]. For example, for a persona attribute 'I have two children and a dog,' we are able to identify a relevant narrative from a story corpus (Figure 8.1). However, such stories may not directly fit fluently in the dialog context. Thus, retrieved stories should be adapted to construct a response that is fluent and relevant to the context. Since existing datasets (such as PersonaChat [Zhang et al., 2018b]) do not contain responses with such background stories, such adaptation has to be done in an unsupervised fashion with decoders trained to generate responses conditioned only on a dialog history and persona.

To adapt a retrieved narrative incident as a relevant background story, we use a decoding procedure which encourages the generated response to (1) be fluent with the dialog history, (2) be consistent with the original persona, and (3) be minimally different from the retrieved story. While fluency with dialog context is encouraged directly by the likelihood as per the underlying language model the remaining two constraints are incorporated via iterative updates to the decoder output distributions at inference time. Our inference-time decoding method is different from the only recent effort by Su et al. [2020] that leverages non-dialog data (forum comments, book snippets) as distant labels to train dialog systems with supervision. Our contributions can be summarized as follows:

- We propose a novel approach to enrich dialog agent personas with relevant backstories, relying only on existing story datasets.
- We propose to use an unsupervised back-propagation based decoding procedure[1] to adapt the relevant stories such that the resulting response is fluent with the dialog history and consistent with the dialog agent persona. Our method works with a model trained just with dialog data i.e. without access to story corpus at training time.
- Our experiments demonstrate that the proposed approach results in much more engaging and specific dialog outputs in a persona-grounded dialog setup. This fills a gap in existing dialog models which often lack the capability to generate responses about specific events and experiences relevant to persona attributes.

5.2 Unsupervised Persona Enrichment with Background Stories

Given dialog history h and persona C consisting of several (typically 3-5, example shown in Figure 8.1) attributes, our goal is to construct a dialog response x. Our underlying model is based on the discrete persona attribute choice model from Majumder et al. [2020a]. To generate a dialog utterance x, we first sample a persona attribute $c \sim p(c|h)$ conditioned on the dialog

---

[1] Code can be found at https://github.com/majumderb/pabst

history h. x is then generated conditioned on the dialog history and the chosen persona attribute. The underlying dialog model's decoder is initialized with a pretrained GPT-2 model, and is fine-tuned on the PersonaChat dataset [Zhang et al., 2018b]. However, in our current setup, we also have to identify relevant background stories and use them to construct fluent responses at decoding time. Therefore, we propose a different decoding procedure.

To generate a response, we first sample a persona attribute $c \sim p(c|h)$. Next we retrieve stories corresponding to the persona attribute c (Section 5.2.1). However, the underlying dialog model is trained to generate responses conditioned only on the dialog history and persona. To incorporate the retrieved story in the response, we perform gradient-based inference (Section 6.3), that only assumes a left-to-right language model trained on dialog context and responses, and the story is handled at decoding time in an unsupervised fashion. We refer to the proposed method as **RExC** (Unsupervised **P**erson**A** enrichment with **B**ackground **ST**ories).

5.2.1 Retrieving Relevant Stories

For a persona attribute c, we aim to identify relevant stories from a story corpus. Toward this goal, we rank the stories using the F1 component of BERT-score [Zhang et al., 2020] based retrieval using the persona attribute c as the query and the highest scoring story is chosen. Note that many of the stories are written in the third person. For use as background stories, we must first transform them to first–person. Following prior work [Brahman and Chaturvedi, 2020], we identify the protagonist of such stories as the most frequently occurring character. Thereafter, we use co-reference resolution [Lee et al., 2017] to identify all words or phrases that refer to the protagonist. Finally, all words or phrases so identified are replaced with suitable first person pronouns (e.g. 'his books' to 'my books').

5.2.2 Gradient-based Inference

Our underlying dialog model is not trained to condition on a retrieved story, and cannot be directly used to construct a desirable response using s. To tackle this, we consider a decoding strategy which, in addition to fluency with history h, encourages response x to follow two soft constraints: (1) be minimally different from story s, and (2) be consistent with persona c.

First, we generate an initial response based only on the dialog history. Then we perform an iterative procedure which alternates between performing a forward pass on the language model to encourage fluency, and a backward pass which updates the response via back-propagation to respect the two soft constraints. However, x is discrete, and cannot be directly updated using gradients from back-propagation. Instead, we maintain and update a soft representation o of x, where o_i corresponds to the last hidden state representation for the ith token position, i.e., $p(x_i) \sim \text{softmax}(Wo_i/\tau)$, where τ is the temperature parameter, W is the embedding matrix, and $Wo_i \in \mathscr{R}^V$ (V is the vocabulary size). Our approach is inspired by recent works that use gradient-based decoding for text generation with soft constraints [Dathathri et al., 2020, Qin et al., 2020]. Next we describe the backward and forward passes of the iterative procedure.

Backward Pass with Soft Constraints

We define the following soft constraints on response x:

(1) **Divergence from story:** We want to encourage x to be *minimally different* from the story s. Following prior work [Qin et al., 2020], we compute a cross entropy loss (denoted by cross-entr henceforth) with story $s = \{s_1, \ldots, s_T\}$ tokens as labels and $Wo_1, \ldots, Wo_T$ as the logits.

(2) **Consistency to persona:** We want x to be *consistent with persona attribute* c. Consider a classifier $q_\phi(o,c)$ which predicts the probability of x (or rather the soft representation o of x) entailing c. The classifier $q_\phi(o,c)$ is a bag-of-words classification head on decoder hidden states o, fine-tuned on the Dialogue-NLI dataset [Welleck et al., 2019] to predict whether pairs of persona attributes and responses are entailed or not. The objective to maximize can be written as:

$$\mathscr{L}(c,s;o) = \lambda_c \log q_\phi(o,c) - \lambda_d \, \text{cross-entr}(s, Wo)$$

where λ_c and λ_d are hyper-parameters. We update o through back-propagation by computing the gradient $\nabla_o \mathscr{L}(c,s;o)$, while keeping the model parameters constant. Let the resulting o after the

Table 5.1. Diversity metrics on the PersonaChat test set. D-1/2 is the % of distinct uni- and bi-grams. ENTR is the geometric mean of n-gram entropy. Grad. Inf. is the unsupervised gradient-based decoding as opposed to Nucleus sampling [Holtzman et al., 2020].

Method	Training	Decoding	D-1	D-2	ENTR
W/o Story Data					
TRANSFERO	PERSONA-CHAT	Nucleus	0.05	0.11	1.21
COMPAC	PERSONA-CHAT	Nucleus	0.15	0.25	1.25
COMPAC	CS-KB	Nucleus	0.87	1.07	2.04
With Story Data					
COMPAC	PSEUDO	Nucleus	0.91	2.45	2.89
COMPAC	MULTITASK	Nucleus	0.99	2.54	2.71
COMPAC	PERSONA-CHAT	RETRIEVAL	2.56	9.67	3.86
PABST (Ours)	PERSONA-CHAT	Grad. Inf.	1.56	3.57	3.21

gradient-based updates be denoted by o^b.

Forward Pass to Encourage Fluency

Next we perform a forward pass of the underlying dialog model, with the goal of regularizing the hidden states towards the unmodified language model values. On computing the forward pass at the jth token, we mix the final hidden states o_j^f from the forward pass with o_j^b computed in the backward pass, via weighted addition to get the resulting $o_j = \gamma \times o_j^f + (1 - \gamma) \times o_j^b$, where $\gamma \in (0,1)$ is a hyperparameter. The resulting o_j is used for computing the logits at the next time step $j+1$.

We initialize the output response by performing greedy decoding from the underlying dialog model, conditioned on the dialog history and persona attribute. Then we iteratively update o by alternate backward and forward passes. We sample the final response $x \sim \text{softmax}(Wo/\tau)$. In practice, we found that 5 iterations are sufficient to generate good quality outputs.

5.3 Experiments

We evaluate methods in terms of their capability to generate diverse, fluent and engaging responses.

Table 5.2. Pairwise comparison (% win/loss cases, ties not reported) between responses from Pabst and from other baselines as well as gold. All differences are significant with $p < 0.05$. Cohen's Kappa [Cohen, 1960] for sensibility and engagement were 0.79 and 0.82 respectively.

Pabst vs.	**Transfero**		**Compac**		**Retrieval**		**Pseudo**		**MultiTask**		**w/o DNLI**		**Gold**	
Aspect	win	loss	win	loss	win	loss	win	loss	win	loss	win	loss	win	loss
Sensible	**83.8**	5.6	**71.4**	10.1	**72.7**	15.7	**94.3**	3.1	**92.8**	1.5	**81.7**	6.8	29.1	**33.6**
Engaging	**96.7**	1.8	**82.6**	6.7	**87.2**	5.3	**94.7**	2.1	**94.1**	2.6	**70.9**	14.8	**70.2**	15.6

Table 5.3. Generations from different models.

Dialog History
Agent: How was your last trip?
User: It was great. I went to Utah and saw the Canyonlands.

Retrieved Story:
I went on a trip. I went to Florida. When we got there, there was no air conditioning. I was irritated. It was too hot.

Compac: That's great. You must be adventurous.

Retrieval: I went on a trip. I went to Florida. When we got there, there was no air conditioning. I was irritated. It was too hot.

Pseudo: I have a house in Florida.

MultiTask: I like Florida. I went to the beaches.

Pabst ($\lambda_d = 5$): I went on a trip. I went to Florida. I have a house there. We do not have air condition. It was hot.

Pabst ($\lambda_d = 1$): That's great. I frequently go to Florida. I have a house there but it gets hot in summer.

Dialog History
Agent: I work as stunt double. What about you?
User: I'm an engineer. I just got back from work. How's yours?

Retrieved Story:
I was a working actor. For my new role I was a stunt double. I started working on new stunts. When I jumped off a building yesterday I got hurt. I had to be rushed to the hospital.

Compac: That's great. I am looking for a job.

Retrieval: I was a working actor. For my new role I was a stunt double. I started working on new stunts. When I jumped off a building yesterday I got hurt. I had to be rushed to the hospital.

Pseudo: I got a new job. I am an actor.

MultiTask: I started a new role. I got hurt.

Pabst ($\lambda_d = 5$): I was working as a stunt double. I jumped off a building. I got hurt. I had to be rushed to the hospital.

Pabst ($\lambda_d = 1$): Yesterday while I was working, I jumped off a building and I got hurt. I had to be taken to the hospital.

Datasets

We experiment with the PersonaChat dialog dataset [Zhang et al., 2018b] consisting of 131,438 utterances for training, 15,602 for validation, and 15,024 for testing. For stories, we use the training split of the ROCStories dataset [Mostafazadeh et al., 2016], that consists of 78,529 stories, each typically of 4 to 5 sentences.

Baselines

We consider two broad groups of models as baselines: **(1)** ***Without access to story corpus***: We use finetuned GPT2 (**Transfero**) on PersonaChat, and the discrete persona attribute choice model (**Compac**) from Majumder et al. [2020a]. We also consider a version of Compac which enriches personas with inferences from a commonsense knowledge base (**cs-kb**). **(2)** ***Baselines***

***using story corpus*:** To allow COMPAC models to generate story-like responses, we adapt an alternative training regime (**PSEUDO**) from [Su et al., 2020], where we randomly replace some of the target dialog responses with retrieved stories—treating them as pseudo labels. Finally, we also consider a **MULTITASK** training setup from [Su et al., 2020], wherein the decoder is trained on PersonaChat as well as with a language modeling objective on ROCStories. We additionally consider a **RETRIEVAL** baseline that uses the retrieved story verbatim as the dialog response.

5.3.1 Automatic Evaluation

We hypothesize that that the proposed approach to leverage external non-dialog data can increase the diversity of the generated responses. Following prior work [Li et al., 2016a], we report the percentage of distinct uni-grams and bi-grams (**D-1** and **D-2** respectively). Note that these values do not capture the actual frequency distribution of different word types. Therefore, we also report the geometric mean of entropy values of empirical frequency distributions of n-grams of words ($n \in \{1,2,3\}$) [Jhamtani et al., 2018], denoted by **ENTR**.

We observe that methods that use story data show much higher diversity compared to methods that do not (Table 5.1). Among methods using story data, gradient-based decoding (PABST) performs better than COMPAC trained with PSEUDO or MULTITASK. Note that just using RETRIEVAL outputs as-is leads to even more diverse outputs than PABST. However, they are much less sensible with the context, as shown in human evaluations.

5.3.2 Human Evaluation

Since we do not have ground truth story-like responses in the dialog dataset, we perform human evaluation with 150 test examples to investigate if PABST generates responses that are 1) **sensible** with the dialog history and 2) **engaging**. We hired two Anglophone (Lifetime HIT acceptance $\% > 85$) annotators for every test sample. The order of the systems present in the interface is randomized. All differences in values from human evaluations are significant with $p < 0.05$ from bootstrap tests on 1000 subsets of size 50. Cohen's Kappa [Cohen, 1960] to measure inter-annotator agreement for sensibility and engagement were 0.79 and 0.82 respectively.

From the results (shown in Table 5.3), we note that in comparison to responses from baselines, responses from PABST are more engaging and more sensible with respect to the dialog history. We further make following observations. Firstly, using the gradient-based decoding approach with retrieved stories (PABST) works significantly better than using distant supervision with stories data (PSEUDO and MULTITASK). Secondly, background stories provide sufficient detail for an engaging conversation compared to COMPAC which expands persona attributes using commonsense knowledge [Majumder et al., 2020a]. Finally, we also observe that PABST performs worse when we do not use the consistency constraint (w/o DNLI).

Choice of λ_d

We also experiment with different values of the weight for the divergence term (λ_d) in $\mathscr{L}$: High ($\lambda_d = 5$), Moderate ($\lambda_d = 1$), and Low ($\lambda_d = 0.05$). We consider 100 samples for this experiment. We attribute a high λ_d to responses strictly copying the story. We find that PABST (moderate λ_d) wins wins 81.2% and 69.1% cases against PABST (high λ_d) on ‘sensible’ and ‘engaging’ response criteria respectively. Similarly, PABST (moderate λ_d) wins 93.2% and 84.7% cases against PABST (low λ_d) in terms of sensibility and engagement respectively.

Qualitative Analysis

Table 5.3 shows responses generated by different baselines. We observe that PABST is able to follow the retrieved story (same as output from RETRIEVAL) while modifying the response to be conversation-like and sensible with dialog history. Responses from other baselines remain verbose or incoherent. Mirroring the human evaluation, we observe that choosing a higher λ_d makes the model to almost repeat the retrieved story but a lower value smooths the output to make it more sensible with the ongoing dialog.

5.4 Conclusion

We propose a method to enrich persona-grounded dialog with background stories at inference time only using an existing corpus of non-conversational narratives—opening up new ways to generate enriched and engaging responses. One of the limitations of PABST is the assumption of the need of a background story at every turn. As future work, we can include a decision step to decide if we need to incorporate a background story or not, given the dialog history. We can further explore ways to use retrieved stories over multiple turns instead of a single turn.

Chapter 5, in part, is a reprint of the material as it appears in "Unsupervised Enrichment of Persona-grounded Dialog with Background Stories" by Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, Harsh Jhamtani, which was published in *Association for Computational Linguistics, Main*, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 6

POKI: Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection

A limitation of current neural dialog models is that they tend to suffer from a lack of specificity and informativeness in generated responses, primarily due to dependence on training data that covers a limited variety of scenarios and conveys limited knowledge. One way to alleviate this issue is to extract relevant knowledge from external sources at decoding time and incorporate it into the dialog response. In this chapter, we propose a post-hoc knowledge-injection technique where we first retrieve a diverse set of relevant knowledge snippets conditioned on both the dialog history and an initial response from an existing dialog model. We construct multiple candidate responses, individually injecting each retrieved snippet into the initial response using a gradient-based decoding method, and then select the final response with an unsupervised ranking step. Our experiments in goal-oriented and knowledge-grounded dialog settings demonstrate that human annotators judge the outputs from the proposed method to be more engaging and informative compared to responses from prior dialog systems. We further show that knowledge-augmentation promotes success in achieving conversational goals in both experimental settings.

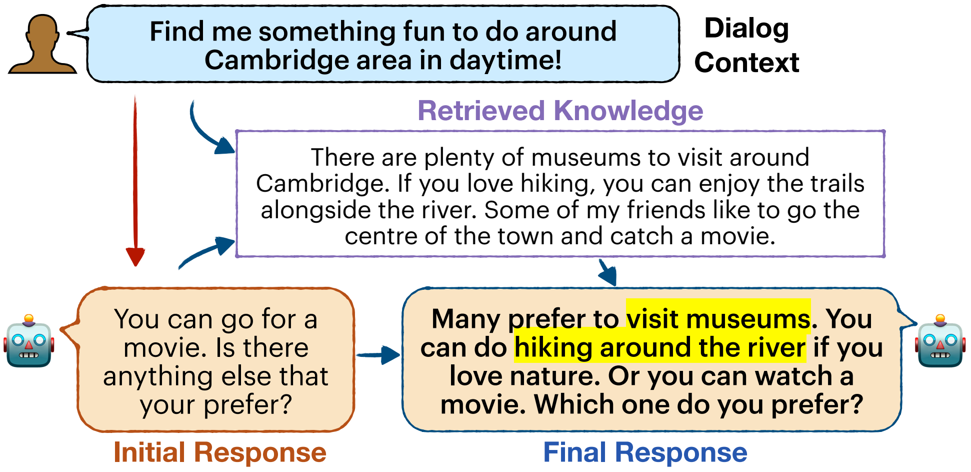

Find me something fun to do around Cambridge area in daytime!
Dialog Context
Retrieved Knowledge
There are plenty of museums to visit around Cambridge. If you love hiking, you can enjoy the trails alongside the river. Some of my friends like to go the centre of the town and catch a movie.
You can go for a movie. Is there anything else that your prefer?
Many prefer to visit museums. You can do hiking around the river if you love nature. Or you can watch a movie. Which one do you prefer?
Initial Response
Final Response

Figure 6.1. Augmenting initial response from an existing dialog model with relevant external knowledge leads to more *engaging* and *informative* responses improving the success in achieving the conversational goal (here, finding a fun activity).

6.1 Introduction

Generic responses which lack specificity have been a major issue in existing dialog models [Hosseini-Asl et al., 2020, Dinan et al., 2019b]. The issue in part stems from bottlenecks in dialog models due to a limited scope of scenarios and access to limited knowledge available during training. On the other hand, encoding all possible world knowledge at training time is not feasible, and even undesirable in cases where knowledge sources are dynamically varying [Ghazvininejad et al., 2018, Majumder et al., 2020b, Zhao et al., 2020, Bruyn et al., 2020, Kim et al., 2020, Prabhumoye et al., 2021]. One possible approach is to incorporate relevant knowledge at decoding-time. For example, in Figure 8.1, the user is seeking options for a fun activity around Cambridge. While the initial dialog response suggests watching a movie as an option, it does not provide any information behind that choice.

We propose and evaluate an approach for unsupervised knowledge injection into a dialog model's response at decoding time[1]—not addressed in any previous work. We first sample a response from the model (trained on dialog data) conditioned on the dialog context. Next, we utilize the dialog context and the sampled response to query external knowledge sources.

---

[1] Code: https://github.com/majumderb/poki

Finally, the retrieved knowledge is used to construct a more informative and engaging response (Figure 8.1). A major advantage of such post-hoc knowledge injection is its flexibility in adding newer knowledge sources especially where the success of achieving conversational goals relies upon the availability of relevant knowledge. Post-hoc injection also promotes efficiency in NLP applications [Schwartz et al., 2020, Strubell et al., 2019]: it mitigates the need to retrain dialog models to accommodate dynamically evolving knowledge.

We experiment with two types of knowledge sources: language models, which we treat as parametric knowledge bases [Petroni et al., 2019, **?**]; and user review datasets such as Yelp reviews [Hajas et al., 2014] as non-parametric knowledge sources (section 6.2). Since it is possible to gather a large amount of related knowledge given a query, we select a relevant and diverse (estimated via information-theoretic measures) subset of knowledge snippets using an unsupervised method (subsection 6.3.1). Then, a gradient-based inference approach is used to construct an updated response that incorporates the selected knowledge (subsection 6.3.2). Note that our framework does not require retraining the existing dialog model—it only relies upon updating the model's output hidden states at decoding time for unsupervised knowledge injection.

We experiment with two scenarios: goal-oriented and knowledge-grounded dialog where the training data covers only a fraction of the needed knowledge. Automatic evaluation reveals that our method is capable of generating highly diverse responses in both settings. In some cases, the generated response shows high overlap with the original target response showing that our unsupervised method bridges the knowledge gap between available knowledge and human-written responses present in the existing dialog corpus. An extensive human evaluation confirms that generated responses are indeed engaging, interesting, and human-like without any loss in fluency.

To pinpoint the usefulness of knowledge injection in the above settings, we design a real-time study (subsection 6.5.3) where users interact with our system to reach a conversational goal (e.g. planning a holiday or knowing more about the solar system). We find that external

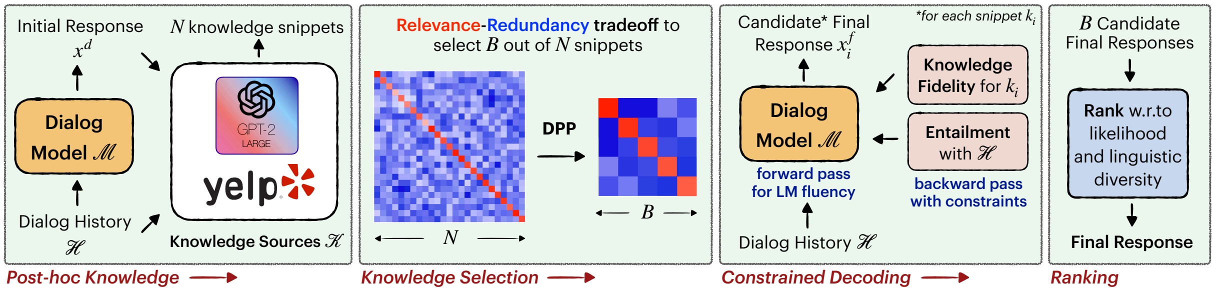

Initial Response x^d
N knowledge snippets
Dialog Model $\mathscr{M}$
GPT-2 LARGE
yelp
Dialog History $\mathscr{H}$
Knowledge Sources $\mathscr{K}$
Relevance-Redundancy tradeoff to select B out of N snippets
DPP
N
B
Candidate* Final Response x_i^f
*for each snippet k_i
Knowledge Fidelity for k_i
Dialog Model $\mathscr{M}$
Entailment with $\mathscr{H}$
forward pass for LM fluency
backward pass with constraints
Dialog History $\mathscr{H}$
B Candidate Final Responses
Rank w.r.to likelihood and linguistic diversity
Final Response
Post-hoc Knowledge
Knowledge Selection
Constrained Decoding
Ranking

Figure 6.2. Pipeline of Poki: It first retrieves **post-hoc knowledge**, then the most relevant and diverse knowledge snippets are **selected** from the retrieved set. Each selected snippet is individually injected through a **constrained decoding** to generate a candidate response. The final response is selected via a **ranking** step.

knowledge enables users to achieve their goals more efficiently. Additionally, we observe that the our approach of sub-selecting relevant but diverse knowledge leads to responses that promote success in achieving conversational goals.

6.2 Post-hoc Knowledge for Dialog

Our goal is to construct a dialog response by injecting knowledge (from external textual sources) at decoding time, without having to retrain the models. Consider a dialog model $\mathscr{M}$ from which we can sample a dialog response x^d given a dialog history $\mathscr{H}$. We shall refer to the response x^d sampled from such a model without any decoding time knowledge injection as the *initial* response.

However, as motivated earlier, samples from such a dialog model often lack detail. To improve such responses, we retrieve and incorporate relevant external knowledge k into the initial response. To achieve our goal, we construct a *query* using both dialog history $\mathscr{H}$ and the initial response x^d, and gather a relevant knowledge candidate k from a knowledge source $\mathscr{K}$. The retrieved snippet can provide useful information to the end-user to achieve the conversational goal (see subsection 6.5.3). We explore both parametric (e.g querying a language model) and non-parametric (e.g. deterministic retrieval using word-overlap) ways to obtain post-hoc knowledge.

6.2.1 Parametric knowledge sources

Pretrained language models (PTLM) are typically trained with a vast amount of text that spans a diverse range of domains. Petroni et al. [2019], **?** showed that such PTLMs can be used as a source of knowledge when queried with suitable textual prompts (e.g. *Seattle is famous for* __). To use PTLMs in our use-case, we construct useful prompts from dialog history and the *initial* response. We assemble simple prompts inspired from various knowledge-seeking situations in dialog [Shwartz et al., 2020] such as `[KP]` *is famous for* __, *Here is what I know about* `[KP]`: __, where `[KP]` is a key-phrase[2] extracted from dialog context. We use `gpt2-large` as the PTLM. For example, a query “Here is what I know about fun things around Cambridge:" results in “*There are plenty of museums to visit around Cambridge. If you love hiking, you can enjoy the trails alongside the river...*" as shown in Figure 8.1. We finally rank each knowledge snippet *k* using the likelihood obtained from the PTLM for a concatenated input of *k* and dialog history and choose the most likely.

6.2.2 Non-parametric knowledge sources

External knowledge in the form of a text corpus can be used as a non-parametric knowledge source available at decoding time. Compared to parametric knowledge sources, such sources do not generate text as knowledge snippets, but offer the advantage of high quality and reliability of human written text. We consider the dialog history and the *initial* response as a query to retrieve relevant knowledge instances from the corpus. Next, we identify the top relevant instances in the given corpus with respect to the constructed query using cosine similarity on TF-IDF based representations [Robertson et al., 1995].

---

[2] It possible that a lack of key-phrases results in no knowledge.

6.3 Unsupervised Knowledge Injection in Generated Dialog

Effectively utilizing the retrieved knowledge snippets to construct an enriched dialog response encompasses two major challenges. Firstly, it is not practical to use potentially hundreds of knowledge snippets obtained from the retrieval step for a single response generation. Thus, we need to find a relevant but diverse subset of the snippets. Secondly, the dialog model $\mathscr{M}$ is trained to condition only on the dialog context, and not on the external knowledge. Hence, to leverage the knowledge snippets, we need a decoding strategy to rewrite the initial response x^d such that the resulting final response x^f should closely follow the knowledge snippet to be injected without a loss in the fluency and consistency. Thus, our method requires no additional training and only assumes a language model trained on dialog context (i.e. $\mathscr{M}$). We refer to our proposed framework (Figure 6.2) as **POKI** (**Po**st-hoc **K**nowledge **I**njection in Generated Dialog).

6.3.1 Relevance-Redundancy Tradeoff for Knowledge Selection

At each turn, we obtain N knowledge snippets from both the parametric and non-parametric sources. We wish to select a subset of B (out of N) relevant but diverse knowledge snippets.

We define relevance score of a snippet k_i with respect to the dialog history H using pointwise mutual information (PMI) as follows:

$$\mathbb{REL}_i = \text{PMI}(k_i, \mathscr{H}) = \log\left(\frac{p(\mathscr{H}|k_i)}{p(\mathscr{H})}\right),$$

Thus, a high PMI score would imply a larger semantic similarity between the snippet k_i and H. To account for redundancy between the snippet pair k_i, k_j we again use the PMI score as follows:

$$\mathbb{RED}_{ij,j>i} = \text{PMI}(k_i, k_j) = \log\left(\frac{p(k_j|k_i)}{p(k_j)}\right).$$

The redundancy score is symmetric i.e. $\mathbb{RED}_{ij} = \mathbb{RED}_{ji}$ as PMI is a symmetric measure.

We estimate probabilities (both conditional and marginal) $p(.)$ in the above equations using GPT2 language model, following past work [Padmakumar and He, 2021]. The PMI measure is often considered better than other n-gram-based overlap metrics to measure the degree of association between two sentences [Kedzie et al., 2018, Padmakumar and He, 2021]. Semantically similar phrases occur in both sentences that can easily be ignored by overlap based metrics.

Selection via Determinantal Point Processes.

To select B knowledge snippets out of N with a relevance-redundancy trade-off, we use a subset selection process named Determinantal Point Process (DPP) [Kulesza and Taskar, 2011]. DPP employs a non-uniform selection that assigns low probability to subsets (here, of knowledge snippets) that are less diverse by modeling the repulsive correlation between independently occurring datapoints (see Figure 6.2).

We build an $N \times N$ kernel matrix $\mathscr{D}$, which is real, symmetric and positive semi-definite. The diagonal entries $\mathscr{D}_{ii}$ are populated by the squared relevance score of the i-th knowledge $\mathbb{REL}_i$ and the off-diagonal entries $\mathscr{D}_{ij}$ are $\beta \times$ squared redundancy scores $\mathbb{RED}_{ij}$. We adjust β in such a way that $\mathscr{D}$ always remains positive semi-definite (more details in [Wilhelm et al., 2018]). To select a subset of B, a DPP assigns a probability of sampling such a subset proportional to the determinant of the submatrix $\mathscr{D}_B$ of $\mathscr{D}$, constructed using the indices of the subsetted items. The DPP probability is geometrically related to the volume of the parallelepiped spanned by the selected knowledge snippets. Diverse knowledge snippets tend to be orthogonal in their space hence span larger volume [Kulesza and Taskar, 2012].

Choosing B-size submatrix from N-size $\mathscr{D}$ is a combinatorial problem and can become prohibitively costly when N is very high. Hence, we use a greedy method [Wilhelm et al., 2018] where we initialize the selection with the most relevant k_i and subsequently select the next k_j that maximizes the determinant of the resultant submatrix.

6.3.2 Gradient-based Constrained Decoding for Knowledge Injection

Upon selecting B knowledge snippets, we want to individually inject each knowledge snippet into x^d to construct a candidate final response x^f at inference time.

Previous works have addressed the problem of unsupervised modification of already-generated text using gradient-based decoding [Dathathri et al., 2020, Qin et al., 2020] that employs an iterative procedure consisting of a forward and a backward pass. The forward pass on the generative model (here, $\mathscr{M}$) encourages fluency of the generated text while the backward pass performs gradient ascent on certain desired constraints. Note that due to the discrete nature of x_d, it is not possible to directly update it via back-propagation. Therefore, we maintain the sequence of hidden representations of each output token as z from the dialog model. Each output token $x^d_{(t)}$ is realized via $p(x^d_{(t)}) \sim \text{softmax}(Wz_{(t)}/\tau)$, where τ is the temperature hyperparameter, W is the output embedding matrix (shared with the input), and $Wz_{(t)} \in \mathscr{R}^V$ (V is the size of the vocabulary).

Constraints. Following Majumder et al. [2021a], we define a **knowledge fidelity** objective that encourages x^f to be minimally different from the knowledge snippet k. We achieve this by minimizing the cross entropy loss (CE) between knowledge tokens $k_{(1)}, \ldots, k_{(T)}$ as labels and $Wz_{(1)}, \ldots, Wz_{(T)}$ as the logits.

We further notice that injected knowledge can influence the generation in such a way that it contradicts with responses uttered during previous turns. Hence, we also want x^f to be entailed with the dialog history $\mathscr{H}$. We build an **entailment** classifier $\theta(z, \mathscr{H})$ that predicts the probability of x^f (ideally, the hidden representation z of x^f) entailing $\mathscr{H}$. The classifier $\theta(z, \mathscr{H})$ is a bag-of-words classification layer with hidden states z from $\mathscr{M}$ and fine-tuned using the DNLI dataset [Welleck et al., 2019] to predict whether the current response is entailed with previous responses or not.

Decoding. In the subsequent forward and backward passes, the hidden representation z is gradually perturbed via gradient ascent on the respective objectives. During backward pass, the objective with constraints is

$$\mathscr{L}(\mathscr{H},k;z) = \alpha \log \theta(z,\mathscr{H}) - \lambda\,\mathrm{CE}(k,Wz)$$

with hyperparameters α and λ. We use back-propagation to update z with the gradient $\nabla_z \mathscr{L}(\mathscr{H},k;z)$ while the parameters of $\mathscr{M}$ remain fixed. The updated latent representations of z after the backward pass are denoted as z^{bw}.

A forward pass with $\mathscr{M}$ is required to regularize the hidden states z toward the original dialog model objective to obtain z^{fw}. Corresponding to the tth token, the hidden states for the $t+1^{\text{th}}$ time step are computed via a weighted addition of backward and forward hidden states, i.e., $z_{(t+1)} = \gamma \times z^{bw}_{(t)} + (1-\gamma) \times z^{fw}_{(t)}$ where $\gamma \in (0,1)$ is a hyperparameter.

During generation, we start by sampling the initial response x^d with greedy decoding from $\mathscr{M}$. The hidden states z (of x^d) are iteratively updated by alternate backward and forward passes. The final response is sampled as $x^f \sim \mathrm{softmax}(Wz/\tau)$. The number of iterations ($=5$) and the γ ($=0.45$) were chosen by maximizing the Z-normalized sum of dialog model perplexity and linguistic diversity (% of distinct bigrams) in a greedy hyperparameter search.

6.3.3 Unsupervised Ranking of Candidate Final Responses

Several previous works often over-generate and use an additional ranking step in order to select the final candidate in unsupervised text generation [Qin et al., 2020, Shwartz et al., 2020, Paranjape and Manning, 2021]. Similarly, here we want to rank the generated candidate final responses according to the diversity of the generated text as well as the conditional likelihood of generation given the dialog history. For diversity, we measure the percentage of distinct bigrams present in the response. For conditional likelihood, we use the pre-trained GPT2 model to obtain the log probability when the dialog history, followed by the generated response, passed as a concatenated input. Since these two scores can have varied scale, we perform Z-normalization on the individual scores and add them to obtain a single score for ranking. The highest ranked candidate response is finally rendered to the user.

6.4 Experimental Setup

6.4.1 Scenarios and Datasets

We experiment with two dialog scenarios: goal-oriented and knowledge grounded. Both setups are knowledge intensive but the training data in such setups often contains only a fraction of the needed knowledge. For the goal-oriented setting, we use the Multi-domain Wizard-of-Oz [Budzianowski et al., 2018] dataset. For knowledge grounded dialog, we use the Wizard-of-Wikipedia [Dinan et al., 2019c] dataset.

Multi-domain Wizard-of-Oz (MultiWOZ) is a multi-domain dialog dataset (we use v2.0 [Hosseini-Asl et al., 2020]) consisting of goal-oriented human-human conversations. The dataset spans seven domains (restaurant, train, attraction, hotel, taxi, hospital, police) and contains 10,438 dialogs with 13.68 average turns. Since, we do not need any training data, we only use an evaluation set (of 7K utterances).

Wizard-of-Wikipedia (WoW) is a knowledge grounded dialog dataset which involves retrieving relevant knowledge from Wikipedia, reading and conditioning on it, and finally generating dialog responses [Dinan et al., 2019c]. The dataset contains 201K utterances from 22K dialogues spanning 1300 diverse topics, from which we use only the test set. The associated Wikipedia knowledge base has 5.4M articles and 93M sentences.

6.4.2 Baselines and Ablations

Baselines for MultiWOZ. For MultiWOZ, we consider several baselines following [Sun et al., 2021] for knowledge injection. First, we use the current state-of-the-art model, SimpleTOD, for goal-oriented dialog [Hosseini-Asl et al., 2020]. Sun et al. [2021] extends SimpleTOD by adding chitchat candidates to dialog histories during training. They also have other variants that either *concatenate* output from SimpleTOD and candidate chitchats (Arranger) or *rewrite* by

Table 6.1. Automatic metrics on the test set of MultiWoZ. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$).

System	Acc	BLEU	BRTSc	D-2	ENTR
KCopy	70.1	4.1	62.3	3.16	2.41
SimpleTOD	70.1	**15.0**	**79.2**	0.56	0.90
SimpleTOD+	69.8	12.1	68.1	0.81	1.11
Arranger	70.2	12.3	68.5	0.93	1.15
Rewriter	70.2	12.1	69.4	1.03	1.45
POKI	**71.1**	13.7	74.5	**3.78**	**2.67**
w/o Entailment	69.9	10.9	67.8	**3.67**	**2.56**
w/o Kw Fidelity	70.0	12.3	71.2	0.95	1.19
Gold	100	100	100	0.78	0.86

Table 6.2. Automatic metrics on the test set of Wizard-of-Wikipedia. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$).

System	BLEU	BRTSc	D-2	ENTR
KCopy	13.4	74.3	**3.64**	3.12
KGuide	16.7	71.5	2.54	2.12
KGround	18.3	72.5	2.87	2.35
BART	**19.8**	73.4	2.97	2.55
RAG	**19.9**	73.1	1.03	1.45
POKI	**19.4**	**76.8**	**3.65**	**3.44**
w/o Entailment	18.1	74.2	3.17	**3.39**
w/o Kw Fidelity	18.8	73.3	2.75	2.54
Gold	100	100	2.98	2.59

combining both output and chitchat snippets (Rewriter). We also have a trivial baseline (KCopy) which appends the retrieved knowledge snippet k from POKI with the initial response x_d.

Baselines for WoW. For WoW, we use two current-best knowledge-grounded models, KGround [Wolf et al., 2019] and BART [Lewis et al., 2020a] that concatenate the associated knowledge snippets (present in WoW) and the dialog history as inputs to generate the response with supervision. KGuide [Zhao et al., 2017] and RAG [Lewis et al., 2020b] have an additional knowledge selection step modeled by a latent variable before response generation similar to knowledge grounded models. We also use the KCopy baseline, as described for MultiWOZ.

Variants of POKI. To investigate the impact of various decoding constraints in POKI, we

Table 6.3. Pairwise comparison (% win/loss cases, tie not reported) between responses from POKI and from other baselines as well as ground truth. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$). κ denotes Cohen's Kappa [Cohen, 1960] between a pair of annotators.

	POKI vs	**SimpleTOD**			**Rewriter**			**w/o Entailment**			**w/o Kw Fidelity**			**Gold**		
	Criteria	win	loss	κ	win	loss	κ	win	loss	κ	win	loss	κ	win	loss	κ
MultiWOZ	Coherent	**93.2**	4.4	0.76	**85.6**	10.2	0.75	**98.7**	0.8	0.72	**77.8**	17.8	0.78	26.2	**34.4**	0.69
	Engaging	**94.3**	4.5	0.78	**89.7**	7.9	0.79	**98.7**	0.6	0.80	**71.5**	20.5	0.80	42.4	37.4	0.78
	Interesting	**92.7**	5.4	0.72	**91.2**	8.3	0.73	**88.6**	8.9	0.68	**98.7**	0.8	0.75	49.7	45.6	0.67
	Humanlike	**85.4**	10.7	0.68	**87.4**	7.3	0.65	**61.9**	30.5	0.71	**81.7**	14.0	0.74	29.7	**37.8**	0.66
		RAG			**BART**			**w/o Entailment**			**w/o Kw Fidelity**			**Gold**		
WoW	Coherent	**95.4**	4.5	0.78	**88.5**	9.6	0.72	**94.3**	3.4	0.68	**83.6**	10.7	0.65	23.8	**25.3**	0.73
	Engaging	**89.3**	7.7	0.72	**87.8**	8.3	0.71	**97.7**	0.8	0.70	**71.5**	25.4	0.69	25.4	26.7	0.73
	Interesting	**96.3**	3.5	0.74	**83.3**	9.9	0.75	**79.8**	17.2	0.70	**93.5**	4.5	0.71	35.9	**37.8**	0.76
	Humanlike	**91.4**	7.1	0.68	**92.4**	6.5	0.66	**84.5**	10.5	0.67	**81.8**	13.5	0.71	42.3	41.9	0.68

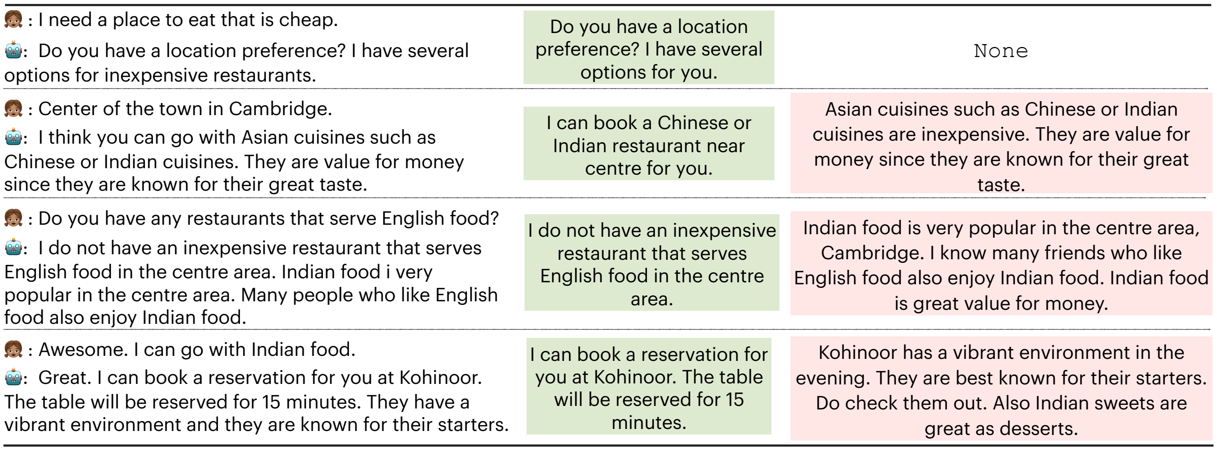
consider the following two variants of POKI—w/o Entailment and w/o Knowledge (Kw) Fidelity (subsection 6.3.2). In POKI, we use SimpleTOD as the base dialog model in goal-oriented scenarios and use BART (which is a state-of-the-art model for WoW) as the base dialog model in the knowledge-grounded scenario. For all variants of POKI, we use gradient-based inference for decoding the final response.

6.5 Results and Discussion

6.5.1 Automatic Evaluation

Our primary goal is to generate responses enriched with relevant external knowledge. Arguably, a system which can effectively leverage additional knowledge at decoding time should generate more diverse responses. We measure percentage of distinct bigrams as Distinct-(D-2) [Li et al., 2016a] and geometric mean of entropy values of empirical frequency distributions of n-grams ($n = 1,2,3$) as Entropy (ENTR) [Jhamtani et al., 2018] for diversity. Additionally, we report overlap between generated responses and corresponding ground truth as per BLEU and BERTScore (BRTSc). For MultiWOZ, we also report the final goal accuracy (Acc) following [Hosseini-Asl et al., 2020].

MultiWOZ. Table 6.1 shows POKI outperforms all the baselines in terms of diversity of



👩🏾 : I need a place to eat that is cheap. 🤖: Do you have a location preference? I have several options for inexpensive restaurants.	Do you have a location preference? I have several options for you.	`None`
👩🏾 : Center of the town in Cambridge. 🤖: I think you can go with Asian cuisines such as Chinese or Indian cuisines. They are value for money since they are known for their great taste.	I can book a Chinese or Indian restaurant near centre for you.	Asian cuisines such as Chinese or Indian cuisines are inexpensive. They are value for money since they are known for their great taste.
👩🏾 : Do you have any restaurants that serve English food? 🤖: I do not have an inexpensive restaurant that serves English food in the centre area. Indian food i very popular in the centre area. Many people who like English food also enjoy Indian food.	I do not have an inexpensive restaurant that serves English food in the centre area.	Indian food is very popular in the centre area, Cambridge. I know many friends who like English food also enjoy Indian food. Indian food is great value for money.
👩🏾 : Awesome. I can go with Indian food. 🤖: Great. I can book a reservation for you at Kohinoor. The table will be reserved for 15 minutes. They have a vibrant environment and they are known for their starters.	I can book a reservation for you at Kohinoor. The table will be reserved for 15 minutes.	Kohinoor has a vibrant environment in the evening. They are best known for their starters. Do check them out. Also Indian sweets are great as desserts.

Figure 6.3. POKI converses with a user who is looking for some restaurant options (left column). In each turn, a knowledge snippet (right column) is injected into an initial response (middle column).

generated responses. More importantly, we see POKI promotes accuracy of reaching the final dialog state i.e. the goal. For ablated versions of POKI, we find the entailment constraint has little effect on diversity while dropping the knowledge adherence constraint negatively influences accuracy and diversity. All variants of SimpleTOD and all versions of POKI show departure from the results obtained by SimpleTOD on BLEU and BERTScore since all of these versions add external knowledge that were not explicitly present in the data. However, we observe that the departure is not significant and POKI achieves a much closer BERTScore to SimpleTOD compared to baselines.

WoW. Despite all systems for WoW use knowledge explicitly in the knowledge-grounded dialog generation task, Table 6.2 shows POKI generates the most diverse responses. Similar to MultiWOZ, the knowledge adherence constraint still remains a significant factor for increasing diversity, one of the main goals of knowledge injection. For WoW, we instead see POKI outperform even BART (previous SOTA) in terms of BERTScore when injected with external knowledge indicating the need of the external knowledge for modeling WoW dialogs.

6.5.2 Human Evaluation

We conduct a comparative human evaluation with 300 samples to evaluate the quality of generated dialog responses following ACUTE-Eval [Li et al., 2019b]. We show a generated response from POKI to an annotator with its associated dialog history to annotate if knowledge injection makes the final response more *engaging*, *interesting* and *humanlike* compared to a baseline response. As sanity check, we also investigate if the response remain *coherent* after knowledge injection. Each sample is evaluated by two annotators.

MultiWOZ. Table 6.3 records the pairwise comparison showing POKI consistently outperforms baselines on all criteria. Responses from POKI are more engaging and interesting compared to SimpleTOD and Rewriter, demonstrating that gradient-based decoding is effective for knowledge injection. In POKI, entailment constraint mostly influences coherence whereas knowledge fidelity constraint is important for engagingness and interestingness.

WoW. Table 6.3 shows POKI outperforms baselines that use grounding knowledge during training in all criteria showing that external knowledge can be useful even in the knowledge-grounded setting to make the conversation engaging and interesting. It also indicates the limitation of the training signal or lack of access to sufficient knowledge and room for improvement in terms of how knowledge is utilized. A large gap in win percentages in favor of POKI for evaluating how 'humanlike' is a response when compared to state-of-the-art methods suggests knowledge injection leads to more natural conversation. Here too, both decoding constraints show similar trends to MultiWOZ.

Qualitative Analysis. Figure 8.3 shows a conversation by POKI with a user who seeks to find restaurant options around Cambridge. We observe that in most of the turns the injected knowledge appeared as an additional justification over the initial responses making the dialog engaging and effective to reach the user's goal (also noted by human judges in subsection 6.5.3). For example, in turn 3, we observe that adding the extra information about Indian cuisine helped user to reach a conclusion when their original choice of English cuisine was absent.

Effect of Response Length. Qualitatively, as seen in Figure 8.3, responses generated by POKI are longer than those from the initial response due to the post-hoc knowledge injection. In the human evaluation sample, we found that 37% of responses from POKI are similar or smaller in length compared to responses from the best baseline. We investigate if response length acted as a confounding factor during human evaluation. Among all the cases where POKI was *lost* over a baseline, 45% (± 2% when bootstrapped with 1000 subsets of size 50) of responses from POKI were longer than those from the comparing baseline. Among *win* cases for POKI, we observe 49% (± 3% when bootstrapped with 1000 subsets of size 50) POKI responses were longer than those from the comparing method. This indicates that human users did not only choose longer responses as better.

6.5.3 User Study for Effectiveness of Knowledge Injection

Relevant knowledge injection has the benefit of adding more justification to terse dialog outputs and hence influencing the task outcome positively. Mirroring observations from [Ghandeharioun et al., 2019], a real-time full conversation evaluation is needed to investigate if POKI could achieve the conversational goal any better than baselines.

We recruited 60 users for this study . One half of the users interacted with POKI, while the other half interacted with the best baseline model that does not augment dialog responses with external knowledge. We construct a *speculative goal* for each user to accomplish via the conversation. We allow users to end the conversation any time they would like and ask them whether the system helped them to reach their conversation goal along with additional comments to justify their annotation. Users who interacted with a knowledge-augmented system also asked if the system provided any knowledge that user has not explicitly asked for but indeed the extra information helped them to reach the conversational goal [Majumder et al., 2021b]. Finally, we also ask if they would like to engage with the system they interacted with in future.

For goal-oriented dialog, we construct speculative goals (e.g. looking for entertainment options) manually from the ground truth for 300 dialog samples. Since we are not using

Table 6.4. Real-time user study with average # of turns for successful goal completion, % of time the goal was achieved, % of success cases users were helped by an *additional* knowledge (Know) that was not explicitly asked to reach their goal, and if users would like to use the system in future.

MultiWOZ	**# turns ↓**	**Goal**	**Know**	**Would use**
Rewriter	8 ± 2	69%	35%	56%
POKI	$\mathbf{4 \pm 3}$	**86%**	**84%**	**76%**
WoW	**# turns ↑**	**Goal**	**Know**	**Would use**
BART	10 ± 2	56%	70%	48%
POKI	$\mathbf{16 \pm 3}$	**76%**	**89%**	**71%**

the underlying databases, we made sure speculative goals do not require specific information (e.g. booking availability, flight information, etc.). For knowledge-grounded dialog, we provide the intended topic of discussion (e.g. science fiction) present in the data; the speculative goal here is to know more about, or to have an engaging conversation about the topic.

Results. First of all, we find that POKI is unanimously preferred by users compared to the baseline during the user study. More importantly, we see that when the user successfully accomplished their goal, 84% of those times they found the additional knowledge helpful in the goal-oriented setting (MultiWOZ) as compared to a baseline (Rewriter) that did not use any external knowledge. Most importantly, POKI takes significantly fewer turns for users to accomplish the goal as compared to Rewriter implicitly indicating injected knowledge (we observe high correlation, 0.67) contributes toward more efficient conversations.

For the knowledge-grounded setting (WoW), both BART and POKI have access to external knowledge sources. However, 89% (compared to 70%) of success scenarios were directly influenced by the additional post-hoc knowledge. For knowledge-grounded dialog, a longer conversation is indicative of engagingness on a particular topic [Gopalakrishnan et al., 2019a], hence users preferred to converse with POKI for more turns as compared to a BART baseline. We quote a comment from a user who found a conversation about the Korean culture with POKI was particularly engaging—"*Before this conversation, I had less knowledge about*

Table 6.5. Evaluation for the quality of the knowledge snippets for random and DPP-based selection.

	Relevant		**Factual**		**BRTSc for WoW**	
Source	**Random**	**DPP**	**Random**	**DPP**	**Random**	**DPP**
Parametric	82%	**89%**	65%	**83%**	74.2	**81.3**
Non-parametric	81%	**83%**	97%	**98%**	65.2	**76.8**

Table 6.6. Mean and std. error of clock-time taken per token

System	**MultiWOZ**	**WoW**
Supervised	17.6 ± 5.2 ms	23.6 ± 4.6 ms
PPCM	30.9 ± 7.5 ms	32.6 ± 4.2 ms
POKI	34.2 ± 8.4 ms	35.7 ± 5.7 ms
POKI, only decoding	31.6 ± 2.7 ms	32.3 ± 3.4 ms

Korean movies and art-forms. This gave me a new perspective and a handful of popular opinions to look at it.".

6.5.4 Discussion

Performance of Knowledge Selection. The knowledge selection step in POKI acts an information bottleneck where the quality of the generated response directly depends on the quality of the selected knowledge We perform a human evaluation on 200 snippets to measure the relevance and the factual correctness in two scenarios: when we randomly select a retrieved snippet or select via DPP. In Table 6.5, we see that the parametric knowledge source (`gpt2-large`) generates more relevant knowledge snippets than a non-parametric one. We attribute this to 1) a large and diverse dataset (webtext) used during pretraining of `gpt2` as compared to yelp reviews (restricted domains) we used for retrieval, and 2) the limited recall of relevant knowledge when using word-overlap based retrieval. However, large language models are still prone to generate non-factual knowledge. We observe that DPP-based selection in POKI is able to sub-select more factual knowledge which then positively influences the final response quality. For WoW, we also compare the selected snippets with the gold knowledge available in the dataset that in turn show high fidelity in terms of BERTScore.

Time Complexity. Madotto et al. [2020] shows that iterative gradient-based decoding could be slower than generating response using single forward pass from an existing model. When we benchmark POKI in an Nvidia 2080Ti GPU, in Table 6.6, we see that knowledge generation (or retrieval) could be a computational bottleneck for POKI. However the greedy selection and the constrained decoding step do not add significant computational load. Furthermore, POKI's performance is comparable with PPCM [Madotto et al., 2020]—a more efficient version of gradient-based decoding. The efficiency of the knowledge retrieval step can be improved with better indexing [Johnson et al., 2021] which we leave as a future work.

6.6 Conclusion

We propose a framework for unsupervised knowledge injection into dialog responses. We show that knowledge can be obtained post-hoc from *any* knowledge sources that can improve users' ability to reach their conversational goal more effectively. In future, our idea can be generalized to setups where external knowledge can justify model's predictions such as conversational recommendation.

Chapter 6, in part, is a reprint of the material as it appears in “Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection” by Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, Julian McAuley, which was published in *Association for Computational Linguistics*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Improving Clarification Question Generation using Global Knowledge

In a conversational user-centric setup it is important to The ability to generate clarification questions i.e., questions that identify useful missing information in a given context, is important in reducing ambiguity. Humans use previous experience with similar contexts to form a global view and compare it to the given context to ascertain what is missing and what is useful in the context. Inspired by this, we propose a model for clarification question generation where we first identify what is missing by taking a difference between the global and the local view and then train a model to identify what is useful and generate a question about it. Our model outperforms several baselines as judged by both automatic metrics and humans.

7.1 Introduction

An important but under-explored aspect of text understanding is the identification of *missing information in a given context* i.e., information that is essential to accomplish an underlying goal but is currently missing from the text. Identifying such missing information can help to reduce ambiguity in a given context which can aid machine learning models in prediction and generation [De Boni and Manandhar, 2003, Stoyanchev et al., 2014]. Rao and Daumé III [2018, 2019] recently proposed the task of clarification question generation as a way to identify such missing information in context. They propose a model for this task which while successful

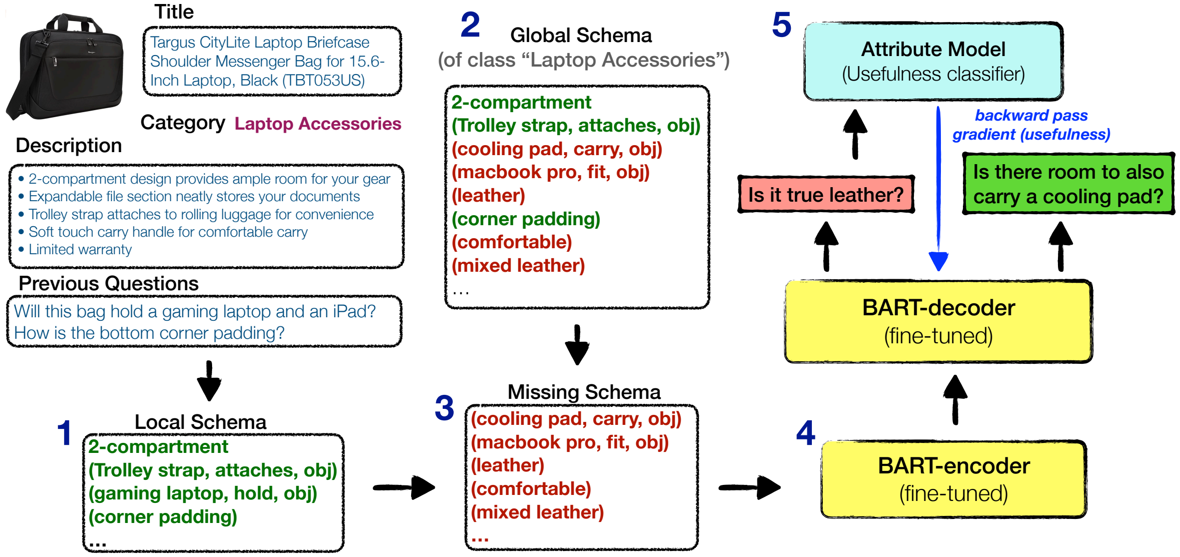

Title
Targus CityLite Laptop Briefcase Shoulder Messenger Bag for 15.6-Inch Laptop, Black (TBT053US)
Category Laptop Accessories
Description
• 2-compartment design provides ample room for your gear
• Expandable file section neatly stores your documents
• Trolley strap attaches to rolling luggage for convenience
• Soft touch carry handle for comfortable carry
• Limited warranty
Previous Questions
Will this bag hold a gaming laptop and an iPad?
How is the bottom corner padding?
1
Local Schema
2-compartment
(Trolley strap, attaches, obj)
(gaming laptop, hold, obj)
(corner padding)
...
2
Global Schema
(of class "Laptop Accessories")
2-compartment
(Trolley strap, attaches, obj)
(cooling pad, carry, obj)
(macbook pro, fit, obj)
(leather)
(corner padding)
(comfortable)
(mixed leather)
...
3
Missing Schema
(cooling pad, carry, obj)
(macbook pro, fit, obj)
(leather)
(comfortable)
(mixed leather)
...
4
BART-encoder
(fine-tuned)
BART-decoder
(fine-tuned)
5
Attribute Model
(Usefulness classifier)
backward pass
gradient (usefulness)
Is it true leather?
Is there room to also carry a cooling pad?

Figure 7.1. Test-time behaviour of our model in a Community-QA setup. 1. We obtain a local schema, 2. the global schema, and 3. estimate the *missing* schema for the clarification question. 4. A BART model generates a question and 5. a PPLM model further tunes it to be more *useful*.

at generating fluent and relevant questions, still falls short in terms of usefulness and identifying missing information. With the advent of large-scale pretrained generative models [Radford et al., 2019, Lewis et al., 2019a, Raffel et al., 2019], generating fluent and coherent text is within reach. However, generating clarification questions requires going beyond fluency and relevance. Doing so requires understanding what is missing, which if included could be useful to the consumer of the information.

Humans are naturally good at identifying missing information in a given context. They possibly make use of *global knowledge* i.e., recollecting previous similar contexts and comparing them to the current one to ascertain what information is *missing* and if added would be the most *useful*. Inspired by this, we propose a two-stage framework for the task of clarification question generation. Our model hinges on the concept of a "schema" which we define as the key pieces of information in a text. In the first stage, we find *what's missing* by taking a difference between the global knowledge's schema and schema of the local context (subsection 7.3.1). In the second stage we feed this missing schema to a fine-tuned BART [Lewis et al., 2019a] model to generate a question which is further made more *useful* using PPLM [Dathathri et al., 2019]

Table 7.1. Product description from amazon.com paired with a clarification question generated by our model.

TITLE:	Sony 18x Optical Zoom 330x Digital Zoom Hi8 Camcorder
DESC:	Sony Hi-8mm Handycam Vision camcorder 330X digital zoom, Nightshot(TM) Infrared 0 lux system, Special Effects, 2.5" SwivelScreen color LCD and 16:9 recording mode, Laserlink connection. Image Stabilization, remote, built in video light.
QUESTION:	Can I manually control the video quality?

(subsection 7.3.2).[1]

We test our proposed model on two scenarios (section 7.2): *community-QA*, where the context is a product description from amazon.com [McAuley and Yang, 2016] (see e.g. Table 7.1); and *dialog* where the context is a dialog history from the Ubuntu Chat forum [Lowe et al., 2015]. We compare our model to several baselines (subsection 7.4.2) and evaluate outputs using both automatic metrics and human evaluation to show that our model significantly outperforms baselines in generating useful questions that identify missing information in a given context (subsection 7.4.4). Furthermore, our analysis reveals reasoning behind generated questions as well as robustness of our model to available contextual information. (section 7.5).

7.2 Problem Setup and Scenarios

Rao and Daumé III [2018] define the task of clarification question generation as: given a context, generate a question that identifies missing information in the context. We consider two scenarios:

Community-QA

Community-driven question-answering has become a common venue for crowdsourcing answers. These forums often have some initial context on which people ask clarification questions. We consider the Amazon question-answer dataset [McAuley and Yang, 2016] where context is a

---

[1] The code is available at https://github.com/microsoft/clarification-qgen-globalinfo

product description and the task is to generate a clarification question that helps a potential buyer better understand the product.

Goal Oriented Dialog

With the advent of high quality speech recognition and text generation systems, we are increasingly using dialog as a mode to interact with devices [Clark et al., 2019]. However, these dialog systems still struggle when faced with ambiguity and could greatly benefit from having the ability to ask clarification questions. We explore such a goal-oriented dialog scenario using the Ubuntu Dialog Corpus [Lowe et al., 2015] consisting of dialogs between a person facing a technical issue and another person helping them resolve the issue. Given a context i.e. a dialog history, the task is to generate a clarification question that would aid the resolution of the technical issue.

7.3 Approach

Figure 7.1 depicts our approach at a high level. We propose a two-stage approach for the task of clarification question generation. In the first stage, we identify the missing information in a given context. For this, we first group together all similar contexts in our data[2] to form the *global schema* for each high-level class. Next, we extract the schema of the given context to form the *local schema*. Finally, we take a difference between the local schema and the global schema (of the class to which the context belongs) to identify the missing schema for the given context. In the second stage, we train a model to generate a question about the most useful information in the missing schema. For this, we fine-tune a BART model [Lewis et al., 2019a] on (missing schema, question) pairs and at test time, we use PPLM [Dathathri et al., 2019] with a usefulness classifier as the attribute model to generate a useful question about missing information.

---

[2] See subsection 7.4.1 for details to combine data splits

7.3.1 Identifying Missing Information

Schema Definition Motivated by [Khashabi et al., 2017] who use essential terms from a question to improve performance of a Question-Answering system, we see the need of identifying important elements in a context to ask a better question. We define schema of sentence s as set consisting of one or more triples of the form (key-phrase, verb, relation) and/or one or more key-phrases.

$$
\begin{gathered}
schema_s = \{\, element \,\}; \text{ where} \\
element \in \{(key\text{-}phrase, verb, relation), \\
key\text{-}phrase\}
\end{gathered}
$$

Schema Extraction Our goal is to extract a schema from a given context. We consider (key-phrase, action verb, relation) as the basic element of our schema. Such triples have been found to be representative of key information in previous work [Vedula et al., 2019]. Given a sentence from the context, we first extract bigram and unigram key-phrases using YAKE (Yet-Another-Keyword-Extractor) [Campos et al., 2020] and retain only those that contain at least a noun. We then obtain the dependency parse tree [Qi et al., 2020b] of the sentence and map the key-phrases to tree nodes.[3] Now, to obtain the required triple, we need to associate a verb and a relation to each key-phrase. This procedure is described in 2. At a high-level, we use the path between the key-phrase and the closest verb in the dependency tree to establish a relation between the key-phrase and the verb. In cases where there is no path, we use only the key-phrase as our schema element. Figure 7.2 shows an example dependency tree for a sentence.

Creating local schema Given a context, we extract a schema for each sentence in the context. The local schema of a context c is a union of schemata of each sentence s in the context.

$$local\_schema_c = \cup_{s \in c}\, schema_s$$

---

[3] In the case of bigram phrases, we merge the tree nodes.

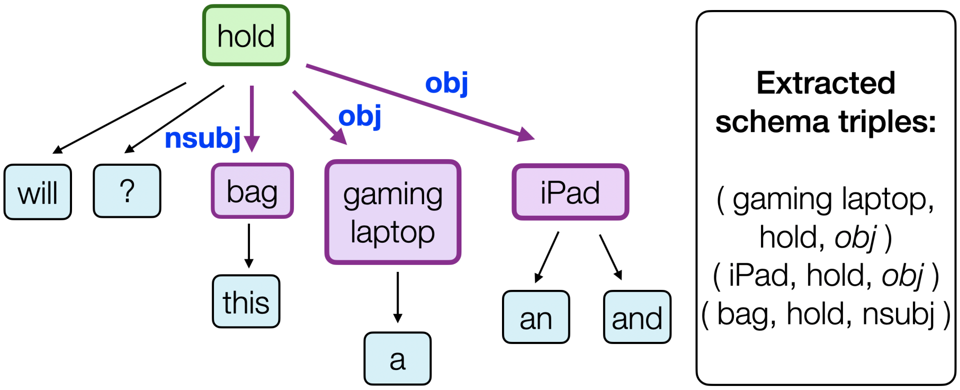


Figure 7.2. Dependency tree and paths showing how we obtain schema triples for a sentence: *"Will this bag hold a gaming laptop and an iPad?"* (from Figure 7.1).

Algorithm 2. Pseudocode for extracting (key-phrase, verb, relation) triple.

```
Initialize with empty path (path length $\infty$) for all possible pairs of verbs ($\in$ {VB, VBG, VBZ}) and key-phrases in the sentence
for Each verb and key-phrase pair do
    Search for the key-phrase among the children of the verb in the dependency tree
    if A key-phrase is found and path is shorter than the stored path then
        Update the path between the key-phrase and the verb pair
    end if
end for
for Each verb and key-phrase pair do
    if The key-phrase is the immediate child of the verb then
        Create the triple (key-phrase, verb, relation) using the relation in the path
    else
        Traverse backward from the key-phrase, stop at the first verb, use the relation with its immediate child in the path to create (key-phrase, verb, relation)
    end if
end for
```

Creating global schema We define global schema at the class level where a 'class' is a group of similar contexts. For *Amazon*, classes consist of groups of similar products and for *Ubuntu*, classes consist of groups of similar dialogs (see subsection 7.4.1 for details). The global schema of a class K is a union of local schemata of all contexts c belonging to K.

$$global\_schema_K = \cup_{c \in K} \, local\_schema_c$$

A naive union of all local schemata can result in a global schema that has a long tail of low-frequency schema elements. Moreover, it may have redundancy where schema elements with similar meaning are expressed differently (e.g. *OS* and *operating system*). We therefore use word embedding based similarity to group together similar key-phrases and retain only the most frequent elements.

Creating a missing schema Given a context c, we first determine the class K to which the context belongs. We then compute its missing schema by taking the set difference between the global schema of class K and the local schema of the context c:

$$missing\_schema_c = global_K \setminus local_c$$

More specifically, we start with the elements in the global schema and remove elements that have a semantic match with any element in the local schema to obtain the missing schema.

7.3.2 Generating Useful Questions

Our goal is to generate a useful question about missing information. In subsection 7.3.1, we explained how we compute the missing schema for a given context; here we describe how we train a model to generate a useful question given the missing schema.

BART-based generation model

Our generation model is based on the BART [Lewis et al., 2019a] encoder-decoder model, which is also a state-of-the-art model in various generation tasks including dialog generation and summarization. We start with the pretrained base BART model consisting of a six layer encoder and six layer decoder. We fine-tune this model on our data where the inputs are the missing schema and the output is the question. The elements of the missing schema in the input are separated by a special `[SEP]` token. Since the elements in our input do not have any order, we use the same positional encoding for all input positions. We use a token type embedding layer with three types of tokens: key-phrases, verbs, and relations.

PPLM-based decoder

We observed during our human evaluation[4] that a BART model fine-tuned in this manner, in spite of generating questions that ask about missing information, does not always generate *useful* questions. We therefore propose to integrate the usefulness criteria into our generation model. We use the Plug-and-Play-Language-Model (PPLM) [Dathathri et al., 2019] during decoding (at test time). The attribute model of the PPLM in our case is a usefulness classifier trained on bags-of-words of questions. In order to train such a classifier, we need usefulness annotations on a set of questions. For the Amazon dataset, we collect usefulness scores (0 or 1) on 5000 questions using human annotation whereas for the Ubuntu dataset we assume positive labels for (true context, question) pairs and negative labels for (random context, question) pairs and use 5000 such pairs to train the usefulness classifier.

7.4 Experiments

We aim to answer the following research questions (RQ):

1. Is the model that uses missing schema better at identifying missing information compared to models that use the context directly to generate questions?
2. Do large-scale pretrained models help generate better questions?
3. Does the PPLM-based decoder help increase the usefulness of the generated questions?

7.4.1 Datasets

Amazon The Amazon review dataset [McAuley et al., 2015] consists of descriptions of products on amazon.com and the Amazon question-answering dataset [McAuley and Yang, 2016] consists of questions (and answers) asked about products. Given a product description and *N* questions asked about the product, we create *N* instances of (*context*, *question*) pairs where *context* consists of the description and previously asked questions (if any). We use the "Electronics" category consisting of 23,686 products. We split this into train, validation and

---

[4] See results of BART+missinfo in Table 7.5

Table 7.2. Number of data instances in the train, validation and test splits of Amazon and Ubuntu datasets (Both datasets are in English.)

Dataset	**Train**	**Validation**	**Test**
Amazon	123,567	4,525	2,361
Ubuntu	102,678	7,864	200

test sets (Table 7.2). The references for each context are all the questions (average=6) asked about the product. A class is defined as a group of products within a subcategory (e.g. DSLR Camera) as defined in the dataset. We restrict a class to have at most 400 products, and a bigger subcategory is broken into lower-level subcategories (based on the product hierarchy) resulting in 203 classes. While creating global schema, we exclude target questions from validation and test examples. The product descriptions and associated metadata come as inputs during test time. Hence, including them from all splits while creating the global schema does not expose the test and validation targets to the model during training.

Ubuntu The Ubuntu dialog corpus [Lowe et al., 2015] consists of utterances of dialog between two users on the Ubuntu chat forum. Given a dialog, we identify utterances that end with a question mark. We then create data instances of (context, question) where the question is the utterance ending with a question mark and the context consists of all utterances before the question. We consider only those contexts that have at least five utterances and at most ten utterances. Table 7.2 shows the number of data instances in the train, validation and test splits. Unlike the Amazon dataset, each context has only one reference question. A class is defined as a group of dialogs that address similar topics. Since such class information is not present in the dataset, we use *k*-means to cluster dialogs into subsequent classes. Each dialog was represented using a TF-IDF vector. After tuning the number of clusters based on sum of squared distances of dialogs to their closest cluster center, we obtain 26 classes. We follow a similar scheme as with Amazon for not including target questions from validation and test sets while building the global schema.

7.4.2 Baselines and Ablations

Retrieval We retrieve the question from the train set whose schema overlaps most with the missing schema of the given context.

GAN-Utility The state-of-the-art model for the task of clarification question generation [Rao and Daumé III, 2019] trained on (context, question, answer) triples.

Transformer A transformer [Vaswani et al., 2017b][5] model trained on (context, question) pairs.

BART We finetune a BART model [Lewis et al., 2019a] on (context, question) pairs.

BART + missinfo We compare to a BART model fine-tuned on (missing schema, question) pairs.

BART + missinfo + WD This is similar to the "BART + missinfo" baseline with the modification that, at test time only, we use a weighted-decoding (WD) strategy [Ghazvininejad et al., 2017] by redefining the probability of words in the vocabulary using usefulness criteria .

BART + missinfo + PPLM This is our proposed model as described in section 7.3 where we fine-tune the BART model on (missing schema, question) pairs and use a usefulness classifier based PPLM model for decoding at test time.

7.4.3 Evaluation Metrics

Automatic Metrics

BLEU-4 [Papineni et al., 2002b] evaluates 4-gram precision between model generation and references. at the corpus level; **METEOR** [Banerjee and Lavie, 2005b] additionally uses stem and synonym matches for similarity; and **Distinct-2** [Li et al., 2016b] measures diversity by calculating the number of distinct bigrams in model generations scaled by the total number of generated tokens.

---

[5] We use original hyperparameters & tokenization scheme.

Table 7.3. Automatic metric results on the full test set of Amazon. The difference between bold and non-bold numbers is statistically significant with $p < 0.001$.

Model	BLEU-4	METEOR	Distinct-2
Retrieval	8.76	9.23	**0.92**
GAN-Utility	14.23	16.82	0.79
Transformer	12.89	14.56	0.60
BART	15.98	16.78	0.78
+ missinfo	16.87	17.11	0.82
+ missinfo + WD	16.23	**17.98**	**0.84**
+ missinfo + PPLM	**18.55**	**18.01**	**0.86**
Reference	–	–	0.95

Table 7.4. Automatic metric results the full test set of Ubuntu. The difference between bold and non-bold numbers is statistically significant with $p < 0.001$.

Model	BLEU-4	METEOR	Distinct-2
Retrieval	4.89	5.12	**0.82**
Transformer	6.89	7.45	0.67
BART	8.23	9.67	0.72
+ missinfo	**9.54**	10.78	**0.75**
+ missinfo + PPLM	**10.02**	**11.65**	**0.79**
Reference	–	–	0.87

Human Judgment

Similar to Rao and Daumé III [2019], we conduct a human evaluation on Amazon Mechanical Turk to evaluate model generation on the four criteria below. Each generated output is shown with the context and is evaluated by three annotators.

Relevance We ask "*Is the question relevant to the context?*" and let annotators choose between Yes (1) and No (0).

Fluency We ask "*Is the question grammatically well-formed i.e. a fluent English sentence?*" and let annotators choose between Yes (1) and No (0).

Missing Information We ask "*Does the question ask for new information currently not included in the context?*" and let annotators choose between Yes (1) and No (0).

Usefulness We perform a comparative study where we show annotators two model-generated questions (in a random order) along with the context. For Amazon, we ask "*Choose which of the two questions is more useful to a potential buyer of the product*". For Ubuntu, we ask "*Choose which of the two questions is more useful to the other person in the dialog*".

7.4.4 Experimental Results

Automatic Metric Results

Amazon Table 7.3 shows automatic metric results on Amazon. Under BLEU-4 and METEOR, the retrieval model performs the worst suggesting that picking a random question that matches the most with the missing schema does not always yield a good question. This strengthens the need of the second stage of our proposed model i.e. BART + PPLM based learning. GAN-Utility, which is state-of-the-art on Amazon, outperforms the Transformer baseline suggesting that training a larger model (in terms of the number of parameters) does not always yield better questions. BART, on the other hand, outperforms GAN-Utility suggesting the benefit of large-scale pretraining (RQ2). BART+missinfo further outperforms BART showing the value in training on missing schemata instead of training directly on the context (RQ1). A variation of this model that uses weighted decoding performs marginally better on METEOR but slightly worse of BLEU-4. Our final proposed model i.e., BART+missinfo+PPLM performs the best among all baselines across both BLEU-4 and METEOR.

Under diversity (Distinct-2), the retrieval model produces the most diverse questions (as also observed by Rao and Daumé III [2019]) since it selects among human written questions which tend to be more diverse compared to model generated ones. Among other baselines, transformer interestingly has the lowest diversity whereas GAN-Utility and BART come very close to each other. Model ablations that use missing schema produce more diverse questions further strengthening the importance of training on missing schema. Our model i.e., BART+missinfo+PPLM, in spite of outperforming all baselines (except retrieval), is still far from reference questions in terms of diversity, suggesting room for improvement.

Table 7.5. Human judgment results (0-1) on 300 randomly sampled descriptions from the Amazon test set

Model	Relevancy	Fluency	MissInfo
GAN-Utility	0.9	0.86	0.81
BART	0.94	0.92	0.77
+ missinfo	0.97	0.92	0.87
+ missinfo + PPLM	**0.99**	**0.93**	**0.89**
Reference	0.96	0.83	0.89

Ubuntu

Table 7.4 shows the results of automatic metrics on Ubuntu.[6] The overall BLEU-4 and METEOR scores are much lower compared to Amazon since Ubuntu has only one reference per context. Under BLEU-4 and METEOR scores, similar to Amazon, we find that the retrieval baseline has the lowest scores. Transformer baseline outperforms the retrieval baseline but lags behind BART, again showing the importance of large-scale pretraining. The difference between the BLEU-4 scores of BART+missinfo and our final proposed model is not significant but their METEOR score difference is significant suggesting that our model produces questions that may be lexically different from references but have more semantic overlap with the reference set. Under Distinct-2 scores, we find the same trend as in Amazon, with the retrieval model being the most diverse and our final model outperforming all other baselines.

Human Judgement Results

Amazon

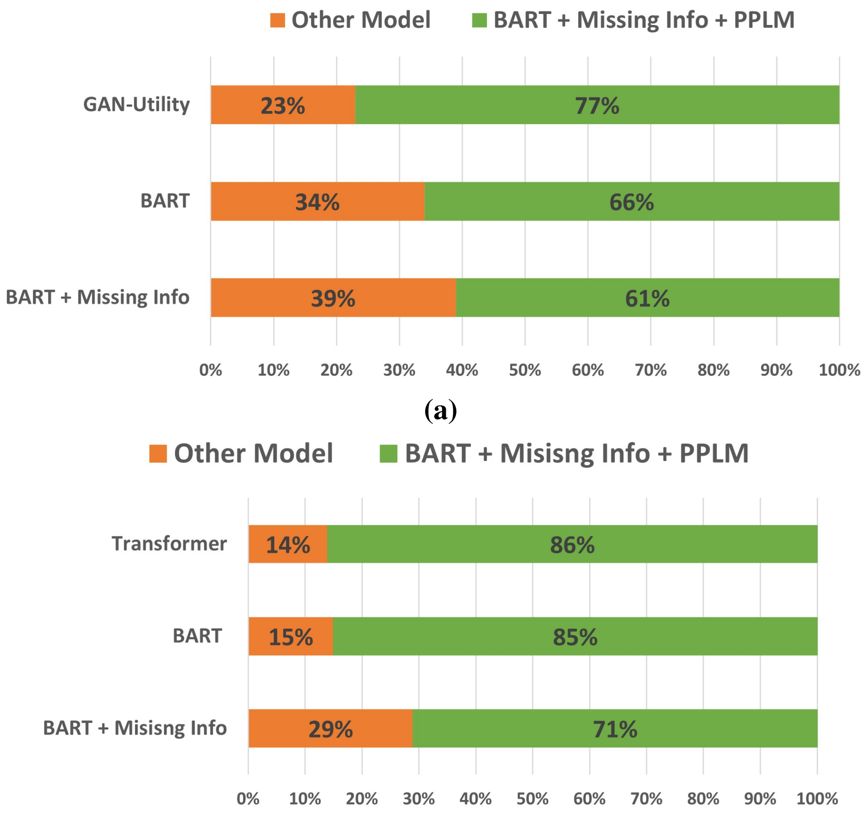
Table 7.5 shows the human judgment results on model generations for 300 randomly sampled product descriptions from the Amazon test set. Under relevancy and fluency, all models score reasonably with our proposed model producing the most relevant and fluent questions. Under missing information, the BART model, fine-tuned on context instead of missing schema, has the lowest score. GAN-Utility outperforms BART but significantly lags behind

---

[6] We do not experiment with the GAN-Utility model (since it requires “answers”) and the BART+missinfo+WD model (since usefulness labels are not obtained from humans).

Table 7.6. Human judgment results (0-1) on 150 randomly sampled dialog contexts from Ubuntu test set

Model	Relevancy	Fluency	MissInfo
Transformer	0.74	0.99	0.99
BART	0.69	0.99	0.96
+ missinfo	0.81	0.95	0.98
+ missinfo + PPLM	0.91	0.83	0.99
Reference	0.85	0.83	0.96



(a)

(b)

Figure 7.3. Results of a pairwise comparison (on usefulness criteria) between our model and baseline generated question on (a) 300 randomly sampled product descriptions from the Amazon test set, (b) 150 randomly sampled dialogs from the Ubuntu test set as judged by humans.

BART+missinfo and BART+missinfo+PPLM reaffirming our finding from the automatic metric results that our idea of feeding missing schema to a learning model helps.

We additionally observe that the human-written questions score lower than model-generated questions under ‘fluency’ and ‘missing information’ criteria, mirroring similar observations from [Rao and Daumé III, 2018, 2019]. We believe the reason for this is that human-written questions often have typos or are written by non-native speakers (leading to lower fluency). Moreover, humans may miss out on reading full product descriptions causing them to ask about details that are already included in the description (leading to lower missing information scores).

Figure 7.3a shows the results of pairwise comparison on the usefulness criteria. We find that our model wins over GAN-Utility by a significant margin with humans preferring our model-generated questions 77% of the time. Our model also beats BART-baseline 66% of the time further affirming the importance of using missing schema. Finally, our model beats BART+missinfo model 61% of the time suggesting that the PPLM-based decoder that uses usefulness classifier is able to produce much more *useful* questions (RQ3).

Ubuntu

Table 7.6 shows the results of human judgments on the model generations of 150 randomly sampled dialog contexts from the Ubuntu test set. In terms of relevance, we find that the transformer and BART baselines produce less relevant questions. With the addition of missing schema (i.e., BART+missinfo), the questions become more relevant and our proposed model obtains the highest relevance score. The reference obtains slightly a lower relevance score which can possibly be explained by the fact that humans sometimes digress from the topic. Under fluency, interestingly, the transformer and BART baselines obtain high scores. With the addition of missing schema, fluency decreases and the score reduce further with the PPLM model. We suspect that the usefulness classifier trained with a negative sampling strategy (as opposed to human labelled data, as in Amazon) contributes to fluency issues. Under missing information, all models perform well which can be explained by the fact that in Ubuntu, the scope of missing information is much larger (since dialog is much more open-ended) than in Amazon.

Figure 7.3b shows the results of pairwise comparison on usefulness criteria. We find that humans choose our model-generated questions 85% of time when compared to either transformer or BART generated questions. When compared to BART+missinfo, our model is selected 71%

Table 7.7. Model generations for an example product from Amazon and an example dialog context from Ubuntu.

Amazon	
Category	Binoculars & Scopes
Title	Nikon 7239 Action 7x50 EX Extreme All-Terain Binocular
Description	The Monarch ATB 42mm with dielectric high-reflective Multilayer Prism coating binocular features brighter, sharper colors, crisp and drastically improved low-light performance. A new body style provides unparalleled strength and ruggedness in a package ...
Missing Schema	{mounting, **center focused**, (Nikon, works, obj), (Canon, works, obj), digital camera, . . . }
GAN-Utility	price?
BART	How is the focus quality?
BART+missinfo	Is it **center focused**?
BART+missinfo+PPLM	Is it **center focused**, or do you have to focus each eye individually?
Ubuntu	
Dialog history	User A: I'm having trouble installing nvidia drivers for my geforce 6200, could anyone perhaps assist? User B: i use the drivers from the website, much better User A: which drivers? from the website? User B: I used add/remove software from the menu to install nvidia proprietary drivers
Missing schema	{(driver, update, nsubj), (new version, install, nsubj), (machine, **reboot**, nsubj), ...}
Transformer	Did you try booting your machine?
BART	where did you download them from?
BART+missinfo	Can you tell the output after you install them?
BART+missinfo+PPLM	Can you try **rebooting** from the start and removing the software after installation?

of the time, further affirming the importance of using the PPLM-based decoder.

Table 7.8. Missing information overlap (in %) between missing schema and output generations

Model	Amazon	Ubuntu
Retrieval	10.5	6.78
GAN-Utility	73.4	–
Transformer	57.2	45.7
BART	60.3	56.9
+ missinfo	97.3	89.2
+ missinfo + PPLM	**98.3**	**90.1**
Reference	99.7	93.7

7.5 Analysis

Robustness to input information

We analyze how a model is robust toward the amount of information present. To measure the amount of information, we look for context length (description length for Amazon, dialog context length for Ubuntu) and the size of global schema since these two directly control how much knowledge regarding potential missing information is available to the model. We measure the difference in BLEU score between two groups of data samples where context length/size of global schema is either high or low. Figure 7.4 shows that our model is the least variant toward the information available hence more robust for the Amazon dataset.[7]

Owing to our modular approach for estimating missing information, we seek to analyze whether a question is really asking about missing information in an automatic fashion. This also allows us to explain the reasoning behind a particular generation as we are able to trace back to the particular missing information that is used to generate the question. We run a YAKE extractor on the generated questions to obtain key-phrases. We calculate the ratio between the number of key-phrases in the output that belong to the original missing schema and the total number of key-phrases present in the output. Table 7.8 shows that when we use our framework of estimating missing information coupled with BART, both models achieve very high missing information overlap, thus suggesting that we can obtain the reasoning behind a generated question reliably by tracing the missing information overlap, as shown in Table 7.7.

Question length

We also observe in Table 7.7 that baseline models tend to generate short and generic questions as compared to our model that often chooses longer schema key-phrases (e.g. bigrams) to generate a more specific question. We further looked into annotated (for usefulness) questions from the Amazon dataset and we observed that 70% of questions that were annotated as useful are longer than not-useful questions. The average length of gold useful questions is 10.76

---

[7] Ubuntu follows similar trends.

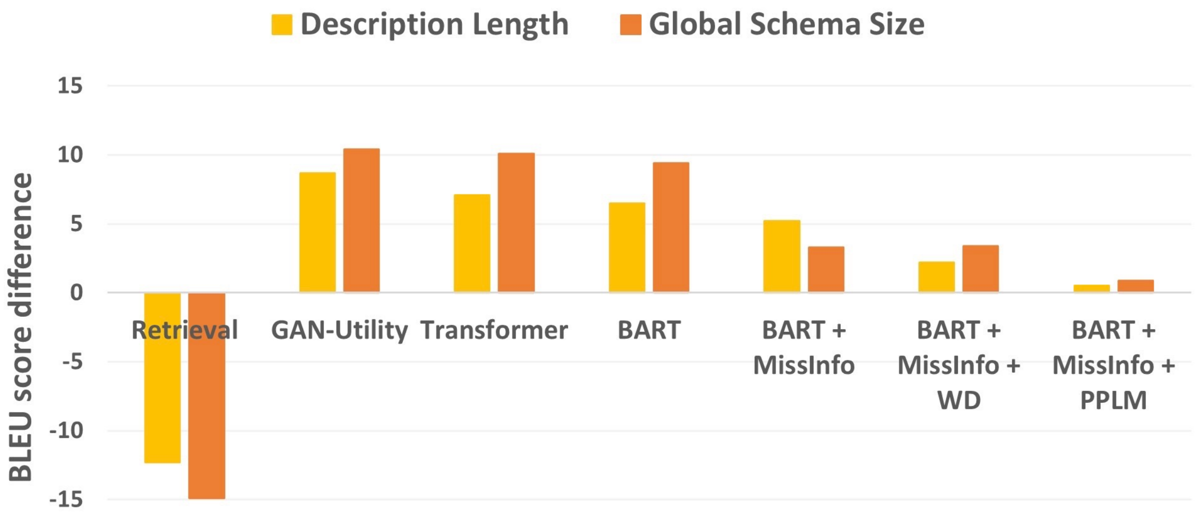


Figure 7.4. Average BLEU score difference between classes having longer (> 200 (median) words) and shorter descriptions; larger (> 200 (median) key-phrases) and shorter global schema for the Amazon dataset. Lower differences indicate more invariance toward the available information.

words and 8.21 for not-useful questions. The average length of generated questions for BART, BART+MissInfo and BART+MissInfo+PPLM (ours) are 5.6, 6.2, 12.3 respectively. We also find a similar trend in the Ubuntu dataset as well.

Dynamic expansion of global schema

We anticipate that even if we build the global schema from the available offline dataset, it is possible that new entries may appear in a real application. We investigate how our framework responds to the dynamic expansion of global schema. We simulate a scenario where we extend the "Laptop Accessories" category in the Amazon dataset, with 100 new products (those that appeared on Amazon.com after the latest entry in the dataset). We obtain key-phrases from their product descriptions and include them in the global schema for the category which amounts to a 21% change in the existing global schema. For 50 random products in the test set from the same category, we found that in 28 out of 50 cases (56%), the model picked a new schema element that is added later. This indicates that our framework is capable of supporting dynamic changes in the global schema and reflecting them in subsequent generations without retraining from scratch.

7.6 Conclusion

We propose a model for generating useful clarification questions based on the idea that missing information in a context can be identified by taking a difference between the global and the local view. We show how we can fine-tune a large-scale pretrained model such as BART on such differences to generate questions about missing information. Further, we show how we can tune these generations to make them more useful using PPLM with a usefulness classifier as its attribute model. Thorough analyses reveal that our framework works across domains, shows robustness towards information availability, and responds to the dynamic change in global knowledge. Although we experiment only with Amazon and Ubuntu datasets, our idea is generalizable to scenarios where it is valuable to identify missing information such as conversational recommendation, or eliciting user preferences in a chit-chat, among others.

Chapter 7, in part, is a reprint of the material as it appears in "Ask what's missing and what's useful: Improving Clarification Question Generation using Global Knowledge" by Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, Julian McAuley, which was published in *North American Chapter of the Association for Computational Linguistics*, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 8

REXC: Knowledge Grounded Self-rationalization via Extractive and Natural Language Explanations

8.1 Introduction

Two approaches that currently predominate for building self-explainable neural models are (i) selecting a subset of input features responsible for a prediction, known as an *extractive rationale* (ER) [Zaidan and Eisner, 2008, Bastings et al., 2019a, Sha et al., 2021], and (ii) generating a *natural language explanation* (NLE) for a prediction [Park et al., 2018, Hendricks et al., 2016, Camburu et al., 2018, Kayser et al., 2021]. For an explanation (ER or NLE), one is interested in two characteristics: *quality* (or *plausibility*) and *faithfulness*. Quality measures the degree of matching between the model's explanations and some ground truth; models with low-quality explanations would be undeployable. *Faithfulness* measures how well the explanations reflect the decision-making processes behind the predictions; unfaithful explanations would be misleading.

ERs are concise and provide quick explanations, which may sometimes be enough for users to assess the trustworthiness of the model. However, ERs may not have the means to provide important details of the reasoning of a model (e.g., relations between features) [Wiegreffe et al., 2021b]. In such cases, NLEs can be complementary, as they allow for detailed justification

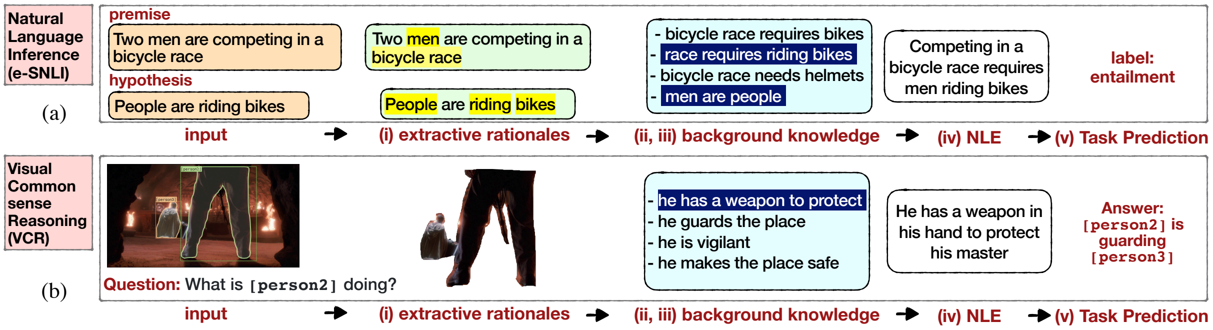

Natural Language Inference (e-SNLI)
(a)
premise
Two men are competing in a bicycle race
hypothesis
People are riding bikes
Two men are competing in a bicycle race
People are riding bikes
- bicycle race requires bikes
- race requires riding bikes
- bicycle race needs helmets
- men are people
Competing in a bicycle race requires men riding bikes
label: entailment
input
(i) extractive rationales
(ii, iii) background knowledge
(iv) NLE
(v) Task Prediction
Visual Common sense Reasoning (VCR)
(b)
Question: What is [person2] doing?
- he has a weapon to protect
- he guards the place
- he is vigilant
- he makes the place safe
He has a weapon in his hand to protect his master
Answer: [person2] is guarding [person3]
input
(i) extractive rationales
(ii, iii) background knowledge
(iv) NLE
(v) Task Prediction

Figure 8.1. Illustrative examples for RExC on (a) natural language and (b) vision-language tasks.

in a form that is most accessible to humans (natural language). However, machine-generated NLEs, like other generated text, are prone to lacking background knowledge (e.g., commonsense) [Camburu et al., 2020, Mao et al., 2019]. This could be because the NLEs are unfaithful or the model did not use the necessary knowledge in its decision-making process. Despite the complementary nature of ERs and NLEs, self-rationalizing models usually provide only one of them, with a few exceptions [Park et al., 2018, Wu and Mooney, 2019]. Moreover, while knowledge grounding has been done for black-box models [Bauer et al., 2018, Chandu et al., 2021, Chen et al., 2020a], we are not aware of any work on knowledge grounding for self-rationalizing models. Furthermore, existing self-rationalizing models are often outperformed by black-box models at solving the task at hand, leading to an undesirable performance-explainability trade-off.

To ground both decision-making and rationalization in background knowledge, as well as to reap the benefits of both ERs and NLEs, we combine these three ingredients in a unified self-rationalization framework. Our framework, which we call **RExC** (Extractive **R**ationales, Natural Language **Ex**planations, and (here) **C**ommonsense)[1], performs five steps: (i) selects a subset of the input features as an ER, (ii) inputs the ER to a knowledge resource to obtain a set of knowledge snippets about the ER, (iii) selects a subset of the snippets as the most relevant ones for solving the instance, (iv) passes the selected snippets to an NLE generator, (v) passes the generated NLE to a predictor that outputs the final answer (see Figs. 8.1 and

---

[1] Code is available at https://github.com/majumderb/rexc

8.2). All steps are learned jointly. REXC does not require direct supervision on the ER and snippet selections, which are modeled by two series of latent variables and variational learning (Section 8.2). Supervision comes from the final answers and NLEs.

REXC is illustrated in Fig. 8.1. In Fig. 8.1b, a subset of super-pixels of an input image form the selected ER for the question-answering instance. To answer that "`Person2` is guarding `person3`" and explain the answer, the model needs to identify that `person2` holds a weapon and have the knowledge that weapons are used to protect.

In our experiments spanning natural language (NL) and vision-language (VL) domains, we find that REXC significantly improves the quality of both ERs and NLEs, while bridging the gap between task performance and explainability. We also show, via perturbation analysis, that the explanations from REXC exhibit necessary conditions of faithfulness. Finally, REXC allows the selection of relevant knowledge snippets even without supervision from the NLEs. As these snippets can act as NLEs, we provide a zero-shot model with NLEs (REXC-ZS), which proves to be competitive with its supervised version.

The contributions of this work are summarized as follows:

- We propose a novel self-rationalizing framework that incorporates background knowledge and provides two complementary types of explanations: ERs and NLEs.
- REXC consistently outperforms previous best models that produce at least one type of explanation and performs on par with the SOTA models that do not provide any explanation, thus bridging the gap between explainability and task performance.
- REXC largely outperforms the previous SOTA in NLE and ER quality.
- REXC passes necessary faithfulness tests.
- REXC allows for a zero-shot setting in terms of NLEs (REXC-ZS), which sometimes outperforms models trained with a full training set of NLEs.

8.2 REXC

We aim to build a model that solves a task and explains its predictions via both ERs and NLEs. Furthermore, we aim for our model to benefit from resources of background knowledge, which could be general commonsense or domain-specific. To this end, REXC combines these three ingredients in the following way: it extracts rationales from the input, uses them to query an incorporated knowledge module to obtain knowledge snippets, selects the most relevant snippets, generates an NLE, and gives the prediction. We use Fig. 8.1a as a running example and Fig. 8.2 for an overview of the architecture.

8.2.1 Extractive Rationales via Binary Latent Variables

We define a neural module $\mathscr{R}$ that selects an ER from the input. An ER is a minimal sufficient subset of input parts (e.g., tokens for text or super-pixels for images) most responsible for the model's prediction [Lei et al., 2016b]. In Fig. 8.1a, we see an example from the natural language inference task [Bowman et al., 2015] (details in Section 8.3), where the ER is {"men", "people", "bicycle race", "riding bikes"}, the most responsible units for the prediction (*entailment*).

We model the selection of ERs using a series of latent variables ranging from $[0,1]$ ($z_i^r \in \mathscr{Z}^r$) over the N input units. A unit becomes a part of the ER *iff* its associated variable takes value 1. Following [Bastings et al., 2019a], we use the Hard Kumaraswamy distribution (referred to as HardKuma) as the reparameterization strategy to learn these latent selectors using backpropagation. The parameters of the neural module $\mathscr{R}$ are denoted by θ^r, which estimate the HardKuma variables for the input units. We also encourage the ERs to be terse, and we control the sparsity using an L_1 relaxation defined by the tractable Kumaraswamy CDF.

8.2.2 Knowledge about an Extractive Rationale

We hypothesize that inferred knowledge about the ERs are the most important bits of information for the predictions and, implicitly, for the NLEs. For example, in Fig. 8.1a, we

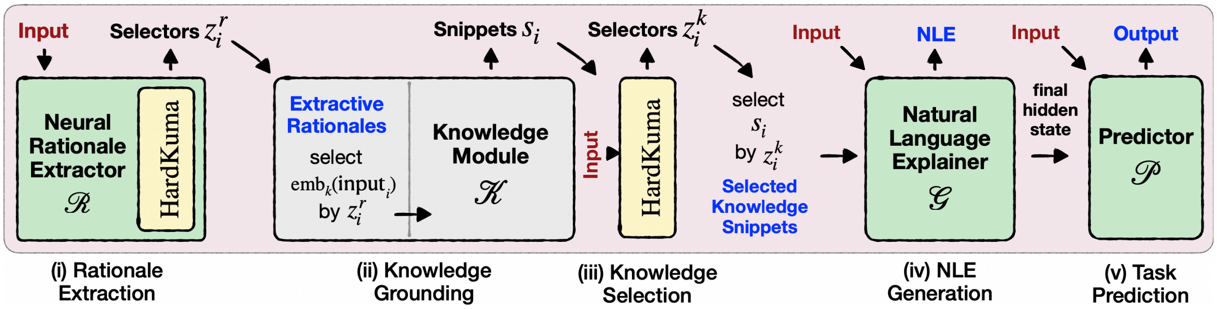


Figure 8.2. Architecture of REXC. The knowledge module is frozen, while the rest of the modules are trained jointly with the signals from the NLEs and outputs. Deliverables from REXC are in blue.

obtain relevant knowledge snippets (*bicycle race requires bikes* and *men are people*) for the ER ("bicycle race", "men", "people"), which influence both the prediction and the NLE.

We use a knowledge module $\mathscr{K}$, which supports input from an appropriate modality (e.g., text or image) for querying. We query $\mathscr{K}$ with each contiguous element of the ER (e.g., "bicycle race") to obtain a large pool of associated knowledge snippets $\mathscr{S}$. We take advantage of recent developments in generative models capable of providing background knowledge about a given entity for the ease of end-to-end training, such as `COMET` [Bosselut et al., 2019] for NL inputs and `VisualCOMET` [Park et al., 2020] for image inputs. The generative knowledge module does not suffer from the no-hit issue that is typically encountered in retrieval settings. However, REXC is flexible to accommodate a retrieval-based knowledge source when equipped with a differential search (see Section 8.4.4). To facilitate end-to-end training, we use soft representations of the elements of the ER—which are encoded using the embedding layer of $\mathscr{K}$ and subsequently selected by z_i^r (when 1) for queries to $\mathscr{K}$. Finally, we denote the parameters of $\mathscr{K}$ as θ^k.

8.2.3 Knowledge Selection

While the knowledge module generates several knowledge snippets ($\mathscr{S}$), not all of them are relevant for the prediction. Hence, we introduce a knowledge selection step. Furthermore, the selected knowledge snippets can appear as supporting evidence in addition to the generated NLE—an advantage of REXC over models that only generate NLEs.

We model the selection step via another set of latent selectors $z_i^k \in \mathscr{Z}^k$, which take a value from the interval $[0,1]$ and are realized by a HardKuma distribution (similarly to Section 8.2.1). More than one knowledge snippet may be relevant, however, we want the knowledge selection to be sparse. Hence, we use L_1 regularization to control the sparsity of the selected knowledge. The parameters predicting the latent selectors z_i^k are denoted as θ^{ks}.

To facilitate end-to-end training, we do not decode knowledge snippets into natural language. Instead, we retain the final hidden representations of each snippet from the knowledge module as $s_i \in S$. Using z_i^k as an indicator of selection, we obtain the vectors of selected knowledge snippets and concatenate them as input to the NLE generator. We also concatenate the representation of the input for the selector to be able to select the most relevant snippets given the input. At inference time, we decode the selected knowledge snippets into language, which could be used as additional supporting evidence along with the NLE. We call this variant REXC+. Human evaluation shows that this additional evidence leads to higher quality explanations (Section 8.4.1).

8.2.4 NLE Generation and Task Prediction

We use a natural language decoder $\mathscr{G}$, which concatenates the soft representations of the knowledge snippets and of the instance input at the input layer and generates an NLE. After $\mathscr{G}$, we add a predictor module $\mathscr{P}$, a linear layer with softmax, which takes the final hidden representation of the NLE and the representation of the instance input, and projects them to the output space for the task prediction. The prediction is thus directly conditioned on the NLE and the input, and, implicitly, on the ER and selected snippets. We denote the parameters of $\mathscr{G}$ and $\mathscr{P}$ as θ^g and θ^p, respectively. We use direct supervision from the ground-truth NLEs and task outputs.

8.2.5 Training

The parameters for $\mathscr{R}$, $\mathscr{G}$, $\mathscr{P}$, and the knowledge selector can be jointly trained end-to-end with backpropagation by summing up the negative log-likelihoods for the predictions and NLEs. We found that updating parameters for the knowledge resource $\mathscr{K}$ led to a minimal improvement; hence, $\mathscr{K}$ is fixed for computational ease.

However, due to the presence of z_i^rs in $\mathscr{R}$, we instead have to optimize a lower bound $\mathscr{E}$ of the original log-likelihood. We follow Bastings et al. [2019a] and optimize $\min_{\theta^r,\theta^g,\theta^{ks},\theta^p} \mathscr{L}_1$ with

$$\mathscr{L}_1 = -\mathscr{E}(\theta^r,\theta^k,\theta^{ks},\theta^g,\theta^p) + \lambda_0^r \sum_{i=1}^{N} z_i^r + \lambda_1^r \sum_{i=1}^{N-1} \left| z_i^r - z_{i+1}^r \right|,$$

where the second term is the L_1 penalty, the third term is a fused Lasso to control the total number of transitions for compactness [Lei et al., 2016b], and λ_0^r and λ_1^r are hyperparameters. Similarly, we have another lower bound for the z_i^k variables in the knowledge selection step, for which we optimize $min_{\theta^{ks},\theta^g,\theta^p} \mathscr{L}_2$ with

$$\mathscr{L}_2 = -\mathscr{E}(\theta^{ks},\theta^g,\theta^p) + \lambda_0^k \sum_{i=1}^{M} z_i^k,$$

where the second term denotes L_1 regularization for sparse knowledge selection. Finally, we combine the lower bounds as $\alpha \times \mathscr{L}_1 + (1-\alpha) \times \mathscr{L}_2$, where $\alpha \in [0,1]$ is a hyperparameter. We estimate the gradient of $\mathscr{E}$ via Monte-Carlo sampling from the reparameterized HardKuma variables [Kingma and Welling, 2014]. All hyperparameters are chosen based on a greedy search over the task prediction accuracy.

Table 8.1. Our tasks: three Natural Language and two Vision-Language.

Task	Dataset	Summary
Commonsense Validation	ComVE [Wang et al., 2019]	Choosing input sentence that defies commonsense
Natural Language Inference	e-SNLI [Camburu et al., 2018]	Textual entailment between premise and hypothesis
Commonsense Question Answering	COSe [Rajani et al., 2019]	Answering multi-choice commonsense questions
Visual Entailment	e-SNLI-VE [Kayser et al., 2021]	Entailment between image premise and text hypothesis
Visual Commonsense Reasoning	VCR [Zellers et al., 2019a]	Commonsense reasoning in visual question-answering

8.3 Experiments

Tasks.

We experiment with three tasks of natural language and two tasks of vision-language understanding as described in Table 8.1.

Implementation Details.

The components of REXC for the **NL tasks** are: **Rationale extraction:** We use the denoising encoder-decoder `bart-large` [Lewis et al., 2020a] with a linear layer and softmax at the end to generate the distribution for latent selectors. **Knowledge source:** We pre-train a `bart-large` model as a proxy for `COMET` (matched with original perplexity, 11.47 vs. 11.14 as from [Bosselut et al., 2019]) that matches the tokenization scheme used in $\mathscr{R}$. **NLE and task output:** We use another `bart-large` model to generate the NLEs, decoded with top-p sampling ($p = 0.95$) [Holtzman et al., 2020]. A linear layer followed by a softmax is used as the task predictor $\mathscr{P}$.

The components of REXC for the **VL tasks** are: **Rationale extraction:** We use a transformer-based VL model, UNITER [Chen et al., 2020b], which uses self-attention to learn contextualized representations for image-text input pairs. We add two MLPs on top of UNITER, which are used to generate the distributions for the latent ER selection from the image and text input; **Knowledge**

Table 8.2. Task performance (Acc.) and NLE quality for the (a) NL and (b) VL tasks. NLE Automatic metrics: METEOR, BERTScore, BLEURT, and NLE human evaluation metrics: e-ViL score, Yes/No %s. **Bold** are the best numbers ($p < 0.001$). Underline indicates best task performance for models with any explanations.

	ComVE							e-SNLI							COSe						
Model	**Acc.**	**MET.**	**BRTSc.**	**BLRT.**	**e-ViL**	**Yes**	**No**	**Acc.**	**MET.**	**BRTSc.**	**BLRT.**	**e-ViL**	**Yes**	**No**	**Acc.**	**MET.**	**BRTSc.**	**BLRT.**	**e-ViL**	**Yes**	**No**
Gold	–	–	–	–	91.6	79.3	1.1	–	–	–	–	98.1	94.1	2.7	–	–	–	–	84.8	74.5	1.8
Task SOTA	97.0	–	–	–	–	–	–	**93.1**	–	–	–	–	–	–	**83.7**	–	–	–	–	–	–
NILE	–	–	–	–	–	–	–	91.9	11.3	75.3	41.2	84.3	80.1	9.4	–	–	–	–	–	–	–
CAGE	–	–	–	–	–	–	–	–	–	–	–	–	–	–	72.1	1.3	43.1	16.9	59.5	35.4	16.7
WT5	96.1	3.4	86.4	27.0	67.7	46.2	11.0	92.1	12.3	75.3	42.3	85.3	82.7	12.8	81.0	2.2	52.0	22.4	73.0	53.9	10.5
RExC-ZS	96.7	7.7	72.4	24.2	65.8	56.5	16.3	92.4	11.9	63.2	40.7	88.3	85.8	5.5	83.1	2.6	38.1	17.1	83.4	73.2	5.6
RExC	**97.2**	**14.1**	**91.9**	**33.7**	87.3	**72.6**	2.8	92.9	**19.6**	**86.8**	**51.3**	**94.9**	93.9	3.6	83.6	**7.2**	**60.3**	**30.5**	87.4	**74.3**	2.1
RExC+	**97.2**	–	–	–	**88.4**	**72.6**	**1.2**	92.9	–	–	–	**95.6**	**94.3**	**2.7**	83.5	–	–	–	**87.9**	**74.7**	**1.8**
RExC-RB	96.4	3.1	89.5	26.1	62.2	43.3	15.1	92.7	13.2	77.4	45.3	87.6	81.2	13.5	82.2	3.7	55.5	23.8	79.3	63.2	9.6
w/o KN-Sel	97.1	11.3	90.2	33.6	84.4	65.3	5.1	92.8	17.9	83.4	51.2	92.8	91.7	5.8	83.2	6.4	58.4	27.9	85.0	70.2	2.5
w/o ER	96.5	5.2	86.1	28.1	67.2	43.4	7.6	92.3	13.1	77.7	43.5	83.4	83.2	15.1	81.4	2.9	52.8	23.8	66.7	45.2	14.9
w/o KN & ER	96.0	4.3	85.2	26.3	66.6	41.3	7.6	92.2	12.4	76.4	41.9	82.9	81.2	15.7	80.8	2.5	51.6	22.4	65.9	44.1	15.9

(a)

	e-SNLI-VE							VCR						
Model	**Acc.**	**MET.**	**BRTSc.**	**BLRT.**	**e-ViL**	**Yes**	**No**	**Acc.**	**MET.**	**BRTSc.**	**BLRT.**	**e-ViL**	**Yes**	**No**
Gold	–	–	–	–	90.6	79.3	1.1	–	–	–	–	95.8	94.1	2.7
Task SOTA	79.5	–	–	–	–	–	–	**81.6**	–	–	–	–	–	–
PJ-X	69.2	14.7	79.1	35.6	70.1	55.2	14.5	39.0	16.4	78.4	43.5	73.9	58.2	10.5
FME	73.7	15.6	79.7	34.5	71.9	56.7	13.2	48.9	17.3	79.4	47.8	73.0	56.2	11.1
RVT	72.0	18.8	81.1	35.3	72.2	55.4	12.8	59.0	11.2	78.9	44.2	73.2	57.4	11.5
e-UG	79.5	19.6	81.7	37.8	75.6	57.9	9.9	69.8	11.8	79.0	45.6	75.1	59.3	10.4
RExC-ZS	78.8	12.3	78.6	35.9	79.8	60.7	10.4	79.2	15.8	78.9	41.5	78.9	65.3	10.4
RExC	**80.8**	**22.9**	**87.7**	**39.6**	81.8	64.2	**6.5**	79.5	**20.9**	**86.6**	**53.1**	80.9	**67.7**	7.3
RExC+	**80.8**	–	–	–	**82.1**	**65.4**	6.3	79.5	–	–	–	**81.8**	67.2	**6.2**
RExC-RB	78.9	20.7	83.5	38.4	78.3	59.3	10.3	78.9	14.7	81.3	47.2	78.4	62.2	11.4
w/o KN-Sel	79.5	22.4	86.8	**39.7**	79.9	62.3	7.9	78.6	19.7	85.5	51.4	79.9	67.6	8.2
w/o ER	79.7	20.1	81.9	38.4	76.5	58.6	9.1	74.5	12.4	79.6	46.4	76.3	60.1	10.2
w/o KN & ER	79.4	19.5	81.7	37.7	75.5	57.9	9.8	69.8	11.9	79.0	45.8	75.1	59.4	10.5

(b)

source: We use `VisualCOMET` [Park et al., 2020] as an image-based commonsense module, which is fine-tuned on ATOMIC [Sap et al., 2019]. For text ERs, we follow the same setup as in the NL setup; **NLE and task output:** We use GPT-2 [Radford et al., 2019], a language decoder, for NLE generation. We adapt GPT-2 to condition on the representations learned by UNITER for VL inputs and use nucleus sampling ($p = 0.95$) for decoding the NLEs. A linear layer followed by a softmax is used for task prediction.

Baselines. We consider existing self-explainable models with the SOTA explanations (NLEs or ERs) as baselines. We also compare RExC with models that are SOTA for task performance (all until now are black-box models for our tasks).

NL Baselines.[2] The current SOTA for NLEs in all three NL tasks was obtained by WT5 [Narang et al., 2020], a general-purpose NLE generation model. We also compare with works that model NLEs specifically for a dataset: WT5 for ComVE, NILE [Kumar and Talukdar, 2020] for e-SNLI, and CAGE [Rajani et al., 2019] for COSe.

VL Baselines. We compare RExC with: PJ-X [Park et al., 2018] and FME [Wu and Mooney, 2019], two self-rationalizing models that provide both NLEs and ERs, and RVT [Marasovic et al., 2020], a post-hoc explainer that uses external knowledge as RExC. We also compare with e-UG [Kayser et al., 2021], the current SOTA in terms of NLE generation on VL tasks.

Ablations of RExC. We ablate RExC to investigate the effects of each component: ER selector (w/o ER), knowledge selector (w/o KN-Sel), and both (w/o KN & ER). We also ablate with the NLE generator (RExC-ZS), while training just using the final answers as supervision and using the selected knowledge snippets as NLEs. This yields a zero-shot model for NLEs. RExC+ adds the selected knowledge to the NLEs, hence is only used in the human evaluation. Finally, we also investigate the advantage of the generative knowledge module by replacing it with a retrieval-based knowledge source: ConceptNet [Speer et al., 2017] and Visual Commonsense Graphs [Zellers et al., 2019a]. To make the replacement, we use Maximum Inner Product Search as in [Lewis et al., 2020b]. We call this version RExC-RB.

8.4 Results

8.4.1 Evaluating the Quality of the Explanations

We evaluate the quality of the ERs and NLEs for RExC in comparison with the baselines.

Automatic Evaluation of NLEs. Following Kayser et al. [2021], we measure the quality of the NLEs by comparing them with the ground truth when the predicted label is correct. Here, we report METEOR [Banerjee and Lavie, 2005a], BERTScore [Zhang et al., 2020], and BLEURT [Sellam et al., 2020], which showed the highest correlation with human evaluation [Kayser et al.,

---

[2] We used the implementations from the original works.

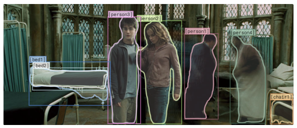
	Input	ER	Knowledge Snippets	NLE	SOTA NLE	Prediction
COSe	**Q:** People do many things to alleviate boredom. If you can't get out of the house you might decide to do what? **A:** a) play cards, b) skateboard, c) meet interesting people, d) listen to music	boredom, house, music	1. Music alleviates boredom 2. Music is listened at home	Music can alleviate boredom when you are alone at home	People listen to music	listen to music
VCR	 **Q:** Where are `[person3]` and `[person2]` right now?	 `[person2]`, `[person3]`	1. Hospital room has hospital beds 2. Hospital has nurses	There are hospital beds and nurses in the room	They are patients in the room	They are in a hospital room

Figure 8.3. Examples of NLEs and ERs generated from REXC along with selected knowledge snippets vs. those from the previous SOTA for the correct predictions for COSe and VCR.

Table 8.3. ER quality. Comparison of previous SOTA models [DeYoung et al., 2020a] for rationale extraction vs. REXC for ER quality. Best numbers are in **bold**.

	e-SNLI			COSe		
System	**Acc.**	**IOU**	**Tok.**	**Acc.**	**IOU**	**Tok.**
SOTA	73.4	70.5	70.2	34.6	38.9	51.9
REXC	**78.4**	**72.9**	**73.5**	**39.6**	**41.7**	**56.1**
w/o KN-Sel.	77.8	72.5	73.1	38.7	40.6	55.7

2021].

For NL tasks, REXC achieves the best values on all three automatic metrics (see Table 8.2a). We see sharp jumps (e.g., ranging from 4.8 to 11 points in METEOR) between REXC and models that do not use knowledge grounding, such as REXC w/o KN & ER and WT5. This confirms that background knowledge is a useful component for better NLEs. The gains for REXC over REXC w/o KN-Sel. show that knowledge selection provides a regularizing effect.

Similarly, REXC outperforms the previous SOTA models for VL tasks (see Table 8.2b). In particular, REXC outperforms RVT, a competitive model providing post-hoc NLEs also using the same commonsense resource as REXC, which possibly indicates that joint training for predictions and NLEs is superior over a post-hoc explainability approach.

Automatic Evaluation of ERs. To evaluate the quality of ERs, we directly compare them with gold ERs using ERASER [DeYoung et al., 2020a]. ERASER uses accuracy (Acc.) and overlap-based metrics such as F1 at Intersection-Over-Union spans (IOU) and token (Tok.) overlap. In Table 8.3, we show results for e-SNLI and COSe, the only ones from our list that have gold ERs available. We observe that REXC leads to significantly superior-quality ERs compared to models that do not use NLEs or background knowledge to influence rationale extraction (e.g., 56. vs. 51.9 F1). Thus, REXC achieves a new SOTA in ERs for both datasets. Possible explanations for this are: (1) additionally optimizing for NLEs constrains REXC to generate more informative ERs, and (2) to obtain better-suited knowledge snippets, REXC must extract high-quality ERs.

Human Evaluation of NLEs. Following Kayser et al. [2021], we asked human annotators to measure the quality of the generated NLEs. For each NLE, we asked: *Given the input, does the explanation justify the answer?* and provide four options: Yes, Weak-Yes, Weak-No, and No. We report the e-ViL score from [Kayser et al., 2021] combining results for each option with a weight of 1, $\frac{2}{2}$, $\frac{1}{3}$, and 0 respectively. We only evaluate NLEs for correct predictions and collect 250 random such examples for each model and each dataset.

For NL tasks, Table 8.2a shows that humans also rated the NLEs from REXC far better than those from the previous SOTA models. Again, REXC without knowledge selection shows large drops, which indicates that the knowledge selection step has positive effects on the quality of the NLEs.

For VL tasks, NLEs from previous SOTA models were rated far lower than ground truths, indicating an even bigger need for improvement. We observe substantial gains for REXC, even when compared to competitive models that already use external knowledge, such as RVT [Marasovic et al., 2020].

Often NLEs generated by REXC are longer than those from the baselines, since they are rich in background knowledge. In the human evaluation sample for e-SNLI, we found that 73% of NLEs from REXC are longer (at least by a token) compared to NLEs from WT5. However, we find that for REXC, length is loosely correlated with the e-ViL score with a Pearson's correlation score of 0.21. This correlation is similar (0.17) for NLEs from WT5. We also find similarly low correlations (0.13, 0.24, 0.14, and 0.20) between length and e-ViL score for ComVE, COSe, e-SNLI-VE, and VCR, respectively, which indicates that NLE length did not act as a confounding factor during human evaluation.

Qualitative Analysis. Fig. 8.3 shows sample outputs from RExC for COSe and VCR . We observe that NLEs from RExC are more grounded in knowledge than those from previous SOTA models. Moreover, previous SOTA NLEs fall short of being comprehensive NLEs (e.g., "People listen to music" for COSe), which could be because they do not condition on ERs (e.g., "boredom").

8.4.2 Task Performance

Until now, the SOTA models in terms of task performance for all five tasks were models that do not offer any explainability [Wang et al., 2020, 2021, Lan et al., 2020, Xie et al., 2019a, Yu et al., 2020]. Models that attempt to offer explanations (NLEs or ERs) faced a drop in accuracy (see Tables 8.2a and 8.2b). RExC bridges this important gap by matching SOTA task performance for 4 out of 5 tasks and even achieving a new SOTA for e-SNLI-VE, while providing two types of explanations, both of which are of higher quality than the previous models with SOTA explanations.

8.4.3 Zero-shot NLEs

Often, there exists a high overlap between the generated NLEs and the selected knowledge snippets. This is expected, since the NLEs and predictions are conditioned on the selected knowledge. This raises the question of whether the selected snippets alone could form sufficient NLEs. We argue that, in general, this is not the case, because the information in a background resource may not provide the whole reasoning behind a prediction. This information is only meant to add value but not replace the NLEs. However, in particular cases where the ground-truth NLEs consist mainly of pieces of background knowledge, selected snippets may be sufficient explanations. To investigate this for our datasets, we look at RExC-ZS, where relevant knowledge was selected *only* using the task prediction loss and concatenated to be used as NLEs. Tables 8.2a and 8.2b show that REXC-ZS performs poorly in automatic metrics, which is mostly due to being out of distribution w.r.t. the ground-truth explanations. However, in human evaluation, we see that even if the NLEs from REXC-ZS were not better than the generated NLEs from REXC, they were largely better than the NLEs from the previous SOTA models (which were trained with full training sets of NLEs) for 4 out of the 5 tasks. These results indicate that: (1) the NLE module in REXC acts as an important conditional generation step that makes NLEs fluent and more comprehensible; and (2) despite being less fluent, concatenated knowledge snippets can act as NLEs in cases where ground-truth NLEs are not present. This shows the potential of REXC for zero-shot natural language rationalization.

8.4.4 Generative vs. Retrieval-based Knowledge Module

One of the reasons for choosing a generative knowledge module (e.g., `COMET`) is to avoid the no-hit issue of indexed knowledge bases. For example, when we replaced `COMET` with ConceptNet [Speer et al., 2017], for e-SNLI, we found that 23% of instances do not retrieve any knowledge snippet. As expected, REXC-RB performed worse than REXC (see Tables 8.2a and 8.2b).

8.5 Evaluating Faithfulness

Evaluating the faithfulness of explanations is a challenging open question for both ERs [Jacovi and Goldberg, 2021] and NLEs [Wiegreffe et al., 2021b]. We analyze REXC for faithfulness based on existing works.

8.5.1 Faithfulness of the NLEs

Evaluating the faithfulness of NLEs is still in its infancy. To our knowledge, Wiegreffe et al. [2021b] is the only work that provides (two) necessary conditions for NLEs' faithfulness: *feature importance agreement* and *robustness equivalence*. Both conditions perturb the input and

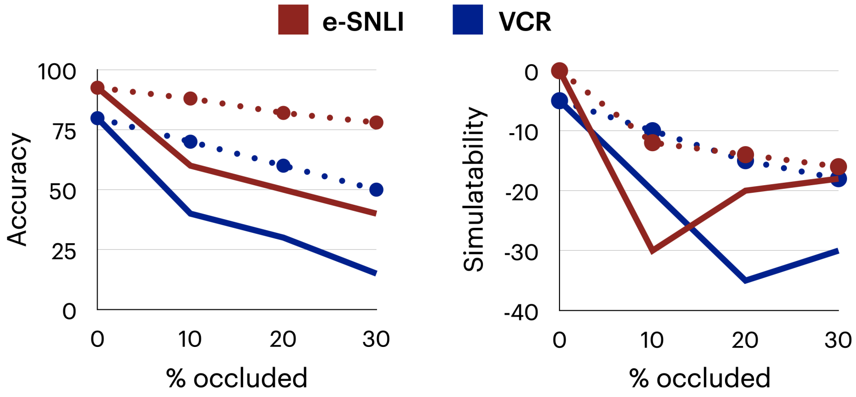


Figure 8.4. Feature importance agreement. Left: Accuracy when important features (solid) vs when random features (dotted) are occluded. Right: Simulatabilities important features (solid) vs when random features (dotted) are occluded.

measure the change in model behavior in order to establish the extent of label-NLE association. As they mentioned, there are currently no sufficient conditions for faithful NLEs, since there can be different realizations of NLEs that significantly (but differently) contribute to the model's prediction process.

Changes in Model Behavior. Change in model behavior can be captured by changes in task accuracy and changes in the predictive ability of NLEs. The predictive ability of NLEs over inputs (formally termed as *simulatability* [Doshi-Velez and Kim, 2017, Hase et al., 2020]) is defined by the change in task accuracy when the generated NLEs are appended to the input. To ensure NLEs' faithfulness, changes in accuracy and in NLEs (via simulatability) should be similarly affected by changes in the input.

Feature Importance Agreement. This condition uses a gradient-based attribution technique to find the most important features with respect to an output (prediction or NLE). For a predicted class, a gradient attribution is the gradient of the predicted class's logit with respect to an input feature. The attribution score is calculated by performing an operation (here, L_1 norm) to turn the gradient into a scalar quantity. For REXC, we identify salient input features (tokens or super-pixels) with attribution scores (top-$\{10, 20, 30\}\%$) with respect to the task prediction. We measure the change in simulatability of NLEs when we remove these features from the input.

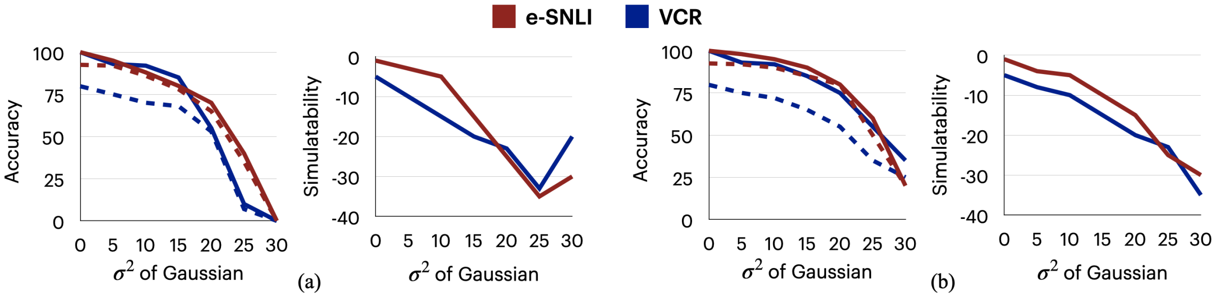


Figure 8.5. Robustness equivalence analysis when noise is added to the (a) input and (b) selected knowledge snippets. In each pair, the left chart shows % of stable (unflipped) labels (solid), and accuracy (dashed). The right chart in a pair depicts the simulatability of NLEs.

Similarly, we measure the change in task accuracy when we remove the features most important for the NLE generation. To ensure faithfulness, both these changes should be significantly higher than the changes that would appear if we were to remove random input features. Fig. 8.4 shows that the removal of salient input features similarly affects both task accuracy and NLEs simulatability when compared to random removal—ensuring that this faithfulness condition is met by RExC on e-SNLI and VCR. Similar trends are also observed in the other datasets.

Robustness Equivalence. The second necessary condition involves perturbing the input by adding zero-mean Gaussian noise $\mathcal{N}(0,\sigma^2)$ to the internal representations of its features and observing the corresponding changes in task accuracy and NLE simulatability for a range of noise values. We are interested in noise regions where labels and NLEs remain stable (small changes) and noise regions where labels and NLE become unstable (large changes). To indicate faithfulness of the NLEs, predicted labels and NLEs should remain stable (or unstable) at the same noise region. For better label-NLE association, the sharpest drop in simulatability and accuracy should align with the sharpest drop in % of stable labels. In Fig. 8.5, we see this condition holds true for RExC. For example, for e-SNLI (in Fig. 8.5(a)), we see that the point of minimum contribution of NLEs to the prediction coincides with the sharpest drop in task accuracy, at $\sigma^2 = 25$. Lower noise than $\sigma^2 = 25$ keeps both labels and NLEs stable, whereas higher noise will make both unstable. Similar trends are observed in other datasets as well.

8.5.2 Faithfulness of the ERs and Knowledge Snippets

For ERs, faithfulness metrics are more studied than NLEs in the literature [DeYoung et al., 2020a, Jacovi and Goldberg, 2021], and both necessary and sufficient conditions for faithfulness exist. DeYoung et al. [2020a] introduced two metrics for measuring faithfulness in ERs: *comprehensiveness* (necessary condition) and *sufficiency*. Comprehensiveness is measured by the change in task accuracy between the case when the full input is used for the prediction by the original model and the case when the ERs (from the original model) are dropped (masked for images) and the model is retrained on these new instances (with dropped ERs). A higher difference (maximum 1) would indicate a higher extent of faithfulness. Sufficiency can be calculated as the difference in accuracy between the case when the full input is used for the prediction and the case when *only* the ERs (from original model) are used to retrain the model. A closer to zero value indicates a higher degree of faithfulness. For REXC, we extend this to the selected knowledge snippets to also analyze their comprehensiveness and sufficiency for the task prediction. Table 8.4 confirms solid comprehensiveness (high values) and sufficiency (close to zero) for both ERs and selected snippets.

A baseline for checking faithfulness of ERs and knowledge selection is to check their sufficiency and comprehensiveness with respect to a *random* selection of input tokens as ER and a random selection of knowledge snippets. Table 8.4 shows that REXC achieves better comprehensive and sufficiency as compared to a random baseline. REXC also outperforms all models reported in DeYoung et al. [2020a] in both metrics.

8.6 Summary and Outlook

In this work, we proposed REXC, a self-rationalizing framework that incorporates background knowledge resources and provides two complementary types of explanations: ERs and NLEs. Using five tasks, from natural language and vision-language domains, we show that REXC obtains a new SOTA performance for both NLEs and ERs. We also close the important

Table 8.4. **Comprehensiveness (Comp.) and Sufficiency (Suff.)** metrics for ERs and selected knowledge snippets generated by REXC vs. random ERs and knowledge snippets

		ComVE	e-SNLI	COSe	e-SNLI-VE	VCR
ERs						
Random	Comp.	0.12	0.11	0.10	0.13	0.14
REXC	Comp.	**0.32**	**0.45**	**0.24**	**0.28**	**0.33**
Random	Suff.	0.44	0.31	0.54	0.51	0.39
REXC	Suff.	**0.14**	**0.08**	**0.05**	**0.10**	**0.13**
Knowledge Snippets						
Random	Comp.	0.12	0.14	0.14	0.10	0.09
REXC	Comp.	**0.56**	**0.49**	**0.36**	**0.27**	**0.35**
Random	Suff.	0.41	0.51	0.43	0.51	0.37
REXC	Suff.	**0.15**	**0.09**	**0.08**	**0.07**	**0.08**

gap between task performance and explainability for the five tasks that we experimented with, and obtained a new SOTA for e-SNLI-VE. While we used commonsense resources, future work could look into adding other types of knowledge resources, including more specialized ones, such as legal and medical. Additionally, while we showed that REXC opens up a promising direction for zero-shot NLE generation, further investigation could reap more benefits from the principals behind REXC for zero-shot and few-shot setups.

Chapter 8, in part, is a reprint of the material as it appears in “Knowledge-grounded Self-rationalization via Extractive and Natural Language Explanations” by Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, Julian McAuley, which was published in *International Conference on Machine Learning*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 9

INTERFAIR: Debiasing with Natural Language Feedback for Fair Interpretable Predictions

9.1 Introduction and Background

Debiasing human written text is an important scientific, and social problem that has been investigated by several recent works [Zhang et al., 2018a, Jentzsch et al., 2019, Badjatiya et al., 2019, Heindorf et al., 2019, Ravfogel et al., 2020, Gonen and Goldberg, 2019, He et al., 2021]. These methods primarily try to eliminate the biased information from the model's internal representations or from the input itself, disregarding the task performance during the process. Ideally and fairly, a model should use the necessary amount of information, irrespective of bias, to achieve an acceptable task performance. This trade-off between task performance and bias mitigation is subjective or varies between users [Yaghini et al., 2021] and is often hard to achieve via learning from data [Zhang et al., 2018a]. Our goal is to perform a predictive task (here, predicting a profession based on a biography) – however, we want to regulate the sensitive information that the model uses while maintaining the model's predictive performance.

Figure 9.1 shows the limit of an algorithmic approach where ignoring all gendered information can lead to a wrong result. However, a user can further tune the model's belief on the bias, leading to a correct prediction while minimally using biased information. While interactive

NLP models recently focused on model debugging [Tandon et al., 2021, 2022], improving explainability in QA [Li et al., 2022b], machine teaching [Dalvi et al., 2022], critiquing for personalization [Li et al., 2022a], and dialog as a more expressive form of explanations [Lakkaraju et al., 2022, Slack et al., 2022], we leverage on the untapped opportunity of adding interaction capabilities for debiasing a model's prediction using natural language feedback. Furthermore, allowing the user to interact with underlying prediction rationales addresses the subjective aspect of fairness and improves transparency.

Here, we propose INTERFAIR, a modular interactive framework that (1) enables users to provide natural language feedback at test time to balance between task performance and bias reduction, (2) provides explanations of how a particular input token contributes to the task performance and exposing bias, and finally (3) achieves better performance than a trained model on full-text input when augmented with feedback obtained via interactions.

9.2 Controlling Bias Exposure via Rationales

An interpretable debiasing algorithm should use a 'fair' amount of sensitive information as part of the input responsible for the prediction (a.k.a rationale). Let the input tokens that are predictive for the task output be called *task rationales* and tokens that reveal sensitive information be called *bias rationales*. We further use energy functions to probabilistically measure the importance of a token for an objective (either task performance or revealing bias). We argue that we can regulate the task energy of input tokes by overlaying bias energy values and blocking high-bias tokens that do not significantly degrade task performance.

9.2.1 Rationales and Energies

We first identify input tokens that carry sensitive information. For an input text $\mathbf{x} = \{x_1, \cdots, x_n\}$ (e.g., biography of a person), we predict an associated bias label (e.g. gender) using a function $f_b(\mathbf{x})$. Now to extract bias rationale, we assume a sequence of latent binary variables $\mathbf{z^b} = \{z_1^b, \cdots, z_n^b\}$, $z_i^b \in \{0, 1\}$ for each input token index i [Lei et al., 2016a], and feed them to f_b

to maximize the prediction probability of the correct bias label by regulating the contribution of each token:

$$\mathbf{z^b} \sim g_b(\mathbf{x}); \quad f_b(\mathbf{x} \odot \mathbf{z^b})$$

where g_b is a bias rationale extractor that predicts the probability of how much each token contributes to predicting the bias label. We construct *bias rationale* as $\mathbf{x} \odot \mathbf{z^b}$ where $\mathbf{z^b}$ is sampled from g_b (modeled as a HardKuma distribution [Bastings et al., 2019b] to maintain differentiability).

Now to measure the significance of each token for bias prediction, we define bias energy as the negative log-likelihood of the non-selection probability of each token [LeCun et al., 2006] for the bias prediction. Higher energy indicates more significance [1]. Hence, the bias energy for the i-th token would be:

$$e_i^t = -\text{log-likelihood}(p(z_i^b = 0)) = -\text{log-likelihood}(1 - g_t(x_i|\phi_t)),$$

where $g_b(x_i)$ is the probability of selection of x_i for the bias prediction.

Similar to bias rationales, we have a task rationale extractor g_t that produces $\mathbf{z_t} = [z_1^t, \cdots, z_n^t]$ and a predictor function f_t that predicts the task label using task rationales $\mathbf{x} \odot \mathbf{z}^t$ with energies as

$$\mathbf{z}^t \sim g_t(\mathbf{x}); \quad f_t(\mathbf{x} \odot \mathbf{z}^t); \quad e_i^t = -\text{log-likelihood}(1 - g_t(x_i)).$$

9.2.2 Training with Energy-based Constraint

Upon obtaining the task and bias energy for the i-th token, we penalize its importance for predicting the task if it has high bias energy. In contrast, tokens with low bias energy can retain

---

[1] Direct probabilities i.e. $p(z_i^b = 0)$ led to unstable performance.

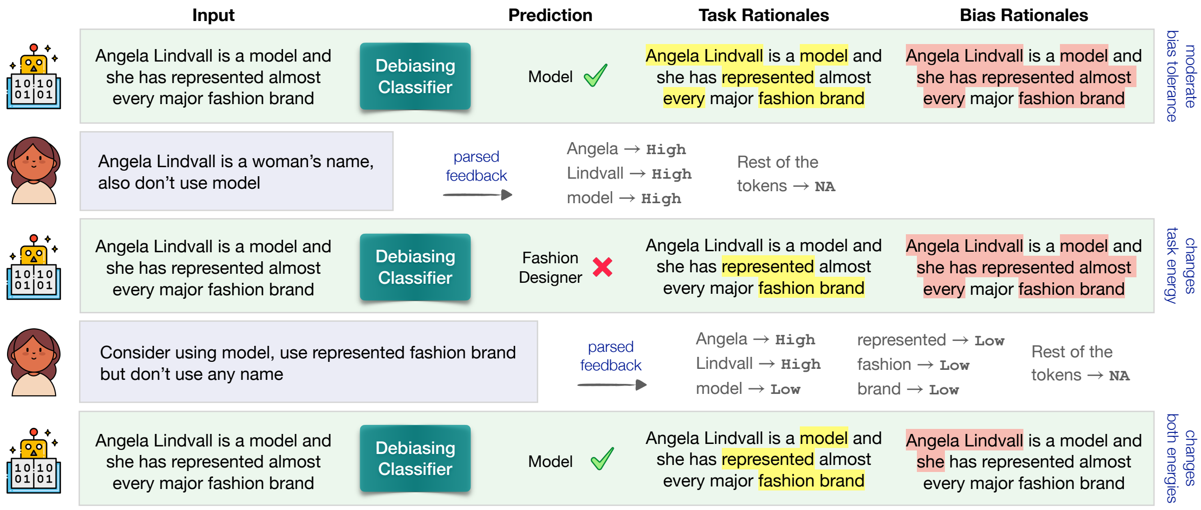


Figure 9.1. Interactive setup for InterFair where user provides feedback to update bias rationales that in turn updates task rationales and subsequent task prediction.

their task energy as is. Hence, we jointly train bias and task prediction modules with a debiasing constraint:

$$DC(i) = e_i^t + (e_i^b - \tau) \quad \text{if } e_i^b > \tau; \quad DC(i) = 0 \quad \text{otherwise}$$

where τ is a hyperparameter indicating the bias tolerance threshold. Ideally, this constraint will promote the use of low-bias energy replacements of the high-bias and high-task important tokens in order to maintain task performance. During training, we first obtain the bias prediction model, keep it fixed, and then update the task prediction model which is optimized with the debiasing constraint.

9.3 Natural Language (NL) Feedback with INTERFAIR

During inference, given an input, the model can produce a debiased output since it is trained to weigh less on the high-bias and high-task energy tokens and replace them with low-bias replacement. However, the bias module we use during training is neither perfect nor does it reflect a user's subjective view of sensitive information. The current (and default) model is regulated b1y the bias tolerance threshold τ, which is a hyperparameter and may or may not reflect the true bias tolerance threshold of the user.

To allow online modification of the output colored by the user's understanding of sensitive information, we need to be able to modify the model output accordingly. We hypothesize a system, **INTERFAIR**, that allows the user to provide new bias energy values for each input token which will override the current output and modify the model's prediction (and task rationales) accordingly. Users can access the task and bias rationales and adjust energy values accordingly. However, providing numerical energy values is often not practical and may limit users' interaction with such a system. Thus, we allow users to provide natural language feedback about input tokens and automatically convert them to relevant energy values for subsequent updates in model prediction.

9.3.1 Parsing NL feedback

The simplest form of feedback would be to provide a critique/correction on the bias energy of a certain input token by indicating whether would they be `high`, or `low`. To convert an NL feedback to actionable feedback for all input tokens, we treat it as a sequence labeling task. Specifically, we use a parser f_p that encodes the NL feedback, the bias variable (e.g. gender) and the original task input and produces a sequence of `High` / `Low` / `NA` labels for the complete input token sequence. An example feedback and its parse are shown in Table 9.1. Such an approach allows us to encode complex feedback where the user wants to provide feedback on multiple input tokens (see Figure 9.1).

Since we do not have large annotated data for the parsing task, we instead adopt a few-shot framework, following [Slack et al., 2022]. We use a large language model (LLM) (e.g. GPT-J[2]) as they have strong priors for language understanding (here, parsing) tasks from their pre-training phase. We use few demonstrative parsing examples for in-context learning of f_p. See complete task prompt in Table 9.1.

---

[2] https://huggingface.co/docs/transformers/model_doc/gptj

9.3.2 Modifying Task Rationales

Upon converting the user feedback to an actionable feedback for each input token, we first calculate the updated bias energies which, in effect, change the task energies and task rationales. We explore two heuristic strategies for to update the bias energies from user feedback: 1) **Coarse** where for the i-th token, we assign the Bernoulli probability $p_{\text{user}}(z_i^b = 1)$ as 1 if the feedback is `High` or 0 if the feedback is `Low`; and 2) **Fine** where we use the probability of feedback labels obtained from the NL parser as a soft score for $p_{\text{user}}(z_i^b = 1)$. Finally, for smoothing, we obtain the new bias energies as:

$$p_{\text{new}}(z_i^b = 1) = \alpha g_b(x_i) + (1-\alpha)p_{\text{user}}(z_i^b = 1); \quad e^b_{i,\text{new}} = -\,\text{log-likelihood}(1 - p_{\text{new}}(z_i^b = 1))$$

Now, to mimic the training-time behavior of how task and bias energies are related via a debiasing constraint, we use the same constraint $DC(i)$ for i-th token to update the task rationale based on updated bias energies during inference. We keep $\tau = 0$ to emphasize the effect of the user feedback.

9.4 Experiments and Results

We break our experiments into two parts: 1) developing the NL parser and 2) interactive debiasing with INTERFAIR. We use BiosBias [De-Arteaga et al., 2019], a dataset made from a large-scale user study of gender in various occupations. It contains short biographies labeled with gender and profession information, and a possible confluence exists between gender and annotated profession labels.

Using INTERFAIR, we would like to predict the profession from biographies without the influence of gender. Following [Ravfogel et al., 2020], we use 393,423 biographies with binary gender labels (male/female) and 28 professions labels (e.g. professor, model, etc.). We initially used 255,710 examples for training and 39,369 for validation. We use 500 examples (a random

Table 9.1. NL feedback parser. Parse examples for IID and compositional (Comp) splits; parsing accuracy on IID, compositional splits and overall test set. GPT-Neo has 2.7B parameters, GPT-J has 6B parameters.

Examples	Models	IID	Comp	Overall
Assign High/Low/NA for each input token given bias and feedback.	GPT-Neo			
`[Input]` Angela Lindvall is a model and she represented (...)	5 shot	40.6	10.2	30.6
`[Bias]` Gender	10 shot	54.2	15.1	34.4
`[Feedback]` Angela Lindvall is a woman's name	20 shot	62.1	17.2	35.9
`[Parse]` `High, High, NA, NA, NA, NA, NA, NA` (...) *IID*	GPT-J			
`[Feedback]` Don't use model, no gendered names or pronouns	5 shot	46.7	14.2	33.1
`[Parse]` `High, High, NA, NA, High, NA, High, NA,` (...) *Comp*	10 shot	62.6	25.8	40.2
	20 shot	**73.8**	**30.1**	**57.1**

sample from the rest 25%) as a test set for interactive debiasing. We perform a user study with two annotators who optionally provide feedback to INTERFAIR to improve task performance or minimize bias in task rationales.

Following [Slack et al., 2022], we use 5, 10, or 20 examples annotated by two independent annotators for the NL parser. We additionally obtain a set of 50 more annotations for testing the parser. While testing the performance of the parser, we use the accuracy metric, i.e. if the parsed feedback matches with the gold parse. We also consider two splits for testing: an IID split where the gold parse contains non-NA labels for one or two contiguous input token sequences and a compositional split (to check LLM's generalization performance [Oren et al., 2020]) where the gold parse has three or more contiguous token sequences.

Table 9.1 shows the parsing accuracy for IID and compositional splits as well for the overall test set. The compositional split is much harder than the IID since it contains more complex feedback for more than three contiguous token sequences (or phrases). Comparatively, GPT-J performs better than GPT-Neo model. Also, more demonstration examples work better for in-context learning. The few-shot parsing using LLMs is faster and easier to adapt with newer user feedback (GPT-J shows reasonable generalization performance) instead of finetuning a supervised model [Slack et al., 2022].

To understand the efficacy of rationale energy-based debiasing, we consider two other

Table 9.2. Evaluation for task accuracy (Acc. ↑), bias (F1 ↓), and faithfulness for task rationales: Comprehensiveness (Compre. ↑) and Sufficiency (Suff. ↓)

Models	**Acc.**	**Bias F1**	**Compre.**	**Suff.**
Full Text	0.81	0.98	–	–
Reranking	0.70	0.45	0.23	0.32
Adv	0.36	0.35	–	–
INTERFAIR				
No feedback	0.80	0.38	0.52	0.01
Coarse	0.84	0.33	0.51	0.01
Fine	**0.85**	**0.31**	**0.48**	**0.00**

variants: *Rerank*, an inference-time debiasing variant where the task rationale is considered based on ascending order of bias energy; (2) *Adv*, a model trained with an adversarial objective [Zhang et al., 2018a] to debias the model's latent space, but incapable of producing any rationales.

Table 9.2 shows that when we use Full Text as task input, the bias in task rationales is very high. Reranking decreases the bias but also incurs a drop in task performance. The adversarial method does not produce any explanation and cannot use any additional feedback, leading to low task performance. INTERFAIR without feedback balances the task performance and bias very well. Even though 81% (Full Text performance) is the upper bound of accuracy for purely training-based frameworks, INTERFAIR allows for further modifications of the rationales to improve task performance while decreasing bias. Indeed, we see both Coarse and Fine feedback enhance task performance beyond what Full Text can achieve standalone while keeping the bias minimal. It indicates that 1) full text based training suffers from spurious correlation or noise that hampers task performance, and 2) interactive debiasing is superior to no feedback since it produces better quality human feedback to refine task performance while eliminating bias. Since test-time interactions modify task rationales, we check their faithfulness using comprehensiveness and sufficiency (check definitions in [DeYoung et al., 2020b]), which shows these still exhibit a high degree of faithfulness.

In summary, INTERFAIR shows the possibility of achieving even more accurate outcomes than simply training a model with full input. Moreover, it shows that prediction rationales are editable without losing faithfulness, thus providing controllability. So far INTERFAIR does not memorize previous feedback at a loss of generalization—this can be addressed via memory-based interactions [Tandon et al., 2022], or model editing [Mitchell et al., 2021] as a future work.

Chapter 9, in part, is a reprint of the material as it appears in “InterFair: Debiasing with Natural Language Feedback for Fair Interpretable Predictions” by Bodhisattwa Prasad Majumder*, Zexue He*, Julian McAuley, which is being prepared for submission. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 10

Related Work

10.1 Related Work for COMPAC

Building personalized dialog agents has been a popular task recently, thanks to Zhang et al. [2018b] who extensively studied the task with a new dataset PERSONA-CHAT, later as a form of a challenge [Dinan et al., 2019a], where the dialog agent is seeded with a predefined persona in the form of multiple sentences of textual description, mirroring a casual human conversation which many times draws snippets from individual personal experiences and facts. Recent works focus on improving persona-grounded dialog generation performance [Wolf et al., 2019, Mazaré et al., 2018, Bao et al., 2019] as well as persona consistency in generated dialog [Welleck et al., 2019, Li et al., 2019a, Song et al., 2019a]. Bao et al. [2019] proposed a reinforcement-learning-based framework that promoting informativeness and persona-consistency via personal knowledge exchange. Xu et al. [2020b] focused on using plausible topical keywords related to the available persona facts using a neural topic model to explore beyond the given knowledge, possibly closest to our work. We rather focus on obtaining commonsense implications of the given persona in the form of text snippets that are more expressive than topical keywords.

Persona-grounded dialog generation is a special case of knowledge-grounded dialog generation. Knowledge grounding in dialog has many real-world applications that are well-studied in recent literature [Zhou et al., 2018, Ghazvininejad et al., 2018, Dinan et al., 2019b, Lewis et al., 2019b]. In this work we use fine-grained grounding/selection on persona which performed better than encoding the entire persona for each response. Such fine-grained selection has been found useful in prior works on text generation such as dialog [Lian et al., 2019] and image captioning [Jhamtani and Berg-Kirkpatrick, 2018]. For dialog generation, a contextual knowledge selection has been successfully applied in prior works [Parthasarathi and Pineau, 2018]. Specifically, Zhao et al. [2017] and later Song et al. [2019b] proposed a conditional-VAE framework to learn latent context given the dialog history to guide knowledge selection.

Finally, few recent works focused on augmenting grounding with commonsense knowledge with successful applications in open-domain topical dialog generation [Ghazvininejad et al., 2018, Moon et al., 2019], story generation [Mao et al., 2019] and sarcasm generation [Chakrabarty et al., 2020b]. In this work, we extend this effort into persona-grounded dialog generation via augmenting grounding persona with commonsense knowledge.

10.2 Related work for INTERVIEW

Media dialog—specifically, the news interview—has seen study primarily in the field of speech transcription, diarization, and speaker role modeling [Chen et al., Laurent et al., 2014]. These works have typically focused on techniques to annotate broadcast audio transcripts [Hutchinson et al., 2010] in order to cluster different news stories from a continuous broadcast stream [Huang et al., 1999]. While Barzilay et al. [2000] and Liu [2006] note that transition points between speaker roles (e.g. anchor and guest) can determine the high-level topical flow of a news conversation, we investigate the impact of discourse patterns on the semantics of specific utterances.

Such research is currently limited by a lack of accessible corpora for the study of media dialog at scale. The Defense Advanced Research Projects Agency has undertaken efforts to collect and transcribe broadcast conversations [Strassel, 2004, Cohen, 2007]. However, it proves difficult to adopt these datasets as widely available benchmarks on dialog modeling tasks, as they come with a substantial cost ($100-$1000 per annum per dataset). More recent efforts to amass such data have either focused on collecting large volumes of conversation fragments with noisy transcripts [Beeferman et al., 2019] or human transcripts for a smaller set of long-form open-domain radio programs [Mao et al., 2020]. We contribute an open-access large-scale corpus of broadcast media dialog annotated with response types, demonstrating that these are useful for modeling interviewer utterances.

We explore the application of discourse analysis [Fairclough and Wodak, 1997] on this large media dialog corpus in order to discover, confirm, and leverage *discourse patterns* regarding interrogative forms, speaker agency, and references to external knowledge. As noted by Weizman [2008] in their deep study of Israeli news television, structure in media dialog (in contrast to spontaneous natural conversation) is uniquely determined by its speaker role dynamics. Wang et al. [2011] investigate the detection of one such dynamic: agreement/disagreement between speakers. Ma et al. [2019] classify discourse relations (e.g. comparative, temporal) between two turns of dialog, but do not study discourse structure. In this work we extend our analysis to other properties of interviewer utterances (e.g. subjectivity, polarity, dialog act patterns) [Heritage, 1985] in the context of generative dialog modeling. Structured approaches for dialog modeling employ a simple concatenation of dialog history in a transformer-based architecture [Zhang et al., 2019]. We draw inspiration from Luan et al. [2017] who demonstrate the usefulness of a multi-task framework for speaker-conditioned dialog modeling. Guu et al. [2020a] propose a framework for jointly learning document retrieval and language modeling, and we propose a similar model to learn task-specific annotation of grounding documents.

10.3 Related Work for PABST

A desired impact of the proposed approach is increase in diversity of the generated responses. To tackle the issue of diversity in dialog model outputs, prior work has focused on decoding strategies such as diversity-promoting sampling [Holtzman et al., 2020]; training strategies such as discouraging undesirable responses via unlikelihood training [Li et al., 2020];

model changes such as using stochastic variables [Serban et al., 2017]; and using external data such as forum data [Su et al., 2020] or external knowledge bases [Majumder et al., 2020a]. In contrast to these, our proposed method generates responses with background stories using a gradient-based decoding approach.

One of the steps in our proposed approach is to retrieve relevant stories from an external corpus. Prior work has explored using retrieval of similar dialog instances as an initial step in improving response diversity and other human-like desiderata in dialog [Roller et al., 2021, Weston et al., 2018]. Distant supervision by using retrieved text snippets as pseudo responses has been explored in prior work [Su et al., 2020, Roller et al., 2021]. We use an external data source to improve dialog responses, a theme shared with some efforts in other tasks such as machine translation [Khandelwal et al.]. The use of narrative text in dialog has been explored in prior work, mostly as a 'script' or template for conversation [Xu et al., 2020a, Zhu et al., 2020]. We adapted a BERT-based retrieval method [Zhang et al., 2020] in our case to retrieve relevant story given dialog context and use retrieved story in the decoding phase.

Gradient-based for text generation with soft constraints has been explored in prior work [Dathathri et al., 2020, Qin et al., 2020]. Song et al. [2020] focused on generating response which are consistent to given persona. Differently, we use a gradient-based decoding to generate a dialog response while honoring constraints such as consistency to persona and similarity to retrieved story.

10.4 Related Work for POKI

Knowledge grounded dialog datasets such as Wizard-of-Wikipedia [Dinan et al., 2019b] and Topical chat [Gopalakrishnan et al., 2019a] typically consist of dialog responses paired with relevant knowledge available as collected annotations. Hence, models trained on such datasets are restricted to the knowledge sources they were exposed to at training time. Past work [Sun et al., 2021, Majumder et al., 2020a, Su et al., 2020, Komeili et al., 2021, Adolphs et al., 2021, Ghazvininejad et al., 2018, Tuan et al., 2020, Lewis et al., 2020c, Guu et al., 2020b] has looked into injecting extra knowledge sources at training time in a bid to add knowledge not available originally as paired to dialog responses. However, such approaches require re-training the model if some new knowledge source were to be used. Moreover, while previous work focuses on just improving specificity of dialog response using external knowledge, we also study the effect of additional knowledge in achieving conversational goals.

Improving the diversity of dialog responses by using diversity-promoting sampling has been explored in past work [Fan et al., 2018, Holtzman et al., 2020]. We use a gradient-based decoding method, building on past work in this direction [Dathathri et al., 2020, Qin et al., 2020, Madotto et al., 2020, Majumder et al., 2021a]. However, we propose new objectives to inject post-hoc knowledge obtained based on already generated dialog—an unsupervised knowledge injection method that has not been explored so far.

10.5 Related Work for Clarification Question Generation

Most previous work on question generation focused on generating reading comprehension style questions i.e., questions that ask about information present in a given text [Duan et al., 2017, Zhang and Bansal, 2019]. Later, Rao and Daumé III [2018, 2019] introduced the task of clarification question generation in order to ask questions about missing information in a given context. ClarQ [Kumar and Black, 2020] entails clarification questions in a question answering setup. However, unlike our work, these works still suffer from estimating the most useful missing information.

Recent works on conversational question answering also focused on the aspect of question generation or retrieval [Choi et al., 2018, Aliannejadi et al., 2019]. Qi et al. [2020a] especially focused on generating information-seeking questions while Majumder et al. [2020b] proposed a question generation task in free-form interview-style conversations. In this work, in addition to improving clarification question generation in a community-QA dataset, we are the first to explore a goal-oriented dialog scenario as well.

Representing context and associated global information in a structure format has been shown to improve performance in generation task [Das et al., 2019, Subramanian et al., 2018, Khashabi et al., 2017] in general and summarization [Fan et al., 2019] and story-generation [Yao et al., 2019] in particular. We also derive inspiration from recent works on information extraction from free-form text [Vedula et al., 2019, Stanovsky et al., 2016] and develop a novel framework to estimate missing information from available natural text contexts.

Finally, for question generation, we use BART [Lewis et al., 2019a], that is state-of-the-art for many generation tasks such as summarization, dialog generation etc. Furthermore, inspired from recent works that use controlled language generation during decoding [Ghazvininejad et al., 2017, Holtzman et al., 2018], we use Plug-and-Play-Language-Model [Dathathri et al., 2019] to tune generations during decoding. While similar approaches for controllable generation [Keskar et al., 2019, See et al., 2019] have been proposed, we extend such efforts to enhance the usefulness of the generated clarification questions.

10.6 Related Work for RExC

Providing explanations for a model's predictions can be done either post-hoc (via methods that aim to explain already trained and fixed black-box models) or by building self-explainable models (by jointly producing predictions and explanations). Post-hoc explanations [Lundberg and Lee, 2017, Ribeiro et al., 2016] can be useful when one only has access to a high-performance[1] but black-box model. However, post-hoc explanatory methods have been shown to have certain downsides [Adebayo et al., 2018, Slack et al., 2020, Laugel et al., 2019, Camburu et al., 2021, Wiegreffe et al., 2021b, Camburu et al., 2019]. Moreover, self-explanatory models may benefit from the rich information in the explanations provided at training time [Schramowski et al., 2020, Stacey et al., 2022, Lazaridou et al., 2022]. In this work, we focus on self-explainable models to

---

[1] High performance on held-out sets does not guarantee that the models do the right thing for the right reasons [McCoy et al., 2019].

produce two predominant types of explanations: NLEs and ERs.

NLEs. A growing number of works in NL and VL focus on designing neural models that produce NLEs for their predictions to make these models accessible to their users [Hendricks et al., 2016, Camburu et al., 2018, Park et al., 2018, Kayser et al., 2021, Kim et al., 2018, Ling et al., 2017, Marasovic et al., 2020, Wang et al., 2019, Rajani et al., 2019, Zellers et al., 2019a]. Recently, Narang et al. [2020] achieved SOTA on NLEs for NL tasks by using a pre-trained language model (of 11B parameters, which can be prohibitively large). However, NLEs are sometimes produced separately from predictions [Marasovic et al., 2020, Brahman et al., 2021, Atanasova et al., 2020], which raises questions about their faithfulness. In some cases, they were even produced as a task in isolation (without predictions) [Ji et al., 2020]. Moreover, the majority of the existing models only produce NLEs, with few exceptions that produce both NLEs and ERs [Park et al., 2018, Wu and Mooney, 2019], as our model does. Furthermore, an analysis on the faithfulness of NLEs is usually missing from the large majority of these works. To our knowledge, only one work recently introduced general necessary conditions for faithfulness in NLEs [Wiegreffe et al., 2021b], while few other works attempted architecture-specific faithfulness measures [Kumar and Talukdar, 2020, Wu and Mooney, 2019].

ERs. An early work [Zaidan and Eisner, 2008] investigated rationale extraction from inputs and later was successfully followed by works for both NL [DeYoung et al., 2020a, Lei et al., 2016b, Bastings et al., 2019a, Sha et al., 2021] and VL [Strout et al., 2019] tasks. We model both ERs and NLEs jointly in a novel framework that improves the quality of both types of explanations.

Knowledge Grounding. Free-text generation tasks heavily rely on background knowledge (e.g., commonsense). Several tasks such as dialog generation [Majumder et al., 2020a], creative text generation [Chakrabarty et al., 2020a, Mao et al., 2019], and counterfactual generation [Bhagavatula et al., 2020] used commonsense for grounding. Recently, Marasovic et al. [2020], Brahman et al. [2021] showed that external knowledge can be useful in separately justifying predictions using NLEs. In this work, we establish that knowledge grounding can be useful in a self-rationalizing framework benefiting both predictions and explanations.

Chapter 11

Conclusion and Future Outlook

In this dissertation, we show the AI systems with underlying language models can be redesigned to start with individual needs. We developed training-time and post-hoc knowledge injection techniques that promoted user efficiency and success in achieving task-specific goals. We developed inquisitive user-centric systems that can clarify for any ambiguous or incomplete context. We showed that knowledge-grounding not only improves model relevance to the user but also improves faithfulness in model explanations. Finally, our human-in-the-loop system system establishes that user can progressively improve model performance by intervening model explanations.

11.1 Future Work for Knowledge Grounding

My work so far significantly reduces the knowledge gap in existing language models by opening up the opportunity to make this process more *pro-user*.

Building user-specific knowledge representations: I am interested in developing novel ways to capture dynamic user knowledge and their possible world knowledge inferences. These methods will enhance personalized query refinement for more accurate knowledge retrieval, better knowledge attribution, and privacy-preserving actions in user-centric agents (e.g., Alexa). For example, I plan to update POKI's knowledge retrieval with user information using a memory-based architecture [Tandon et al., 2022].

Long-tail knowledge augmentation: I also want to improve current knowledge augmentation techniques with a special focus on long-tail documents. I am interested in developing augmentation methods (using meta-knowledge from LLMs) that are equally sensitive toward sparse (domain-specific, user-specific, task-specific) knowledge. Such latent knowledge selection can further be made user-aware by infusing it with user metadata.

11.2 Future Work for Generating Explanations

My work on knowledge grounding in model explanations proves to be an efficient regularizer to steer away from spurious correlations learned from the data. My future goal is to make this effect more *pro-user* by focusing on improving soundness in reasoning.

Reasoning in generated explanations: Most knowledge-grounded explainable models, including REXC, lack explicit reasoning structures in generated explanations (such as a hierarchy of justifications, assumptions, and exceptions). I want to focus on categorizing model reasoning errors and potentially building a meta-learner to uncover high-level reasoning ontologies. The variety of reasoning structures will be critical at explaining a prediction differently (a.k.a. personally) to different users.

Personalized explanations: Explanations are subjective. For example, one would explain a science topic *differently* to a child than an adult. Similarly, explaining ethical situations varies according to the user's stance (utilitarian, hedonistic, etc.). I am currently building demography-centric NLE datasets and developing novel methods with psychological, social, and cultural evidence.

11.3 Future Work for Interactive Systems

My work on interactive systems dealing with explanations is the first-step towards next-generation intelligent systems that incorporate users as an important aspect of the learning process. My future goal is to make this *pro-user* process more effective.

Teachable AI models: Updating the model's belief during test time for improved predictions can be extended to any model capable of producing explanations for its predictions. This model debugging process can also be viewed as teaching AI models. My ongoing work on machine teaching Dalvi et al. [2022] to improve the model's capability of better domain understanding (e.g., science) has promise for building never-ending learning systems to improve continually over time.

Achieving generalization with user feedback: To address the debate about the possibility of persisting user feedback for better generalization on unseen data points, I plan to explore two predominant approaches: memory-based architectures [Tandon et al., 2022, Dalvi et al., 2022] and model editing approaches [Mitchell et al., 2022a,b]. For example, INTERFAIR-like applications can benefit from user feedback by learning and memorizing how humans perceive biases.

Critiquable AI models: It is also non-trivial to extend the critiquing process to any AI model (e.g., SVM, decision trees). I want to develop a generalized framework for critiquing explanations that unify various explanation formats (e.g., feature-based, natural language) and feedback synthesis methods (e.g., post-hoc, persistent). Expanding INTERFAIR-like critiquing techniques to traditional AI models (decision trees) will directly apply to domains such as healthcare, where end users (patients or providers) can verify, locate, and update model biases to ensure fair outcomes.

In summary, our research takes a user-centric approach to achieve subjectivity and personalization in AI models—a long-standing goal of artificial general intelligence. It is possible transformative technical and societal impacts by redesigning AI development as a closed loop: knowledge $\rightarrow$ explanations $\rightarrow$ interactions; where users remain an integral part. Subsequently, this will make more positive impacts in society using intelligent systems to address learning differences among users, build skills, and reduce ambiguity in communication, among many others.

Bibliography

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, volume 31, 2018.

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. Reason first, then respond: Modular generation for knowledge-infused dialogue. *CoRR*, abs/2111.05204, 2021. URL https://arxiv.org/abs/2111.05204.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *SIGIR*. ACM, 2019. doi: 10.1145/3331184.3331265. URL https://doi.org/10.1145/3331184.3331265.

Reinald Kim Amplayo, Seonjae Lim, and Seung-won Hwang. Entity commonsense representation for neural abstractive summarization. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *NAACL-HLT*, 2018. doi: 10.18653/v1/n18-1064. URL https://doi.org/10.18653/v1/n18-1064.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In *ACL*, pages 7352–7364, 2020.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *WWW*, 2019.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL*, pages 65–72, 2005a.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005b.

Siqi Bao, Huang He, Fan Wang, Rongzhong Lian, and Hua Wu. Know more about each other: Evolving dialogue strategy via compound assessment. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL*, 2019. doi: 10.18653/v1/p19-1535. URL https:

//doi.org/10.18653/v1/p19-1535.

Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker. The rules behind roles: Identifying speaker role in radio broadcasts. In *AAAI*, 2000. URL http://www.aaai.org/Library/AAAI/2000/aaai00-104.php.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *ACL*, pages 2963–2977, 2019a.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *ACL*, Florence, Italy, July 2019b. doi: 10.18653/v1/P19-1284. URL https://aclanthology.org/P19-1284.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. Commonsense for generative multi-hop question answering tasks. In *EMNLP*, pages 4220–4230, 2018.

Monika Bednarek. *Evaluation in media discourse: Analysis of a newspaper corpus*. A&C Black, 2006.

Doug Beeferman, William Brannon, and Deb Roy. Radiotalk: A large-scale corpus of talk radio transcripts. In *INTERSPEECH*, 2019. doi: 10.21437/Interspeech.2019-2714.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *ICLR*, 2020.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. In *ICLR*, 2017. URL https://openreview.net/forum?id=S1Bb3D5gg.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL*, 2019. doi: 10.18653/v1/p19-1470. URL https://doi.org/10.18653/v1/p19-1470.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *SIGNLL*, 2016.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642, 2015.

Faeze Brahman and Snigdha Chaturvedi. Modeling protagonist emotions for emotion-aware storytelling. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, pages 5277–5294, 2020. doi: 10.18653/v1/2020.emnlp-main.426. URL https://doi.org/10.18653/v1/2020.emnlp-main.426.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *AAAI*, pages 12592–12601, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. BART for knowledge grounded conversations. In Giuseppe Di Fabbrizio, Surya Kallumadi, Utkarsh Porwal, and Thrivikrama Taula, editors, *Converse@KDD*, volume 2666. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2666/KDD_Converse20_paper_7.pdf.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *EMNLP*, 2018. URL https://www.aclweb.org/anthology/D18-1547/.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *NeurIPS*, pages 9560–9572, 2018.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. Can I trust the explainer? Verifying post-hoc explanatory methods. In *NeurIPS Workshop Safety and Robustness in Decision Making*, 2019.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. Make up your mind! Adversarial generation of inconsistent natural language explanations. In *ACL*, pages 4157–4165, 2020.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets. In *AAAI Workshop on Explainable Agency in Artificial Intelligence*, 2021.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 2020.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *ACL*, pages 7976–7986, 2020a.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. *CoRR*, abs/2004.13248, 2020b. URL https://arxiv.org/abs/2004.13248.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. Grounding 'grounding' in NLP. In *Findings of ACL*, pages 4283–4305, 2021.

Langzhou Chen, Lori Lamel, Jean-Luc Gauvain, and Gilles Adda. Dynamic language modeling for broadcast news. In *INTERSPEECH*. URL http://www.isca-speech.org/archive/interspeech_2004/i04_0997.html.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. KGPT: knowledge-grounded pre-training for data-to-text generation. In *EMNLP*, pages 8635–8648, 2020a.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. Distilling the knowledge of BERT for text generation. *CoRR*, abs/1911.03829, 2019. URL http://arxiv.org/abs/1911.03829.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation learning. In *ECCV*, pages 104–120, 2020b.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *EMNLP*, 2018. doi: 10.18653/v1/d18-1241. URL https://doi.org/10.18653/v1/d18-1241.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip R. Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What makes a good conversation?: Challenges in designing truly conversational agents. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *CHI*, 2019. doi: 10.1145/3290605.3300705. URL https://doi.org/10.1145/3290605.3300705.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Jordan Cohen. The gale project: A description and an update. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 237–237. IEEE, 2007.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. Towards teachable reasoning systems. *arXiv* *preprint arXiv:2204.13074*, 2022.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. Building dynamic knowledge graphs from text using machine reading comprehension. In *ICLR*, 2019. URL https://openreview.net/forum?id=S1lhbnRqF7.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *ICLR*, 2019.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *ICLR*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *CACM*, 2015. doi: 10.1145/2701413. URL https://doi.org/10.1145/2701413.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *FAT*, 2019.

Marco De Boni and Suresh Manandhar. An analysis of clarification dialogue for question answering. In *NAACL-HLT*, 2003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL*, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL*, pages 4443–4458, 2020a.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*, 2020b.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098, 2019a. URL http://arxiv.org/abs/1902.00098.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*, 2019b. URL https://openreview.net/forum?id=r1l73iRqKm.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*, 2019c. URL https://openreview.net/forum?id=r1l73iRqKm.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *EMNLP*, 2017. doi: 10.18653/v1/d17-1090. URL https://doi.org/10.18653/v1/d17-1090.

Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. Human conversational behavior. *Human nature*, 8(3):231–246, 1997.

Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. The EPAC corpus: Manual and automatic annotations of conversational speech in french broadcast news. In *LREC*, 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/650.html.

Norman Fairclough. Discourse representation in media discourse. *Sociolinguistics*, 17(2): 125–139, 1988.

Norman Fairclough and Ruth Wodak. Critical discourse analysis. *Discourse studies: A multidisciplinary introduction*, 2:258–284, 1997.

Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *ACL*, 2018. doi: 10.18653/v1/P18-1082. URL https://www.aclweb.org/anthology/P18-1082/.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP-IJCNLP*, 2019. doi: 10.18653/v1/D19-1428. URL https://doi.org/10.18653/v1/D19-1428.

Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *CoRR*, abs/2003.05162, 2020. URL https://arxiv.org/abs/2003.05162.

Christian Federmann, Oussama Elachqar, and Chris Quirk. Multilingual whispers: Generating paraphrases with translation. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *W-NUT@EMNLP*, 2019. doi: 10.18653/v1/D19-5503. URL https://doi.org/10.18653/v1/D19-5503.

Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130, 1988. doi: 10.1016/0893-6080(88)90014-7. URL https://doi.org/10.1016/0893-6080(88)90014-7.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Àgata Lapedriza, and Rosalind W. Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fc9812127bf09c7bd29ad6723c683fb5-Abstract.html.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. Hafez: an interactive poetry generation system. In *ACL, System Demonstrations*, 2017.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16710.

Augusto Gnisci and Marino Bonaiuto. Grilling politicians: Politicians' answers to questions in television interviews and courtroom examinations. *Journal of language and social psychology*, 22(4):385–413, 2003.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL-HLT*, 2019.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-chat: Towards knowledge-grounded open-domain conversations. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech*, 2019a. doi: 10.21437/Interspeech.2019-3079. URL https://doi.org/10.21437/Interspeech.2019-3079.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Interspeech*, 2019b. doi: 10.21437/Interspeech.2019-3079. URL http://dx.doi.org/10.21437/Interspeech.2019-3079.

Clinton Gormley and Zachary Tong. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. "O'Reilly Media, Inc.", 2015.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional LSTM networks for improved phoneme classification and recognition. In *ICANN*, 2005. doi: 10.1007/11550907_126. URL https://doi.org/10.1007/11550907_126.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. A knowledge-enhanced pretraining model for commonsense story generation. *TACL*, 2020. URL https://transacl.org/ojs/index.php/tacl/article/view/1886.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909, 2020a. URL https://arxiv.org/abs/2002.08909.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909, 2020b. URL https://arxiv.org/abs/2002.08909.

Pentti Haddington. Stance taking in news interviews. *SKY Journal of Linguistics*, 17:101–142, 2004.

Peter Hajas, Louis Gutierrez, and Mukkai S. Krishnamoorthy. Analysis of yelp reviews. *CoRR*, abs/1407.1443, 2014. URL http://arxiv.org/abs/1407.1443.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of EMNLP*, pages 4351–4367, 2020.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding. In *Findings of EMNLP*, pages 4173–4181, 2021.

Zexue He, Yu Wang, Julian McAuley, and **Bodhisattwa Prasad Majumder**. Controlling bias exposure for fair interpretable predictions. *Findings of EMNLP*, 2022.

Stefan Heindorf, Yan Scholten, Gregor Engels, and Martin Potthast. Debiasing vandalism detection models at wikidata. In *WWW*, 2019.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *ECCV*, pages 3–19, 2016.

John Heritage. Analyzing news interviews: Aspects of the production of talk for an'overhearing'audience. *Handbook of Discourse Analysis, vol. III: Discourse and Dialogue*, 1985.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.

9.8.1735.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. In Iryna Gurevych and Yusuke Miyao, editors, *ACL*, 2018. doi: 10.18653/v1/P18-1152. URL https://www.aclweb.org/anthology/P18-1152/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e946209592563be0f01c844ab2170f0c-Abstract.html.

Qian Huang, Zhu Liu, Aaron E. Rosenberg, David C. Gibbon, and Behzad Shahraray. Automated generation of news content hierarchy by integrating audio, video, and text information. In *ICASSP*, 1999. doi: 10.1109/ICASSP.1999.757478. URL https://doi.org/10.1109/ICASSP.1999.757478.

Brian Hutchinson, Bin Zhang, and Mari Ostendorf. Unsupervised broadcast conversation speaker role labeling. In *ICASSP*, 2010. doi: 10.1109/ICASSP.2010.5494958. URL https://doi.org/10.1109/ICASSP.2010.5494958.

Robert L. Logan IV, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL*, 2019. doi: 10.18653/v1/p19-1598. URL https://doi.org/10.18653/v1/p19-1598.

Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution. *TACL*, pages 294–310, 2021.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. Semantics derived automatically from language corpora contain human-like moral choices. In *AIES*, 2019.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *EMNLP*, 2018. doi: 10.18653/v1/d18-1436. URL https://doi.org/10.18653/v1/d18-1436.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *ACL 2018*, 2018. URL https://www.aclweb.org/anthology/P18-1154/.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, and Minlie Huang. Generating commonsense explanation by extracting bridge concepts from reasoning paths. In *AACL/IJCNLP*, pages 248–257, 2020.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 2003.

Lauri Karttunen. Syntax and semantics of questions. *Linguistics and philosophy*, 1(1):3–44, 1977.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-ViL: A dataset and benchmark for natural language explanations in vision-language tasks. In *ICCV*, 2021.

Chris Kedzie, Kathleen R. McKeown, and Hal Daumé III. Content selection in deep learning models of summarization. In *EMNLP*, 2018. doi: 10.18653/v1/d18-1208. URL https://doi.org/10.18653/v1/d18-1208.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858, 2019. URL http://arxiv.org/abs/1909.05858.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. *CoRR*. URL https://arxiv.org/abs/2010.00710.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. Learning what is essential in questions. In *CoNLL*, 2017.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. Sequential latent knowledge selection for knowledge-grounded dialogue. In *ICLR*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Hke0K1HKwr.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *ECCV*, pages 577–593, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2014. URL http://arxiv.org/abs/1312.6114.

Walter Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, 1988.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *CoRR*, abs/2107.07566, 2021. URL https://arxiv.org/abs/2107.07566.

Xiang Kong, Bohan Li, Graham Neubig, Eduard H. Hovy, and Yiming Yang. An adversarial approach to high-quality, sentiment-controlled neural dialogue generation. *CoRR*, abs/1901.07129, 2019. URL http://arxiv.org/abs/1901.07129.

Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*. Omnipress, 2011. URL https://icml.cc/2011/papers/611_icmlpaper.pdf.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5(2-3):123–286, 2012. doi: 10.1561/2200000044. URL https://doi.org/10.1561/2200000044.

Sawan Kumar and Partha P. Talukdar. NILE: Natural language inference with faithful natural language explanations. In *ACL*, pages 8730–8742, 2020.

Vaibhav Kumar and Alan W. Black. Clarq: A large-scale and diverse dataset for clarification question generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *ACL*, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.651/.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*, 2022.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*, 2020.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *IJCAI-19*, pages 2801–2807, 7 2019.

Antoine Laurent, Nathalie Camelin, and Christian Raymond. Boosting bonsai trees for efficient features combination: application to speaker role identification. In *INTERSPEECH*, 2014. URL http://www.isca-speech.org/archive/interspeech_2014/i14_0076.html.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. *CoRR*, abs/2203.05115, 2022.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. *To appear in "Predicting Structured Data*, 1:0, 2006.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *EMNLP*, 2017. doi: 10.18653/v1/d17-1018. URL https://doi.org/10.18653/v1/d17-1018.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *EMNLP*, Austin, Texas, November 2016a. doi: 10.18653/v1/D16-1011. URL https://aclanthology.org/D16-1011.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Rationalizing neural predictions. In *EMNLP*, pages 107–117, 2016b.

Douglas Lenat and CycL Lisp. Knowledge base.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019a.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019b. URL http://arxiv.org/abs/1910.13461.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *ACL*, 2020a. doi: 10.18653/v1/2020.acl-main.703. URL https://doi.org/10.18653/v1/2020.acl-main.703.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020c. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and

Owen Rambow, editors, *NAACL HLT*, 2016a. doi: 10.18653/v1/n16-1014. URL https://doi.org/10.18653/v1/n16-1014.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pages 110–119, 2016b.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. *CoRR*, abs/1911.03860, 2019a. URL http://arxiv.org/abs/1911.03860.

Margaret Li, Jason Weston, and Stephen Roller. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087, 2019b. URL http://arxiv.org/abs/1909.03087.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *ACL*, 2020. doi: 10.18653/v1/2020.acl-main.428. URL https://doi.org/10.18653/v1/2020.acl-main.428.

Shuyang Li, Bodhisattwa Prasad Majumder, and Julian McAuley. Self-supervised bot play for transcript-free conversational recommendation with rationales. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 327–337, 2022a.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, 2017.

Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie CK Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. *arXiv preprint arXiv:2204.03025*, 2022b.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. Learning to select knowledge for response generation in dialog systems. In Sarit Kraus, editor, *IJCAI*, 2019. doi: 10.24963/ijcai.2019/706. URL https://doi.org/10.24963/ijcai.2019/706.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *ACL*, pages 158–167, 2017.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *EMNLP*, 2016. doi: 10.18653/v1/d16-1230. URL https://doi.org/10.18653/v1/d16-1230.

Yang Liu. Initial study on automatic identification of speaker role in broadcast news speech. In *NAACL-HLT*, 2006. URL https://www.aclweb.org/anthology/N06-2021/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDial*, 2015.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. Multi-task learning for speaker-role adaptation in neural conversation models. In *IJCNLP*, 2017. URL https://www.aclweb.org/anthology/I17-1061/.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*. 2017.

Mingyu Derek Ma, Kevin Bowden, JiaQi Wu, Wen Cui, and Marilyn A. Walker. Implicit discourse relation identification for open-domain dialogues. In *ACL*, 2019. doi: 10.18653/v1/p19-1065. URL https://doi.org/10.18653/v1/p19-1065.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. Plug-and-play conversational models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of EMNLP*, 2020. doi: 10.18653/v1/2020.findings-emnlp.219. URL https://doi.org/10.18653/v1/2020.findings-emnlp.219.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian J. McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020a. doi: 10.18653/v1/2020.emnlp-main.739. URL https://doi.org/10.18653/v1/2020.emnlp-main.739.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian J. McAuley. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *EMNLP*, 2020b. doi: 10.18653/v1/2020.emnlp-main.653. URL https://doi.org/10.18653/v1/2020.emnlp-main.653.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian J. McAuley, and Harsh Jhamtani. Unsupervised enrichment of persona-grounded dialog with background stories. In *ACL*, 2021a. doi: 10.18653/v1/2021.acl-short.74. URL https://doi.org/10.18653/v1/2021.acl-short.74.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian J. McAuley. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. *NAACL*, 2021b. URL https://arxiv.org/abs/2104.06828.

Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian J. McAuley, and Garrison W. Cottrell. Improving neural story generation by targeted common sense grounding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP*, 2019. doi: 10.18653/v1/D19-1615. URL https://doi.org/10.18653/v1/D19-1615.

Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison W. Cottrell. Speech recognition and multi-speaker diarization of long conversations. *CoRR*, abs/2005.08072, 2020.

Ana Marasovic, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *EMNLP Findings*, pages 2810–2829, 2020.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *EMNLP*, 2018. doi: 10.18653/v1/d18-1298. URL https://doi.org/10.18653/v1/d18-1298.

Julian McAuley and Alex Yang. Addressing complex and subjective product-related queries with customer reviews. In *WWW*, 2016.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*, 2019.

Sarah McLeod, Ivana Kruijff-Korbayova, and Bernd Kiefer. Multi-task learning of system dialogue act selection for supervised pretraining of goal-oriented dialogue policies. In *SIGDial*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-5947.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 2019.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *ICLR*, 2022a.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *ICML*, 2022b.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios,

A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *AAAI*, 2015.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *CoRR*, abs/1909.08855, 2019. URL http://arxiv.org/abs/1909.08855.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL*, 2019. doi: 10.18653/v1/p19-1081. URL https://doi.org/10.18653/v1/p19-1081.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. *CoRR*, abs/1604.01696, 2016. URL http://arxiv.org/abs/1604.01696.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. WT5?! Training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546, 2020.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *EMNLP*, 2017. doi: 10.18653/v1/d17-1238. URL https://doi.org/10.18653/v1/d17-1238.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. Improving compositional generalization in semantic parsing. *arXiv preprint arXiv:2010.05647*, 2020.

Vishakh Padmakumar and He He. Unsupervised extractive summarization using pointwise mutual information. In *EACL*, 2021. URL https://aclanthology.org/2021.eacl-main.213/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002a. URL https://www.aclweb.org/anthology/P02-1040/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002b.

Ashwin Paranjape and Christopher D. Manning. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *NAACL-HLT*, 2021. doi: 10.18653/v1/2021.naacl-main.61. URL https://doi.org/10.18653/v1/2021.naacl-main.61.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, pages 8779–8788, 2018.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visual-COMET: Reasoning about the dynamic context of a still image. In *ECCV*, pages 508–524, 2020.

Prasanna Parthasarathi and Joelle Pineau. Extending neural generative conversational model using external knowledge sources. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *EMNLP*, 2018. doi: 10.18653/v1/d18-1073. URL https://doi.org/10.18653/v1/d18-1073.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP-IJCNLP*, 2019. doi: 10.18653/v1/D19-1250. URL https://doi.org/10.18653/v1/D19-1250.

Paul Placeway, Scotte Chen, Maxine Eskénazi, Uday Jain, Vipul N. Parikh, Bhiksha Raj, Mosur Ravishankar, Rogério Rosenfeld, Kristie Seymore, Matthew A. Siegler, Richard M. Stern, and Eric H. Thayer. The 1996 hub-4 sphinx-3 system. In *Proc. DARPA Speech recognition workshop*, volume 97, 1997.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W. Black, and Ruslan Salakhutdinov. Focused attention improves document-grounded generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *NAACL-HLT*, 2021. doi: 10.18653/v1/2021.naacl-main.338. URL https://doi.org/10.18653/v1/2021.naacl-main.338.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020a. URL https://www.aclweb.org/anthology/2020.findings-emnlp.3/.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In Asli Çelikyilmaz and Tsung-Hsien Wen, editors, *ACL Demo*, 2020b. URL https://www.aclweb.org/anthology/2020.acl-demos.14/.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020. doi: 10.18653/v1/2020.emnlp-main.58. URL https://doi.org/10.18653/v1/2020.emnlp-main.58.

Alec Radford. Improving language understanding by generative pre-training. 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *Tech report*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! Leveraging language models for commonsense reasoning. In *ACL*, pages 4932–4942, 2019.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *ACL*, 2018.

Sudha Rao and Hal Daumé III. Answer-based Adversarial Training for Generating Clarification Questions. In *NAACL*, June 2019. URL https://www.aclweb.org/anthology/N19-1013.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *ACL*, 2020.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *SIGKDD*, 2016.

Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Inf. Process. Manag.*, 31(3):345–360, 1995. doi: 10.1016/0306-4573(94)00051-4. URL https://doi.org/10.1016/0306-4573(94)00051-4.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *EACL*, 2021. URL https://www.aclweb.org/anthology/2021.eacl-main.24/.

Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0. URL https://doi.org/10.1016/0306-4573(88)90021-0.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. In *ACL*, 2019. doi: 10.18653/v1/p19-1004. URL https://doi.org/10.18653/v1/p19-1004.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019. doi: 10.1609/aaai.v33i01.33013027. URL https://doi.org/10.1609/aaai.v33i01.33013027.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. In *Nature Machine Intelligence*, 2020.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Commun. ACM*, 63 (12):54–63, 2020. doi: 10.1145/3381831. URL https://doi.org/10.1145/3381831.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL-HLT*, 2019. doi: 10.18653/v1/n19-1170. URL https://doi.org/10.18653/v1/n19-1170.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: Learning robust metrics for text generation. In *ACL*, pages 7881–7892, 2020.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In Satinder P. Singh and Shaul Markovitch, editors, *AAAI*, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567.

Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. Learning from the best: Rationalizing predictions by adversarial information calibration. In *AAAI*, pages 13771–13779, 2021.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *EMNLP*, 2017. doi: 10.18653/v1/d17-1235. URL https://doi.org/10.18653/v1/d17-1235.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020. doi: 10.18653/v1/2020.emnlp-main.373. URL https://doi.org/10.18653/v1/2020.emnlp-main.373.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AIES*, 2020.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Talktomodel: Understanding machine learning models with open ended dialogues. *arXiv preprint* *arXiv:2207.04154*, 2022.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. Generating persona consistent dialogues by exploiting natural language inference. *CoRR*, abs/1911.05889, 2019a. URL http://arxiv.org/abs/1911.05889.

Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. Exploiting persona information for diverse generation of conversational responses. In Sarit Kraus, editor, *IJCAI*, 2019b. doi: 10.24963/ijcai.2019/721. URL https://doi.org/10.24963/ijcai.2019/721.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. Generating persona consistent dialogues by exploiting natural language inference. In *AAAI*, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6417.

Brian G Southwell, Emily A Thorson, and Laura Sheble. *Misinformation and mass audiences*. University of Texas Press, 2018.

Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451, 2017.

Joe Stacey, Yonatan Belinkov, and Marek Rei. Natural language inference with a human touch: Using human explanations to guide model attention. In *AAAI*, 2022.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. Getting more out of syntax with props. *CoRR*, abs/1603.01648, 2016. URL http://arxiv.org/abs/1603.01648.

Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. Towards natural clarification questions in dialogue systems. In *AISB symposium on questions, discourse and dialogue*, volume 20, 2014.

Stephanie M Strassel. Linguistic resources for effective, affordable, reusable speech-to-text. In *LREC*, 2004.

Julia Strout, Ye Zhang, and Raymond J. Mooney. Do human rationales improve machine explanations? *CoRR*, abs/1905.13714, 2019.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *ACL*, 2019. doi: 10.18653/v1/p19-1355. URL https://doi.org/10.18653/v1/p19-1355.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. Diversifying dialogue generation with non-conversational text. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *ACL*, 2020. doi: 10.18653/v1/2020.acl-main.634. URL https://doi.org/10.18653/v1/2020.acl-main.634.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. Neural models for key phrase extraction and question generation. In Eunsol Choi, Minjoon Seo, Danqi Chen, Robin Jia, and Jonathan Berant, editors, *ACL MRQA*, 2018. doi: 10.18653/v1/W18-2609. URL https://www.aclweb.org/anthology/W18-2609/.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. Adding chit-chats to enhance task-oriented dialogues. *NAACL*, 2021. URL https://arxiv.org/abs/2010.12757.

Niket Tandon, Aman Madaan, Peter Clark, Keisuke Sakaguchi, and Yiming Yang. Interscript: A dataset for interactive learning of scripts through error feedback. *CoRR*, abs/2112.07867, 2021. URL https://arxiv.org/abs/2112.07867.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. *NAACL Findings.(to appear)*, 2022.

Yi-Lin Tuan, Wei Wei, and William Yang Wang. Unsupervised injection of knowledge into dialogue generation via language models. *CoRR*, abs/2004.14614, 2020. URL https://arxiv.org/abs/2004.14614.

Peter D. Turney. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *TACL*, 2013. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/128.

Teun A van Dijk. *Discourse and communication: New approaches to the analysis of mass media discourse and communication*, volume 10. Walter de Gruyter, 2011.

Suus MJ Van Hekken and Wim Roelofsen. More questions than answers: A study of question–answer sequences in a naturalistic setting. *Journal of Child Language*, 9(2):445–460, 1982.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017a. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeuRIPS*, 2017b.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. Towards open intent discovery for conversational text. *CoRR*, abs/1904.08524, 2019. URL http://arxiv.org/

abs/1904.08524.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. Does it make sense? And why? A pilot study for sense making and explanation. In *ACL*, July 2019.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. SemEval-2020 Task 4: Commonsense validation and explanation. In *SemEval*, 2020.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *CoRR*, abs/2104.14690, 2021.

Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey. Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *ICASSP*, 2011. doi: 10.1109/ICASSP.2011.5947618. URL https://doi.org/10.1109/ICASSP.2011.5947618.

Elda Weizman. *Positioning in media dialogue: Negotiating roles in the news interview*, volume 3. John Benjamins Publishing, 2008.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL*, 2019. doi: 10.18653/v1/p19-1363. URL https://doi.org/10.18653/v1/p19-1363.

Jason Weston, Emily Dinan, and Alexander H. Miller. Retrieve and refine: Improved sequence generation models for dialogue. In Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail S. Burtsev, editors, *SCAI@EMNLP*, 2018. doi: 10.18653/v1/w18-5713. URL https://doi.org/10.18653/v1/w18-5713.

Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. Measuring association between labels and free-text rationales. In *EMNLP*, 2021a.

Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. Measuring association between labels and free-text rationales. In *EMNLP*, pages 10266–10284, 2021b.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. In *CIKM*. ACM, 2018. doi: 10.1145/3269206.3272018. URL https://doi.org/10.1145/3269206.3272018.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4), 1992.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149, 2019. URL http://arxiv.org/abs/1901.08149.

Jialin Wu and Raymond J. Mooney. Faithful multimodal explanation for visual question answering. In *ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, 2019.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *ACL*, 2020a. URL https://www.aclweb.org/anthology/2020.acl-main.515/.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In Christian Bessiere, editor, *IJCAI*, 2020b. doi: 10.24963/ijcai.2020/521. URL https://doi.org/10.24963/ijcai.2020/521.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706, 2019a.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation. *CoRR*, abs/1904.12848, 2019b. URL http://arxiv.org/abs/1904.12848.

Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. Enhancing dialog coherence with event graph grounded content planning. In Christian Bessiere, editor, *IJCAI*, 2020a. doi: 10.24963/ijcai.2020/545. URL https://doi.org/10.24963/ijcai.2020/545.

Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. A neural topical expansion framework for unstructured persona-oriented dialogue generation. *CoRR*, abs/2002.02153, 2020b. URL https://arxiv.org/abs/2002.02153.

Mohammad Yaghini, Andreas Krause, and Hoda Heidari. A human-in-the-loop framework to construct context-aware mathematical notions of outcome fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 1023–1033, 2021.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *AAAI*, 2019. doi: 10.1609/aaai.v33i01.33017378. URL https://doi.org/10.1609/aaai.v33i01.33017378.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16573.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph. *CoRR*, abs/2006.16934, 2020.

Omar Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *EMNLP*, pages 31–40, 2008.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*, 2018.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019a.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL*, 2019b. doi: 10.18653/v1/p19-1472. URL https://doi.org/10.18653/v1/p19-1472.

Daojian Zeng, Haoran Zhang, Lingyun Xiang, Jin Wang, and Guoliang Ji. User-oriented paraphrase generation with keywords controlled network. *IEEE Access*, 7, 2019. doi: 10.1109/ACCESS.2019.2923057. URL https://doi.org/10.1109/ACCESS.2019.2923057.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AIES*, 2018a.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *ACL*, 2018b. doi: 10.18653/v1/P18-1205. URL https://www.aclweb.org/anthology/P18-1205/.

Shiyue Zhang and Mohit Bansal. Addressing semantic drift in question generation for semi-supervised question answering. *CoRR*, abs/1909.06356, 2019. URL http://arxiv.org/abs/1909.06356.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536, 2019. URL http://arxiv.org/abs/1911.00536.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In Regina Barzilay and Min-Yen Kan, editors, *ACL*, 2017. doi: 10.18653/v1/P17-1061. URL https://doi.org/10.18653/v1/P17-1061.

Tingting Zhao, Hirotaka Hachiya, Gang Niu, and Masashi Sugiyama. Analysis and improvement of policy gradient estimation. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, 2011. URL http://papers.nips.cc/paper/4264-analysis-and-improvement-of-policy-gradient-estimation.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. Knowledge-grounded dialogue generation with pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020. doi: 10.18653/v1/2020.emnlp-main.272. URL https://doi.org/10.18653/v1/2020.emnlp-main.272.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In Jérôme Lang, editor, *IJCAI*, 2018. doi: 10.24963/ijcai.2018/643. URL https://doi.org/10.24963/ijcai.2018/643.

Yutao Zhu, Ruihua Song, Zhicheng Dou, Jian-Yun Nie, and Jin Zhou. Scriptwriter: Narrative-guided script generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *ACL*, 2020. doi: 10.18653/v1/2020.acl-main.765. URL https://doi.org/10.18653/v1/2020.acl-main.765.